ON THE PREDICTION OF LUNG CANCER RISK BASED ON LIFESTYLE AND HEALTH FACTORS OF PATIENTS

BY

NOAH, ROFIAT OLUWADAMILOLA

20/56EG086

BEING PROJECT REPORT SUBMITTED TO THE DEPARTMENT OF STATISTICS IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF SCIENCE (HONOURS) DEGREE IN STATISTICS OF THE UNIVERSITY OF ILORIN, ILORIN, NIGERIA.

JULY, 2025.

ATTESTATION

This is to certify that the project work entitled **"ON THE PREDICTION OF LUNG CANCER RISK BASED ON LIFESTYLE AND HEALTH FACTORS OF PATIENTS"** is an original work carried out by **NOAH, ROFIAT OLUWADAMILOLA** with Matriculation Number **20/56EG086**, under my supervision.

……..	……..
NOAH, ROFIAT OLUWADAMILOLA	**DR. M. K. GARBA**
20/56EG086	**SUPERVISOR**

CERTIFICATION

This work has been read and approved as meeting the partial requirement for the award of Bachelor of Science Degree (B.Sc.) in Statistics, University of Ilorin, Ilorin, Nigeria.

…….. ……

DR. M. K. GARBA **Date**

SUPERVISOR

…… ……

PROF. A. A. ABIODUN **Date**

HEAD OF DEPARTMENT

…… ……

PROF. O. O. ALABI **Date**

EXTERNAL EXAMINER

DEDICATION

This project is dedicated to my lovely parents, Mr. and Mrs. Noah, as well as my siblings.

ACKNOWLEDGEMENTS

All praise and glory be to Almighty God, the source of life, wisdom, strength, and understanding, for guiding me through every stage of this academic journey. His constant presence, mercy, and grace were my anchor during moments of challenge, uncertainty, and self-doubt.

I express my deepest gratitude to my supervisor, Dr. M. K. Garba, for his invaluable contribution to the success of this research. His profound expertise, patience, insightful feedback, and genuine dedication to my academic development were instrumental in bringing this project to fruition. It was truly a privilege to work under his guidance.

My sincere appreciation also goes to all the lecturers and staff of the Department of Statistics, University of Ilorin, for their relentless efforts in imparting knowledge and nurturing my academic growth. Their passion for teaching and commitment to students' successes have left a lasting impact on me, and I am proud to have been part of this academic family.

I am especially grateful to my wonderful parents, Mr. and Mrs. Noah, and my beloved siblings, whose unwavering love, prayers, sacrifices, and encouragement have been the bedrock of my educational journey. Their belief in me, even when I struggled to believe in myself, has made all the difference.

To my amazing friends: Abogunrin Quasim O. (Comr. Horla), Ayinla Faruq A. (Ayinla), Ojo Kehinde T. (Kenny), Ajiboye muibat O., Yakub Olawale M. (Discrete), Sulyman Olaitan S. (Expert), Afolabi Mubarak O. (U Feel), Olatinwo Jamiu O. (Format),), and my other coursemates, thank you for the shared experiences, companionship, and support that made this academic journey more fulfilling and enjoyable. Their presence made a real difference in my academic life, and I will always cherish our moments together.

ABSTRACT

Lung cancer is the leading cause of cancer-related mortality worldwide, often detected at advanced stages due to limited early screening tools. This study aims to develop a predictive model for lung cancer risk based on lifestyle and health-related factors using Chi-Square analysis and Binary Logistic Regression. A dataset comprising 309 individuals was analyzed to assess the relationship between lung cancer status and variables such as age, gender, smoking, yellow fingers, chronic diseases, coughing, shortness of breath, and chest pain.

Chi-Square tests revealed statistically significant associations between lung cancer and several predictors, including smoking ($p = 0.036$), yellow fingers ($p = 0.001$), coughing ($p < 0.001$), chest pain ($p = 0.001$), and shortness of breath ($p = 0.026$). Logistic Regression analysis confirmed that coughing (OR = 6.413), chronic disease (OR = 5.305), yellow fingers (OR = 3.816), chest pain (OR = 3.447), and smoking (OR = 2.009) significantly increased the odds of developing lung cancer. The model showed a classification accuracy of 87.7%, with a Nagelkerke R^2 of 0.314, indicating moderate explanatory power. The Hosmer-Lemeshow test ($p = 0.357$) confirmed a good model fit.

These findings suggest that statistical models using easily obtainable clinical and lifestyle variables can effectively predict lung cancer risk. This approach provides a cost-effective, interpretable, and practical tool for early identification of high-risk individuals, especially in resource-constrained settings where advanced diagnostic tools may not be readily available.

TABLE OF CONTENTS

CHAPTER ONE

GENERAL INTRODUCTION

1.1 BACKGROUND TO THE STUDY

Lung cancer stands as one of the most devastating forms of cancer, both in terms of its prevalence and its mortality. It is the foremost cause of cancer-related deaths globally, accounting for nearly 1.8 million deaths each year, which represents approximately 18% of all cancer deaths (World Health Organization, 2021). The disease is characterized by the uncontrolled growth of abnormal cells within the lung tissue, which, if not detected early, can invade surrounding tissues and metastasize to distant organs such as the liver, brain, adrenal glands, and bones. This capacity for aggressive spread contributes to the often-poor prognosis associated with lung cancer, especially when diagnosed at an advanced stage. Unfortunately, the majority of cases are not identified until they have reached a late phase, where treatment options are limited and largely palliative rather than curative (Siegel 2022).

The etiology of lung cancer is multifaceted, involving a combination of behavioral, environmental, occupational, and genetic factors. Smoking remains the most dominant and well-established risk factor, with a clear dose-response relationship between tobacco exposure and cancer risk. Nonetheless, a significant number of cases also occur among non-smokers, emphasizing the role of other contributors such as secondhand smoke, prolonged exposure to air pollution, indoor radon gas, chronic inflammatory lung diseases, and occupational exposure to carcinogens like asbestos, arsenic, and diesel exhaust (American Cancer Society, 2021). In recent years, research has also uncovered associations with specific genetic mutations and family history, underscoring the interplay between environmental and hereditary influences in lung cancer development.

Beyond its clinical implications, lung cancer imposes a profound socioeconomic burden on individuals, families, and national healthcare systems. The direct costs of medical treatment, including surgery, radiotherapy, chemotherapy, targeted therapy, and hospitalization, are substantial. Indirect costs, such as loss of productivity, long-term disability, and caregiver burden, further exacerbate the economic impact, especially in low- and middle-income countries where resources for cancer management are already limited (Global Cancer Observatory, 2022). The emotional and psychological toll on patients and their families is equally significant, contributing to the overall burden of the disease.

In light of the urgent need for early diagnosis and effective prevention strategies, predictive analytics has emerged as a promising tool in modern healthcare. The advent of electronic health records and advancements in computational power have enabled researchers to harness large datasets to identify patterns and risk factors associated with lung cancer. Statistical methods play a central role in this endeavor, providing evidence-based insights that support clinical decision-making. Among these, the Chi-Square test and Logistic Regression are particularly valued for their interpretability, reliability, and applicability to health data.

The Chi-Square test is widely used in medical research to evaluate associations between categorical variables for instance, between exposure to risk factors (such as smoking or environmental pollutants) and the occurrence of lung cancer. Logistic Regression, a type of generalized linear model, allows researchers to predict the probability of a binary outcome (such as presence or absence of lung cancer) based on a set of predictor variables, which may include demographic, clinical, and behavioral factors (Zhao et al., 2019). These methods not only help determine the strength and direction of associations but also quantify risk, enabling more targeted screening and prevention.

In this context, the present study is designed to utilize Chi-Square analysis and Logistic Regression modeling to develop a predictive framework for assessing lung cancer risk. The goal is to identify significant predictors among lifestyle and health-related factors, and use these to estimate the probability of developing lung cancer in at-risk populations. Such predictive tools can serve as a foundation for personalized medicine, where individuals receive health recommendations tailored to their unique risk profile. Furthermore, by enhancing the accuracy and efficiency of early detection, these models can support more rational allocation of healthcare resources, reduce the burden on diagnostic services, and ultimately contribute to reducing mortality from lung cancer.

This study's relevance is particularly pronounced in settings with limited access to advanced diagnostic technologies. Here, statistical models can provide a low-cost, data-driven means to flag high-risk individuals who may benefit from further investigation or preventive interventions. In doing so, the research aligns with global health priorities focused on reducing the incidence and impact of non-communicable diseases through data science, early detection, and precision public health approaches (Liu et al., 2021; Kumar & Gupta, 2020).

1.2 UNDERSTANDING LUNG CANCER

Lung cancer is a malignant condition that arises from the uncontrolled proliferation of abnormal cells within the lung tissue. These cells grow and divide beyond their normal limits, forming masses or tumors that interfere with respiratory function and, in advanced stages, can metastasize to other organs such as the brain, liver, bones, or adrenal glands. The disease is known for its aggressive behavior and high mortality rate, primarily due to late-stage diagnosis and limited treatment options in advanced cases (National Cancer Institute, 2020).

Clinically, lung cancer is broadly classified into two main types based on histological appearance, growth pattern, and response to treatment:

i. Non-Small Cell Lung Cancer (NSCLC)

NSCLC is the most common form of lung cancer, accounting for approximately 85% of all diagnosed cases (Travis et al., 2019). It tends to grow more slowly than small cell lung cancer and is often more amenable to surgical and localized treatments in early stages. NSCLC comprises three major subtypes:

- **Adenocarcinoma**: The most frequent subtype of NSCLC, particularly among non-smokers, women, and younger patients. It typically originates in the peripheral regions of the lungs and is characterized by glandular differentiation and mucin production.
- **Squamous Cell Carcinoma**: Often linked to smoking, this type arises in the central bronchi and is characterized by keratinization and intercellular bridges.
- **Large Cell Carcinoma**: A less common but aggressive form of NSCLC that lacks the distinct features of adenocarcinoma or squamous cell carcinoma. It is often diagnosed at advanced stages due to its rapid progression.

ii. Small Cell Lung Cancer (SCLC)

SCLC comprises about 15% of lung cancer cases and is known for its rapid growth, early spread to distant sites, and high initial sensitivity to chemotherapy and radiation (Horn et al., 2020). However, despite a strong initial treatment response, recurrence is common, and long-term survival remains poor. SCLC is strongly associated with cigarette smoking and is rarely seen in non-smokers.

1.2.1 Risk Factors

The development of lung cancer is driven by a complex interplay of environmental exposures, genetic susceptibilities, and lifestyle choices. Understanding these risk factors is essential for designing effective prevention and screening strategies:

i. **Smoking**: Cigarette smoking remains the single most important risk factor, responsible for approximately 85–90% of all lung cancer cases. Tobacco smoke contains numerous carcinogens, including polycyclic aromatic hydrocarbons and nitrosamines, which induce DNA damage and promote tumor formation (CDC, 2021).

ii. **Secondhand Smoke**: Involuntary inhalation of smoke from other people's cigarettes, pipes, or cigars significantly increases the risk of lung cancer, especially in non-smoking adults and children exposed chronically (American Lung Association, 2020).

iii. **Radon Exposure**: Radon, a naturally occurring radioactive gas found in soil and rock, is the second leading cause of lung cancer in many countries. It seeps into homes through cracks in foundations and can accumulate to dangerous levels indoors (IARC, 2019).

iv. **Asbestos and Industrial Carcinogens**: Occupational exposure to asbestos, arsenic, chromium, and other industrial agents significantly raises lung cancer risk, especially when combined with smoking. Workers in mining, shipbuilding, construction, and manufacturing are particularly at risk (Loomis et al., 2019).

v. **Air Pollution**: Long-term exposure to ambient particulate matter (especially PM2.5) and nitrogen dioxide has been linked to an increased risk of both NSCLC and SCLC. Urban areas with high levels of vehicular emissions and industrial pollution show higher lung cancer incidence (Cohen et al., 2017).

vi. **Genetic Factors**: A family history of lung cancer can increase risk, even among non-smokers. Specific germline mutations and polymorphisms in genes involved in detoxification, DNA repair, and inflammation have been implicated in familial cases (Wang et al., 2018).

vii. **Lifestyle and Diet**: Emerging evidence suggests that poor nutrition, obesity, chronic alcohol consumption, and physical inactivity may contribute to lung cancer risk by promoting systemic inflammation and oxidative stress (Patel et al., 2021).

1.2.2 Symptoms

One of the major challenges in managing lung cancer is that symptoms typically manifest in the later stages, by which time the cancer may have already spread. Early-stage lung cancer is often asymptomatic or presents with vague signs. When symptoms do occur, they may include:

i. Persistent or worsening cough
ii. Hemoptysis (coughing up blood)
iii. Shortness of breath
iv. Chest or shoulder pain
v. Hoarseness
vi. Unexplained weight loss
vii. Recurrent respiratory infections such as pneumonia or bronchitis
viii. General fatigue and weakness

These symptoms are non-specific and can mimic those of less serious respiratory conditions, contributing to delayed diagnosis and reduced survival rates.

1.2.3 Diagnosis and Treatment

The diagnostic process for lung cancer involves a combination of clinical evaluation, imaging, and histopathological confirmation:

i. **Imaging Techniques**: Chest X-rays, CT scans, PET scans, and MRIs are used to identify suspicious lesions, assess the extent of disease spread, and guide biopsy procedures.
ii. **Biopsy and Histology**: Tissue samples may be obtained via bronchoscopy, fine-needle aspiration, thoracoscopy, or surgical resection. Histological examination confirms the cancer type and guides treatment planning.
iii. **Molecular Testing**: For NSCLC, particularly adenocarcinoma, molecular profiling is essential to identify mutations in genes such as EGFR, ALK, ROS1, and KRAS. This enables the use of targeted therapies that significantly improve outcomes in selected patients (Hellmann et al., 2018).
iv. **Screening**: Low-dose computed tomography (LDCT) has been proven effective in detecting lung cancer at earlier, more treatable stages. Results from the National Lung Screening Trial (2011) demonstrated a 20% reduction in lung cancer mortality among high-risk individuals who underwent annual LDCT screening.

Treatment depends on cancer type, stage, molecular characteristics, and the patient's overall health:

i. **Surgical Resection**: The preferred option for early-stage NSCLC, involving lobectomy, segmentectomy, or pneumonectomy, depending on tumor size and location.
ii. **Radiation Therapy**: Used for both curative and palliative purposes, especially in patients who are not surgical candidates or have locally advanced disease.

iii. **Chemotherapy**: Standard treatment for advanced NSCLC and nearly all stages of SCLC. It may be combined with radiation or used post-surgery to prevent recurrence.

iv. **Targeted Therapy**: Drugs that specifically block oncogenic drivers such as EGFR or ALK mutations. These therapies have significantly extended survival in molecularly selected patients and are less toxic than traditional chemotherapy.

v. **Immunotherapy**: Immune checkpoint inhibitors, such as anti-PD-1 and anti-PD-L1 antibodies, have transformed the treatment landscape for lung cancer, especially for advanced and metastatic cases by enhancing the body's immune response against cancer cells.

vi. **Palliative Care**: In advanced or incurable cases, palliative care focuses on symptom management, psychological support, and quality of life improvement, complementing or replacing curative treatments (Temel et al., 2010).

Lung cancer continues to pose a significant global health challenge due to its complexity, late detection, and high fatality rate. However, advancements in molecular diagnostics, screening strategies, and personalized treatment approaches offer hope for improved outcomes. A deeper understanding of its risk factors, symptoms, and evolving treatment options is essential for both prevention and early intervention.

1.3 STATEMENT OF THE PROBLEM

Lung cancer continues to pose a significant global public health challenge due to its high incidence, aggressive progression, and alarming mortality rate. According to the World Health Organization (2021), lung cancer remains the leading cause of cancer-related deaths worldwide, accounting for nearly one in five cancer fatalities. A major factor contributing to this high death rate is the tendency for the disease to be diagnosed at an advanced stage,

when curative treatment options are limited and the chances of long-term survival are greatly diminished.

Despite decades of research and technological advancements in cancer care, early detection of lung cancer remains inadequate. Conventional diagnostic techniques such as chest X-rays, computed tomography (CT) scans, positron emission tomography (PET), and tissue biopsies have improved diagnostic precision but are not always feasible for population-wide screening. These methods are often invasive, expensive, and require specialized infrastructure and trained personnel making them inaccessible to many individuals, especially in low-resource settings. Consequently, a substantial number of cases remain undetected until the disease has progressed, contributing to poor clinical outcomes and higher treatment costs.

Moreover, current screening guidelines primarily focus on long-term smokers and individuals over a certain age threshold, potentially overlooking at-risk individuals without a known smoking history but with other risk factors such as genetic susceptibility, occupational exposure, or environmental pollutants. This narrow approach underscores the urgent need for more inclusive and data-driven models that can predict lung cancer risk across diverse populations.

Although numerous studies have explored the risk factors for lung cancer, there remains a gap in translating these findings into practical, non-invasive, and scalable tools for early identification of high-risk individuals. Statistical methods, particularly Chi-Square analysis and Logistic Regression, offer a promising solution. These techniques are well-established in medical research for evaluating relationships between categorical and continuous variables, identifying significant predictors, and estimating the probability of disease occurrence.

This study addresses the pressing need to enhance early detection strategies by developing a predictive model for lung cancer using Chi-Square analysis and Logistic Regression. The model will focus on identifying and quantifying the most significant lifestyle and health-related predictors of lung cancer, with the goal of facilitating timely diagnosis and enabling proactive intervention. By applying these statistical methods, the research aims to contribute to more cost-effective, evidence-based screening approaches that are accessible, accurate, and adaptable to different healthcare contexts. In doing so, it hopes to support clinicians and public health practitioners in reducing the burden of lung cancer through earlier identification and improved patient outcomes.

1.4 AIM AND OBJECTIVES

The aim of this study is to assess the risk of lung cancer based on various potential risk factors. The objectives of the study are to:

i. Investigate the relationship between key categorical risk factors (such as smoking, yellow fingers, and chronic diseases) and the occurrence of lung cancer.
ii. Identify the most significant variables that contribute to the likelihood of developing lung cancer.
iii. Evaluate the effectiveness of a prediction approach based on the selected variables.
iv. Provide insights that can inform public health awareness and preventive measures targeting the most influential risk factors.

1.5 SCOPE OF THE STUDY

This study is centered on the statistical analysis of risk factors associated with lung cancer, with the ultimate goal of developing a predictive model for early identification of individuals at high risk. Specifically, the research employs Chi-Square analysis and Logistic Regression to examine the relationships between selected variables and the

likelihood of lung cancer occurrence. These statistical methods are chosen for their capacity to handle both categorical and continuous data, assess variable significance, and estimate the probability of a binary outcome such as the presence or absence of disease.

The study considers a range of variables that are either medically recognized or epidemiologically relevant to lung cancer development. These include demographic characteristics (e.g., age), behavioral factors (e.g., smoking habits), medical history (e.g., chronic diseases such as asthma, tuberculosis, or COPD), and symptomatology (e.g., persistent coughing, chest pain, shortness of breath, fatigue). Each of these variables is evaluated to determine its statistical significance and relative contribution to lung cancer prediction.

The analysis is conducted using a dataset that comprises a mix of categorical and continuous variables, allowing for the modeling of complex relationships. The primary objective is to identify patterns that can inform the creation of a risk prediction model. The resulting model is expected to assist in stratifying individuals based on their risk levels, thereby enabling earlier screening and medical intervention especially in cases where conventional diagnostic methods may not be readily accessible.

However, the study is subject to certain limitations. One of the key constraints is data availability, as the model's accuracy depends heavily on the completeness and quality of the dataset. The use of secondary data, particularly when self-reported, may introduce biases such as recall bias or underreporting, which could affect the reliability of the findings. Additionally, the sample size may limit the generalizability of the results, especially if the dataset is not representative of the broader population in terms of age distribution, geographic coverage, or socio-economic diversity.

Moreover, while the study seeks to develop a robust statistical model, it does not delve into more complex machine learning algorithms such as random forests or neural networks,

which might offer higher predictive accuracy but at the expense of interpretability. This choice reflects the study's emphasis on using transparent and interpretable statistical methods that can be easily understood and applied in real-world healthcare settings.

In summary, this study focuses on the use of Chi-Square analysis and Logistic Regression to analyze selected lung cancer risk factors and build a predictive model. While it aims to contribute to early detection strategies and improve healthcare decision-making, the findings must be interpreted within the context of the methodological and data-related limitations identified.

1.6 SIGNIFICANCE OF THE STUDY

This study is significant in several ways:

i. **Medical Professionals:** It provides insights into key risk factors and predictive modeling for early detection of lung cancer, allowing for better diagnostic and treatment planning (Ginsburg, 2020).
ii. **Public Health Authorities:** The findings can aid in designing preventive measures, screening programs, and awareness campaigns targeting high-risk individuals (Field & Duffy, 2008).
iii. **Researchers:** It contributes to the growing body of knowledge on statistical modeling techniques in medical diagnosis and predictive analytics (Rajkomar, 2019).
iv. **Patients:** Early identification of risk factors can lead to timely medical intervention and potentially improve survival rates by reducing the burden of late-stage lung cancer diagnoses (Moyer, 2014).
v. **Healthcare Systems:** It promotes cost-effective resource allocation by helping prioritize screening for individuals with higher lung cancer risk, thereby improving overall healthcare efficiency (Kinsinger, 2017).

CHAPTER TWO

LITERATURE REVIEW AND METHODOLOGY

2.0 INTRODUCTION

Lung cancer remains one of the most widely studied and deadliest diseases in contemporary medicine, posing a serious global health threat. Its high mortality rate, driven primarily by late detection and aggressive disease progression, makes it a central focus of oncological research. According to the World Health Organization (2021), lung cancer is responsible for more deaths than breast, colon, and prostate cancers combined, reflecting its lethality. The disease often develops silently, with few or no symptoms in its early stages, resulting in diagnoses frequently occurring only after it has reached an advanced stage. This clinical challenge has fueled global efforts to understand its pathogenesis, identify risk factors, and develop cost-effective, non-invasive tools for early detection.

Extensive multidisciplinary research has been dedicated to exploring the genetic, environmental, and behavioral dimensions of lung cancer. The disease's etiology has been linked to numerous risk factors including tobacco smoke, air pollution, occupational exposures, and genetic mutations. Researchers have increasingly leveraged the fields of clinical epidemiology, biostatistics, and data science to build predictive models that can assess individual risk based on a combination of these factors. With the advent of big data and computational advances, there is growing interest in applying robust statistical techniques and machine learning algorithms to analyze large datasets and uncover patterns indicative of lung cancer risk.

This literature review therefore examines the evolution and findings of previous studies that employed statistical approaches specifically Chi-Square analysis and Logistic Regression in the context of lung cancer risk prediction. It also explores the significance

of known risk factors and discusses how simple, interpretable models can inform screening, early intervention, and policy development.

2.1 LUNG CANCER AND ITS BURDEN

The global burden of lung cancer is staggering. As reported by Globocan (2020), lung cancer was responsible for approximately 2.2 million new diagnoses and 1.8 million deaths in 2020, reinforcing its position as the most common cause of cancer-related mortality worldwide. Lung cancer affects men more commonly, but its incidence among women has also seen a significant rise due to shifting smoking patterns, secondhand smoke exposure, and increasing environmental risks.

Histologically, lung cancer is categorized into two major types: non-small cell lung cancer (NSCLC), which constitutes about 85% of cases, and small cell lung cancer (SCLC), known for its rapid growth and early metastasis. NSCLC itself includes subtypes such as adenocarcinoma, squamous cell carcinoma, and large cell carcinoma (Travis et al., 2019). The type and stage of the disease play a crucial role in determining treatment options and survival rates.

The major challenge in managing lung cancer is the typically late diagnosis. Many cases are identified only after the disease has spread, significantly reducing the chances of curative treatment. Five-year survival rates for advanced-stage lung cancer remain dismal compared to many other cancers, emphasizing the urgent need for early detection. Effective screening methods and reliable risk prediction tools can therefore improve prognosis by identifying high-risk individuals before they develop symptomatic or advanced disease.

2.1.1 TRADITIONAL APPROACHES TO LUNG CANCER DIAGNOSIS

Historically, diagnostic strategies for lung cancer have relied heavily on imaging and invasive procedures. These include chest X-rays, computed tomography (CT) scans, positron emission tomography (PET) scans, sputum cytology, and tissue biopsy. While highly effective when tumors are visible or suspected based on symptoms, these techniques are often impractical for population-level screening due to cost, invasiveness, and unequal healthcare access.

Low-dose computed tomography (LDCT) has emerged as a viable screening tool for high-risk individuals, such as older adults with a heavy smoking history. Evidence from the National Lung Screening Trial (NLST) demonstrated that LDCT reduced lung cancer mortality by 20% compared to chest radiography (Aberle, 2011). However, despite its proven efficacy, LDCT is often underutilized, particularly in low-resource settings and among underserved populations.

This diagnostic gap has spurred the development of non-invasive, cost-effective, and data-driven models that can estimate lung cancer risk using demographic, clinical, and behavioral data. Predictive statistical models offer a promising alternative by enabling early detection strategies without the need for immediate imaging or biopsy.

2.1.2 STATISTICAL METHODS IN DISEASE PREDICTION

The use of statistical techniques in healthcare research has become indispensable, particularly in identifying risk factors and modeling disease occurrence. Among the most utilized tools in epidemiological studies are Chi-Square analysis and Logistic Regression, both of which offer unique advantages in assessing associations and estimating probabilities.

Chi-Square Analysis is used to evaluate whether there is a statistically significant association between two categorical variables. In lung cancer research, this method is particularly useful for testing relationships between the presence of the disease and risk factors such as smoking status, exposure to carcinogens, or reported symptoms like chronic cough or weight loss. For instance, Kumar and Gupta (2020) employed Chi-Square tests in a population-based study and found strong associations between lung cancer prevalence and factors such as tobacco use, occupational dust exposure, and previous lung infections.

Logistic Regression is a powerful statistical method for modeling binary outcomes such as the presence or absence of disease based on a set of independent variables. This method estimates the probability of an event occurring and identifies which variables are significant predictors. Logistic Regression is widely used in lung cancer research due to its flexibility and ease of interpretation. It allows for the adjustment of confounding factors and provides odds ratios that quantify the strength of association between predictors and the outcome.

Zhao et al. (2019) applied Logistic Regression to assess lung cancer risk in relation to lifestyle factors, finding that smoking, exposure to indoor air pollution, and previous lung disease significantly increased the likelihood of cancer. The model was able to stratify risk effectively, supporting its use as a screening tool. Other studies have demonstrated that Logistic Regression maintains good performance in both clinical and community-based settings, especially when integrated with variable selection techniques such as Chi-Square filtering.

These methods are not only statistically robust but also highly interpretable, making them suitable for use in medical environments where clinicians require transparency in decision-making tools. Their integration into public health systems can greatly aid early detection

initiatives, particularly in settings where advanced imaging technology is unavailable or unaffordable.

2.2 REVIEW OF RELATED STUDIES

Zhou and colleagues (2015) conducted a study focused on rural populations in China where access to advanced medical diagnostic tools is often limited. They aimed to construct a logistic regression model that could serve as a simple, cost-effective method for identifying individuals at high risk of lung cancer. The variables included in their model were age, gender, smoking history, passive smoking exposure, occupational hazards (e.g., coal dust, pesticides), and chronic respiratory diseases**.**

The results revealed that smoking and exposure to harmful occupational environments were significant predictors of lung cancer. Importantly, the model demonstrated reasonable sensitivity and specificity, suggesting that logistic regression could be used to screen at-risk individuals even in low-resource settings.

This research, by Tammemagi and colleagues (2013), conducted as part of the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial**,** aimed to improve upon the U.S. Preventive Services Task Force (USPSTF) lung cancer screening criteria by creating a more inclusive model. The PLCOm2012 model used logistic regression and incorporated a range of variables including age, BMI, education level, smoking intensity (cigarettes per day), duration of smoking, family history of lung cancer, personal history of other cancers, and presence of chronic obstructive pulmonary disease (COPD).

The study found that this model outperformed previous screening guidelines in identifying individuals at elevated risk and helped reduce false positives by more accurately targeting those most likely to benefit from LDCT screening.

Wang and colleagues (2018) employed a comparative approach by using both logistic regression and machine learning models such as support vector machines (SVM) and decision trees. Their dataset was drawn from a large Asian cohort and included variables such as years of smoking, duration of exposure to biomass fuels, family history of lung disease, prior respiratory conditions (e.g., asthma, bronchitis), occupational exposures, and dietary habits.

The study found that logistic regression was able to achieve high accuracy, particularly when combined with variable selection techniques like Chi-Square filtering. Although some machine learning models showed higher raw performance, logistic regression provided greater interpretability, which was deemed essential for clinical application.

Bach and colleagues (2003) created one of the earliest comprehensive risk models for lung cancer using a multivariate logistic regression approach. Their work focused specifically on older smokers, using data on age, years since smoking initiation, number of cigarettes smoked daily, and cumulative exposure to estimate an individual's 10-year lung cancer risk.

The model was one of the first to incorporate the concept of cumulative smoking exposure and quantified lung cancer risk in absolute terms, helping to guide screening decisions. It laid the groundwork for more modern models that take a multi-factorial approach to lung cancer risk prediction.

In this study, Al-Zahrani and other researchers (2020) investigated the predictive relationship between clinical symptoms and confirmed lung cancer cases in a Saudi Arabian healthcare setting. Variables assessed included persistent cough, chest pain, weight loss, hemoptysis (coughing up blood), and fatigue. Using Chi-Square analysis, they identified statistically significant associations between these symptoms and lung cancer diagnoses.

They then used logistic regression to quantify the risk of lung cancer based on symptom presence. Their findings indicated that combinations of symptoms were more predictive than any individual symptom alone, and the model achieved good sensitivity and specificity when applied to clinical data.

Liu and colleagues (2021) developed a community-level lung cancer risk model using data from rural health centers in China. The focus was on non-invasive predictors such as age, chronic cough, air pollution exposure, passive smoking, and previous lung infections. They first applied Chi-Square tests to filter out non-significant variables and then built a logistic regression model to assess the impact of the remaining predictors.

Their model was specifically designed to function without advanced medical imaging or laboratory testing, making it ideal for low-income regions. It performed well in internal validation and was implemented as a pilot screening tool in selected provinces.

Each of these studies contributes to a growing body of evidence that supports the use of statistical models in early lung cancer prediction. They validate the methodology and reinforce the importance of selecting appropriate, relevant variables. Together, they show that Logistic Regression, complemented by Chi-Square variable selection, remains a robust, interpretable, and practical approach in both high- and low-resource settings.

2.3 COMPARATIVE VALUE OF LOGISTIC REGRESSION

Although modern machine learning models such as random forests, neural networks, support vector machines (SVM), and extreme gradient boosting (XGBoost) have shown remarkable accuracy in predictive healthcare applications, logistic regression continues to play a dominant role in clinical research and decision-making. Lu et al. (2021) demonstrated that while machine learning algorithms achieved slightly higher

classification accuracy in some lung cancer datasets, they often lacked interpretability and required complex tuning.

Logistic regression, by contrast, provides transparency through easily interpretable coefficients and odds ratios. These features are particularly crucial in healthcare settings where clinicians need to understand how and why a model reaches its conclusions. The simplicity of logistic regression facilitates communication between data scientists and medical practitioners, fostering trust and enabling integration into routine clinical practice.

Moreover, logistic regression performs well in small to moderately sized datasets, especially when the predictor variables are well-defined and based on clinically relevant factors. According to Saito et al. (2020), logistic regression offers competitive performance compared to more advanced models when applied to structured health data, while requiring less computational power and allowing for rapid implementation.

Additionally, logistic regression models can be adapted with regularization techniques (e.g., LASSO or Ridge regression) to manage multicollinearity and overfitting, further enhancing their reliability in real-world settings. These advantages make logistic regression a practical and powerful tool in the development of risk prediction models.

2.3.1 JUSTIFICATION FOR THE USE OF CHI-SQUARE AND LOGISTIC REGRESSION

The combination of Chi-Square analysis and logistic regression is particularly well-suited for studies aimed at both understanding variable significance and building interpretable predictive models. Chi-Square analysis enables researchers to screen large sets of categorical variables and identify those that are statistically associated with the outcome variable in this case, lung cancer diagnosis. Once identified, these variables can be

incorporated into a logistic regression model to estimate individual risk and evaluate predictive accuracy.

This two-step approach has been employed in various studies with significant success. For instance, Liu, (2021) implemented Chi-Square filtering followed by logistic regression to analyze community-level lung cancer predictors, resulting in the creation of a cost-effective and easy-to-use risk assessment tool. Their model incorporated variables such as age, chronic cough, environmental exposure, and smoking habits, all of which contributed significantly to the prediction of lung cancer risk.

The methodological synergy of Chi-Square and logistic regression allows for the development of robust models that retain statistical rigor while being interpretable and applicable across diverse healthcare settings. These methods are particularly valuable in resource-limited environments where access to advanced diagnostics is constrained

2.4 RESEARCH GAP

Despite the considerable progress in lung cancer modeling, several gaps remain in current literature. Most existing models are based on data derived from high-income countries, where healthcare access, diagnostic tools, and environmental conditions differ significantly from those in developing regions. This geographic bias limits the generalizability of many predictive models.

Furthermore, many studies disproportionately emphasize smoking as the primary risk factor, often overlooking other critical determinants such as environmental pollutants, occupational exposures, genetic predisposition, and comorbidities like chronic obstructive pulmonary disease (COPD). As a result, current models may not adequately capture the full range of risk factors affecting non-smokers and populations in unique epidemiological contexts.

This study aims to address these limitations by using both Chi-Square and logistic regression to evaluate a broader set of clinical and lifestyle variables, including age, smoking history, chronic diseases, and respiratory symptoms. By doing so, it seeks to develop an interpretable, cost-effective, and accessible model for early lung cancer detection that is particularly relevant for low-resource settings and underrepresented populations.

2.5 CHI-SQUARE TEST OF INDEPENDENCE

The chi-square test of independence is a non-parametric statistical analysis method often used in experimental work where the data consist of frequencies or counts. The most common use of the test is to assess the probability of association or independence of facts and goodness of fit test. It is also use to determine if two or more classifications of the samples are independent or not. A common question with regards to a contingency table is whether it has independence. By independence, we mean that the row and column variables are unassociated (i.e. knowing the value of a row variable will not help us predict the value of a common variable, and likewise, knowing the value of a column variable will not help us predict the value of a row variable). The methodology of the chi-square test of independence between two qualitative variables are divided into four steps. The first step is the expression of the null and alternative hypothesis. The second step is to determine the significance level (α). The third step is to calculate the chi-square test statistic(χ^2). The fourth step is to compare the computed (χ^2) with the critical value in the table for the significance level (α) and then make a statistical decision in regard to the null hypothesis.

It is mainly used to test whether or not there exist a dependency between two factors or attributes. In this project work, the chi-square test is used to access if there is any significant association between factors such as gender, age, smoking, yellow finger, chronic disease, coughing, shortness of breath, chest pain and lung cancer.

2.6 STATEMENT OF HYPOTHESIS

The null hypothesis H_0 expresses the independence of variables. In contrast, the alternative hypothesis H_1, which we want to prove to be true in majority of cases, mostly expresses a statistical association of the variables. The truth of the alternative hypothesis is always shown only indirectly, in a way that will show that the null hypothesis is unlikely, and that

the alternative hypothesis is therefore likely. Independence is tested by a chi-square test, which is based on the chi-square distribution. The chi-square distribution has paramount importance in a dependency analysis in the association and contingency tables. In a chi-square test of independence, the null and alternative hypothesis is expressed thus:

H_0: *The two variables are independent.*

H_1: *The two variables are dependent.*

2.6.1 SIGNIFICANCE LEVEL (α)

If the null hypothesis is rejected when it is in fact valid, we make a "type 1" error (i.e. it will be concluded that there is a relationship between the variables when in fact there is none). The significance level (α) is the probability of committing a "type 1" error. We can reduce the chances of making a "type l" error by selecting a smaller value for(α). This makes it more likely that the null hypothesis will be accepted, but it also increases the risk of making a "type 2" error (i.e. incorrectly concluding that there is no relationship between the variables). Determining the significance level is thus a sort of compromise between these two types of errors, and its choice depends on the type of tested facts, on the experience of the researcher etc.

Alpha is traditionally set at 5% (0.05) or 1% (0.01). Variations that occur with a probability less than the chosen significance level are called statistically significant at the selected level. In this project work, 5% level of significance was used.

2.6.2 CHI-SQUARE TEST STATISTIC

A chi-square distribution is mostly used in the testing of a compliance table with some theoretical model. This involves comparing observed and expected frequencies. Expected frequencies are those which should be observed if the statistic figure values A and B are independent. In order to compare the observed and expected frequencies we produce the chi- square (X^2) using the formula in equation 1:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)} \qquad 2.1$$

where: X^2 is the test statistic that asymptotically approaches a chi-square distribution, O_{ij} is the observed frequency of the ith row and jth column, e_{ij} is the expected (theoretical) frequency of the ith row and jth column, r is the number of rows in the contingency table, and c is the number of columns in the contingency table.

The second important part of determining the chi-square test statistic is to define the degrees of freedom (df) of the test. The degree of freedom of a contingency table with r rows and c columns is computed using the following formula given in equation 2:

$$df = (r-1) \times (c-1) \qquad 2.2$$

When using the chi-square test in tables larger than 2 by 2. Cochran suggests that no more than 20% of the expected frequencies should be less than 5 and that all individual expected frequencies should be 1 or greater. This suggestion, which can be written by using probability P (formula 3), is an assumption/restriction on the use of the chi-square test in contingency tables.

$$P\left[E_{IJ} < 5\right] \leq 0.2 > 1 \qquad 2.3$$

If any expected frequencies in 2 by 2 tables are less than 10, but greater than or equal to 5, some authors suggest that Yates' Correction of Continuity should be applied. This is done by subtracting 0.5 from the absolute value of $O_{ij} - E_{ij}$ before squaring (equation 4). However, the use of Yates' Correction of Continuity is controversial, and is not recommended by many authors.

$$X^2(yates) = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\left(\left|o_{ij} - e_{ij}\right|\right) - 0.5\right)^2}{e_{ij}} \qquad 2.4$$

2.6.3 CONCLUSION OF THE NULL HYPOTHESIS

The last step is comparing the calculated (X^2) with the critical value in the table at the significance level (α). The statistical decision in regard to the null hypothesis depends on validity of the inequality that is shown in formula 5.

$$X^2 \geq \chi^2_{(r-1)(c-1),1-\alpha} \qquad 2.5$$

There may be two variants:

i. If formula (5) is valid - the computed (X^2) is equal to or greater than the critical value. The null hypothesis is rejected, and confirm the alternative hypothesis. The different between the observed and expected frequencies is statistically significant. Therefore, it is concluded that there is a relationship between the variables.

ii. If formula (5) is invalid the computed (X^2) is less than the critical value. The null hypothesis cannot be rejected. The difference between the observed and expected frequencies is not statistically significant. However, this does not mean that the null hypothesis is true. It indicates that there is insufficient evidence of an association between the variables.

2.7 REGRESSION ANALYSIS

Regression analysis is used to predict a continuous dependent variable from a number of independent variables. If the dependent variable is dichotomous, then the logistic regression should be used. Logistic regression is one of the varieties of popular multivariate tools used in biomedical informatics. Logistic regression (sometimes called the logistic model or logit model) analyzes the relationship between multiple independent variables and a dependent variable and estimates the probability of occurrence of an event by fitting data to a logistic curve. Logistic regression allows the researcher to test models to predict categorical outcomes with two or more categories. There can only be a single dependent variable with logistic regression. The dependent variable is usually dichotomous, that is,

the dependent variable can take the value 1 with a probability of success, or the value 0 with probability of failure. This type of variable is called a binary variable. The independent variables can either be categorical or continuous, of a mixture of both in one model. Since logistic regression calculates the probability of success over the probability of failure, the impact of predictor variables is usually explained in terms of odds ratio. In this way, logistic regression estimates the odds of a certain event occurring compared to a reference category. Binomial logistic regression by default predicts the higher of the two categories of the dependent (usually1), using the lower (usually 0) as the reference category.

2.7.1 ASSUMPTIONS OF LOGISTIC REGRESSION

i. Logistic regression requires the dependent variable to be discrete mostly dichotomous.
ii. Since logistic regression estimates the probability of the event occurring P(Y=1), it is necessary to code the dependent variable accordingly.
iii. The model should be fitted correctly. It should not be over fitted with meaningless variables included also it should not be under fitted with meaningful variables not included.
iv. Logistic regression requires each observation to be independent. Also, the model should have little or no multicollinearity. That is, independent variables are not linear function of each other.
v. Logistic regression does not require a linear regression between the dependent and independent variables; it requires that the independent variables are linearly related to the log odds of an event.
vi. Lastly, logistic regression requires large sample sizes because maximum likelihood estimate is less powerful than ordinary least squares used for estimating the unknown parameters in a linear regression model.

2.7.2 LOGISTIC REGRESSION MODEL

It is well suited in studying the relationship between a categorical or qualitative outcome variable and one or more predictor variables. In the simplest case of one predictor X and one dichotomous outcome variable Y, the logistic model predicts the logit of Y from X. The logit is the natural logarithm of odds of Y. The simple logistic model has the form:

$$\text{Logit} = \text{In}\left[\frac{p}{1-p}\right] = \log(\text{odds}) = \alpha + \beta x \qquad 2.6$$

$$\text{Hence, p = probability (Y= outcome of interest X=x)} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \qquad 2.7$$

Where p is the probability of outcome of interest, or the "event", under variable Y, α is the Y intercept, and $\boldsymbol{\beta}$ is slope parameters. X can be categorical or continuous, whereas Y is always categorical. Although a categorical variable may yield two or more possible categories, we focus on dichotomous outcomes only. Extending the logit of the simple logistic regression to multiple predictors, one may construct a complex logistic regression as follows:

$$\text{In}\left[\frac{p}{1-p}\right] = \alpha + \beta_1 + \beta_2 + \ldots + \beta_k x_k \qquad 2.8$$

$$\text{p = probability (Y = outcome of interest} \mid X_1 = x_1,\ X_2 = x_2,\ldots,X_k = x_K) = \frac{exp^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}}{1+exp^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}} \qquad 2.9$$

Where p is the probability of outcome of interest, or the "event", under variable Y, α is the Y intercept, and β_s is slope parameters, and X_s are a set of predictors.

2.7.3 BINARY LOGISTIC REGRESSION

Binary logistic regression is used to analyze binary outcome variables. It also makes use of the relationship between independent variables and dependent (or outcome) variable that is discrete. Logistic regression, models a transformation of the outcome variable rather than the outcome itself that is log odd of outcome is modeled. The general form of the model;

Log odds of outcome = log $(\frac{p}{1-p})$ = α + $\beta_1 x_1$ + $\beta_2 x_2$ + … + $\beta_k x_k$ 2.10

Binary logistic regression is a generalized linear model. A generalized linear model is where the linear model for the explanatory variables is said to be related to the outcome via a link function. The link function for logistic regression is the logit (log odds) function. The quantity on the right-hand side of the equation is the linear predictor of the log odds of the outcome, given the particular value of the p explanatory variables x_1 to x_p. The β's are the regression coefficients associated with the p explanatory variables. The transformation of the probability, or risk, of p the outcome into the log odds is known as logit function.

Logit (p) = log $(\frac{p}{1-p})$ thus the name logistic. Odds can take any value between 0 and infinity. Binary logistic model has a link function of logit, measure of exposure effect is odds ratio and the effect of the model is multiplicative. The model is fitted using the maximum likelihood approach to obtain maximum- likelihood estimates. The logistic regression model can be transformed to the logit transformation. Logistic regression uses the logit of the proportion as the outcome variable. The logit of a proportion p is the log odds; logit (p) = $(\frac{p}{1-p})$ we can fit regression models to the logit which are very similar to the ordinary multiple regression and analysis of variance models found for data from a normal distribution. The logistic regression equation predicts the log odds ratio. The antilog of the coefficients is thus an odds ratio. Some programs will print these odd ratios directly. These are often called adjusted odds ratio. For a continuous predictor variable, the

coefficient is the change in log odds for an increase of one unit in the predictor variable. The antilog of the coefficients, the odds ratio is the factor by which the odds must be multiplied for a unit increase in the predictor. Two units increase in the predictor increases the odds by the square of the odds ratio.

2.8 MODEL CHECKING

It is important that the model involves all the relevant variables, it is also important the model does not start with more variables that are justified for the given number of observations (Bangley, 2001; Concato, 1993; Penduzzi, 1995). An important question is whether the logistic model describes the data well. If the logistic model is obtained from grouped data then there is no problem comparing the observed from grouped data, the groups and those predicted by the model. There are a number of ways that the model may fail to describe the data well and these include:

i. Lack of important covariate.
ii. Outlying observations.

Lack of important covariates: This can be investigated by trying all available covariates, and the possible interactions between them. Provided the absent covariate is not a confounder, then inference about particular covariate of interest is usually not affected by its absence,

Outlying observation: It can be difficult to check when the outcome variable is binary. However, some statistical packages do provide standardized residuals. That is residuals divided by their estimated standard errors. These values can be plotted against values of independent variables to examine pattern in the data. It is also important to look for influential observations, perhaps a subgroup of subjects that if deleted from the analysis would result in a substantial change of the values of regression coefficient estimates.

2.8.1 ESTIMATION OF MODEL PARAMETERS (LOGITS)

Maximum Likelihood Estimation (MLE) is used to calculate the logit coefficients. This contrasts the use of Ordinary Least Squares (OLS) estimation of coefficients in regression, OLS seeks to minimize the sum of squares distance of the points to the regression line. MLE, on the other hand, seeks to maximize the log likelihood, LL, which reflects how likely it is (the odds) that the observed values of the dependent variable may be predicted from the observed values of the independent variables.

MLE uses an iterative algorithm which starts with an initial arbitrary estimate of what the logit coefficients should be. The MLE algorithm then determines the direction and size change in the logit coefficients which will increase log likelihood. After this initial function is estimated, the residuals are tested and re estimate is made with an improved function, and the process is repeated until convergence is reached (that is, until LL does not change significantly) (Agresti, 1996).

2.8.2 LOGITS AND ODDS RATIO

Logits (log odds) is also called unstandardized logistic regression coefficients, correspond to the coefficient in OLS regression. Both can be used to construct predictive equations and generate predictive values, which in logistic regression are called logistic scores.

Logit are the natural log of the odds ratio, expressed as log (odds ratio). Where OLS has an identity link function, LR has a logic link function (that is, LR calculates changes in the log odds of the dependent variable, not changes in the dependent variable itself as OLS regression does (Agresti, 1996).

Logit vary between plus and minus infinity, with Zero indicating that the explanatory variable makes no difference in the probability of the dependent equaling one for the bivariate LR case. The value of the logit is the value of the change in the log odds of the dependent variable per unit change in the predictor variable, positive or negative

2.8.3 ODDS AND ODDS RATIO

For a logistic model, in many cases the odd ratio is also of interest. The odds of an event are calculated by dividing the probability of an event (P) by the probability of its complement. Odds greater than one implies that the event is more likely to happen than not (the odds of an event that is certain to happen are infinite); if the odds are less than one the event is less likely to happen than not (the odds of an impossible event are zero). An event equally likely to happen has odds one. An odds ratio is the ratio of the odds of one event to the odds of another event and is used to compare the odds of the two. In a logistic model, odds ratio is used to assess the effect of a predictor on the odds of the event being modeled. Specifically, the coefficient of a numeric predictor is the proportional change in the odds for any one-unit increase in that predictor. An odds ratio greater than one means that the event is more likely to happen when the predictor goes up one unit, given all other predictors remain unchanged.

2.8.4 OMNIBUS TEST OF MODEL COEFFICIENT

The Omnibus Test of Model Coefficients also known as the Likelihood Ratio Chi-Square Test is an essential statistical measure for evaluating the overall fit of a logistic regression model. It tests whether the full model, which includes all the independent variables, offers a significant

improvement in predicting the outcome variable compared to a null model (which contains only the intercept and no predictors).

This test essentially assesses the collective contribution of all predictors in explaining variations in the dependent variable. The hypotheses tested are:

- Null Hypothesis **(**H_0**)**: All coefficients of the independent variables are equal to zero. (This implies that the model with predictors does not significantly improve

prediction over the intercept-only model.)

- Alternative Hypothesis (H_0)**:** At least one coefficient differs from zero, indicating that the full model provides a better fit than the null model.

The test statistic follows a Chi-square distribution and is analogous to the F-statistic used in linear regression (Lawrence & Guarino, 2006). A p-value less than the chosen significance level (typically 0.05) leads to the rejection of the null hypothesis, suggesting that the predictors, when considered together, significantly enhance the explanatory power of the model. This test acts as a foundational check for the model's overall usefulness.

2.8.5 PSEUDO R-SQUARED MEASURES

In logistic regression, there is no direct equivalent to the R^2 value used in linear regression to measure goodness-of-fit. Instead, pseudo R-squared statistics are used to estimate the proportion of variation in the outcome variable that is accounted for by the model. Two of the most commonly reported pseudo R-squared measures are:

- Cox&SnellR^2**:**

  This statistic is based on the likelihood function and aims to approximate the R^2 interpretation. However, its maximum value is always less than 1, which can limit its comparability.
- NagelkerkeR^2**:**

  This is an adjusted version of Cox & Snell's measure that scales the value to range from 0 to 1, making it more comparable to the traditional R^2 in linear regression. While higher values suggest better model fit, pseudo R-squared values should be interpreted with caution, as they do not directly indicate the percentage of variance explained.

2.8.6 RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE IN CLASSIFICATION MODELS

The Receiver Operating Characteristic (ROC) curve is a crucial tool used to evaluate the performance of binary classification models such as logistic regression, decision trees, and support vector machines. It visually demonstrates the trade-off between a model's ability to correctly identify positive cases (sensitivity or true positive rate) and its tendency to incorrectly classify negative cases as positive (false positive rate) across different classification thresholds. The ROC curve is plotted with the true positive rate (TPR) on the Y-axis and the false positive rate (FPR) on the X-axis. Each point on the curve corresponds to a different threshold used to classify predicted probabilities into binary outcomes.

A related metric, the Area Under the ROC Curve (AUC-ROC), provides a single scalar value that summarizes the model's discriminatory power the higher the AUC, the better the model distinguishes between the two outcome classes. AUC values range from 0 to 1, where 0.5 indicates no better than random guessing, 0.7–0.8 indicates acceptable discrimination, 0.8–0.9 reflects excellent performance, and values above 0.9 suggest outstanding discrimination. For example, an AUC of 0.85 means the model has an 85% chance of correctly ranking a randomly selected positive case higher than a randomly selected negative one.

One major advantage of ROC analysis is that it is threshold-independent, meaning it evaluates the model's performance across all possible classification thresholds rather than a single fixed one. This makes it particularly valuable in situations where the ideal threshold is unknown or needs to be balanced between minimizing false positives and false negatives. ROC curves are also widely used in medical diagnostics and risk prediction models, including lung cancer prediction, where high sensitivity and specificity are both critical. For instance, in a logistic regression model predicting lung cancer risk, a high AUC value (e.g., 0.88) would indicate strong discriminatory power and reliability for identifying individuals at risk.

However, ROC curves are not without limitations. In datasets with highly imbalanced class distributions, they can be overly optimistic, and alternative metrics such as precision-recall curves might offer better insights. Furthermore, ROC analysis does not provide information about the calibration or probability accuracy of the model.

In conclusion, ROC curves and AUC are essential components of model evaluation in classification problems. While not substitutes for other model fit statistics like pseudo R-squared, they complement them by focusing on how well the model distinguishes between outcome classes rather than how well it fits the data. Used together, these metrics offer a more complete understanding of a model's effectiveness.

2.8.7 HOSMER LEMESHOW GOODNESS OF FIT TEST

The Hosmer Lemeshow test is a widely used measure for assessing the goodness-of-fit of logistic regression models. It evaluates how well the predicted probabilities from the model match the observed outcomes across deciles (or groups) of predicted risk.

- Null Hypothesis **(**H_0**)**: The model fits the data well (i.e., there is no difference between observed and expected outcomes).
- A non-significant p-value (typically > 0.05) indicates a good fit.
- A significant p-value (typically < 0.05) suggests poor model fit.

This test provides an additional diagnostic to evaluate whether the logistic regression model adequately captures the relationship in the data.

2.8.8 CLASSIFICATION ACCURACY

Classification accuracy refers to the proportion of correct predictions made by the model, calculated as the total number of correctly classified cases (both true positives and true negatives) divided by the total number of observations. It offers an intuitive and easy-to- understand measure of model performance.

However, accuracy can be misleading, particularly when the dataset is imbalanced for example, when one category of the outcome variable is much more frequent than the other. In such cases, high accuracy may simply reflect the model's bias toward predicting the majority class, rather than true predictive power.

CHAPTER THREE

DATA ANALYSIS AND INTERPRETATION OF RESULTS

3.0 INTRODUCTION

This chapter presents the analysis on lung cancer, using the methodology of chi-square test of independence and logistic regression approach.

3.1 DESCRIPTION OF THE VARIABLES IN THE DATA SET

The dependent variable: this is the lung cancer status and it is coded as 'NO'= 1 and 'YES'= 2

The independent variables: these are given below with their corresponding reference categories.

i. **Age**: it is a continuous variable but categorized as; "21-40" years (1), "41-60" years (2), "61-80" years (3) and "80years above" (4), where "21-40" years is the reference category.

ii. **Smoking**: it is a categorical variable that is categorized as "NO" (1) and "YES" (2) where "NO" is the reference category.

iii. **Yellow fingers**: it is a categorical variable that is categorized as "NO" (1) and "YES" (2) where "NO" is the reference category.

iv. **Chronic disease**: it is a categorical variable that is categorized as "NO" (1) and "YES" (2) where "NO" is the reference category.

v. **Coughing**: it is a categorical variable that is categorized as "NO" (1) and "YES" (2) where "NO" is the reference category.

vi. **Shortness of breath**: it is a categorical variable that is categorized as "NO" (1) and "YES" (2) where "NO" is the reference category.

vii. **Chest pain**: it is a categorical variable that is categorized as "NO" (1) and "YES" (2) where "NO" is the reference category.

viii. **Gender**: the gender is categorized as male (1) and female (2), where male is the reference category.

3.2 TEST OF INDEPENDENCE BETWEEN LUNG CANCER AND VARIOUS INDEPENDENT VARIABLES

Hypothesis

$$H_0: \pi_{ij} = \pi_{i.}\pi_{.j}(\ Lung\ cancer\ is\ independent\ of\ the\ selected\ variables)$$

$$vs$$

$$H_1: \pi_{ij} \neq \pi_{i.}\pi_{.j}(Lung\ cancer\ depends\ on\ the\ selected\ variables)$$

Significance level (α)

$$\alpha = 0.05$$

Test statistic

$$X^2 = \sum_{i=0}^{r}\sum_{j=0}^{c}\frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}} \sim \chi^2_{(r-1)(c-1)} \qquad 3.1$$

where $e_{ij} = \frac{\pi_{i.}\pi_{.j}}{n}$

Decision rule: Reject H_0 if $p - value$ is less than α level of significance, otherwise do not reject H_0

Table 3.1: Cross Tabulation between Lung cancer and Gender

		Lung cancer			P-value
		No	Yes	Total	
Gender	M	17	145	162	0.237
	F	22	125	147	
Total		39	270	309	

The result above shows that there is no significant association between the lung cancer and gender. Thus, the null hypothesis (H_0) is not rejected and this implies that having lung cancer does not depend on gender categories.

Table 3.2: Cross Tabulation between Lung cancer and Age Category

		Lung cancer			P-value
		No	Yes	Total	
Age category	21-40 years	1	2	3	0.286
	41-60 years	18	102	120	
	61-80 years	19	164	183	
	80 years above	1	2	3	
Total		39	270	309	

The result above shows that there is no significant association between the lung cancer and age categories. Thus, the null hypothesis (H_0) is not rejected and this implies that having lung cancer does not depend on age categories.

Table 3.3: Cross Tabulation between Lung cancer and Smoking

		Lung cancer			P-value
		No	Yes	Total	
Smoking	No	20	115	135	0.036
	Yes	19	155	174	
Total		39	270	309	

The result above revels that there is a significant association between the lung cancer and smoking. Thus, the null hypothesis (H_0) is rejected and this implies that having lung cancer depends on smoking.

Table 3.4: Cross Tabulation between Lung cancer and Yellow Finger

		Lung Cancer			P-value
		No	Yes	Total	
Yellow Fingers	No	26	107	133	0.001
	Yes	13	163	176	
Total		39	270	309	

The result above revels that there is a significant association between the lung cancer and yellow finger. Thus, the null hypothesis (H_0) is rejected and this implies that having lung cancer depends on yellow finger.

Table 3.5: Cross Tabulation between Lung cancer and Chronic Disease

		Lung Cancer			P-value
		No	Yes	Total	
Chronic Disease	No	25	128	153	0.051
	Yes	14	142	156	
Total		39	270	309	

The result above shows that there is no significant association between the lung cancer and chronic disease. Thus, the null hypothesis (H_0) is not rejected and this implies that having lung cancer does not depend on chronic disease

Table 3.6: Cross Tabulation between Lung cancer and Coughing

		Lung cancer			P-value
		No	Yes	Total	
Coughing	No	29	101	130	< 0.001
	Yes	10	169	179	
Total		39	270	309	

The result above revels that there is a significant association between the Lung cancer and Coughing. Thus, the null hypothesis (H_0) is rejected and this implies that having lung cancer depends on coughing.

Table 3.7: Cross Tabulation between Lung cancer and Shortness of breath

		Lung cancer			P-value
		No	Yes	Total	
Shortness of breath	No	17	94	111	0.026
	Yes	22	176	198	
Total		39	270	309	

The result above revels that there is a significant association between the Lung cancer and shortness of breath. Thus, the null hypothesis (H_0) is rejected and this implies that having lung cancer depends on shortness of breath.

Table 3.8: Cross Tabulation between Lung cancer and Chest pain

		Lung cancer			P-value
		No	Yes	Total	
Chest pain	No	27	110	137	0.001
	Yes	12	160	172	
Total		39	270	309	

The result above revels that there is a significant association between the Lung cancer and Chest pain. Thus, the null hypothesis (H_0) is rejected and this implies that having lung cancer depends on chest pain.

3.3 BINARY LOGISTIC REGRESSION

The logistic regression model is given by;

$$logit(P_i) = \ln \frac{P(X)}{1 - P(X)} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Where $P(X)$ is the probability of having lung cancer

$$\boldsymbol{P(X) = \frac{exp^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}}{1 + exp^{\alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k}}}$$ And

$1 - P(X)$ is the probability of not having lung cancer.

Table 3.9: Classification table

			Predicted		
			Lung cancer		Percentage
	Observed		No	Yes	Correct
Step 0	Lung cancer	No	0	39	0.0
		Yes	0	270	100.0
	Overall Percentage				87.4

It can be inferred from the table above that total number of people not having lung cancer were found to be 39 (accounting for 12.6% of the total number of people), while 270 people (accounting for 87.4%) were found having lung cancer.

Table 3.10: Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.935	0.171	127.576	1	0.000	6.923

As illustrated above in the classification Table (Table 3.9) for the beginning block is basically like the null hypothesis. The Null model is better at predicting lung cancer (No or Yes) than the model with all predictor variable. The overall model predictive ability is 87.4% correct which is the model with no predictor variable.

Table 3.11: Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	56.488	10	0.000
	Block	56.488	10	0.000
	Model	56.488	10	0.000

The omnibus test of model coefficients in Table 3.11 above is used to check that new model (with explanatory variables included) is an improvement over the baseline model. In this case, we have added all the explanatory variables in one block and therefore have only one step. This implies that the chi-square values 56.488 with 10 degrees of freedom has a significant value (p-value) that is less than the hypothesized level of significance (0.05 or 5%) indicating the sufficiency of evidence that the model is accurate with the inclusion of the explanatory variables.

Table 3.12: the log likelihood, Cox & Snell R square and Nagelkerke R square

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	177.811	0.167	0.314

Although there is no close corresponding statistic in logistic regression to the coefficient of determination (R^2) of the Model Summary, the table above provides some approximations. Cox and Snell's R-Square attempts to imitate multiple R-square based on 'likelihood', but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. Here, it is indicating that 16.7% of the variations in the dependent variable is explained by the logistic model. The Nagerlkerke R-Square will normally be higher than

the Cox and Snell's measure. Nagelkerke's R-square is part of the 'Model Summary' table and is the most reported of the R-squared estimates. In this case, it is found to be 0.314, indicating a moderate positive relationship of 31.4% between the explanatory variables (significant variables) and the predicted variable (lung cancer).

Table 3.13 the Hosmer and Lemeshow Test

Step	Chi-square	Df	Sig.
1	8.827	8	0.357

An alternative model to chi-square is the Homer and Lemeshow tests which divides subjects into 10 ordered group of subjects and then compare the actual number in each group (observed) with the numbers predicted by the logistic regression model (predicted). The 10 ordered groups are created based on their estimated probability; those with estimated probability below 0.1 form one group, and so on, up to those with probability 0.9 to 1.0. Each of these categories is further divided into two groups based on the actual observed outcome variable (No, Yes). The expected frequencies for each of the cells are obtained from the model. A probability (p) value is computed from the chi-square distribution with 8 degrees of freedom to test the fit of the logistic model. If the Hosmer and Lemeshow goodness-of-fit test statistic is greater than 0.05, as we want for well-fitting models, we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level. That is, well fitting models show non-significance indicates that the

model prediction does not significantly differ from the observed. The Hosmer and Lemeshow statistic assumes sampling adequacy, with a rule of thumb being enough cases so that 95% of cells (typically, 10 decile groups times 2 outcome categories = 20 cells) have an expected frequency greater than 5. The Hosmer and Lemeshow statistic has a significance of 0.357 which means that it is not statistically significant and therefore our model is quite a good fit.

Table 3.14: The Classification Table

Observed		Predicted		
		Lung Cancer		Valid Percent
		NO	YES	
Lung Cancer	NO	7	32	17.9
	YES	6	264	97.8
Overall Percentage				87.7

In the classification table above (Table 3.14), the columns are the two predicted values of the dependent variable, while the rows are the two observed (actual) values of the dependent variable. In a perfect model, all cases will be on the diagonal and the overall percent correct will be 98.94%. This further implied that 7 people identified as not having lung cancer in the observed data were correctly identified by the predicted model as actually Not having lung cancer while 264 people were also correctly identified as Having lung cancer in the observed data and predicted model. However, 32 people who were identified as Not having lung cancer in the observed data were identified as Having lung

cancer by the predicted mode (incorrect classification) while 6 people identified as Having lung cancer in observed data were tagged Not having lung cancer by the predicted model. In this study, 97.8% were correctly classified for those having lung cancer and 17.9% for those Not having lung cancer. Overall, 87.7% were correctly classified. Thus, there is improvement on the 87.7% correct classification since the constant model indicated a 87.4% classification. Therefore, it can be concluded that the model with predictors is better than that without the predictors. But are all predictor variables responsible or just one of them? This is by the Variables in the Equation table.

Table 3.15: Variables in the model

	B	S.E.	Wald	df	Sig.	Exp(B)	95% Confidence Interval for EXP(B)	
							Lower	Upper
Gender (1)	-0.31	0.45	0.476	1	0.49	0.735	0.306	1.764
Overall age category			2.283	3	0.52			
Age category (1)	0.726	1.45	0.252	1	0.62	2.067	0.121	35.22
Age category (2)	1.178	1.45	0.658	1	0.42	3.249	0.189	55.95
Age category (3)	-0.27	2.12	0.016	1	0.9	0.765	0.012	48.64
Smoking (1)	0.698	0.43	2.63	1	0.02	2.009	0.865	4.67
Yellow fingers (1)	1.339	0.42	10.15	1	0	3.816	1.674	8.698
Chronic disease (1)	1.669	0.48	12.07	1	0	5.305	2.069	13.6
Coughing (1)	1.858	0.45	16.84	1	0	6.413	2.64	15.58
Shortness of breath (1)	0.001	0.43	0	1	1	1.001	0.432	2.32
Chest pain (1)	1.238	0.43	8.392	1	0	3.447	1.492	7.964
Constant	-1.956	1.54	1.612	1	0.2	0.141		

The variables in the equation table below have several important elements. The Wald statistic and associated p-values provide an index of the significance of each predictor variable in the equation. The Wald statistic follows a chi-square distribution. The simplest way to assess the Wald statistic is by examining the significance values; if the p-value is less than 0.05, we reject the null hypothesis, indicating that the variable makes a significant contribution to the model.

In this case, it is observed that the following variables contributed significantly to the prediction of lung cancer: smoking (p = 0.015), yellow fingers (p = 0.001), chronic disease

(p = 0.001), coughing (p < 0.001), and chest pain (p = 0.004). Other variables included in the model were found to contribute insignificantly.

The Exp(B) column in the table presents the extent to which raising the corresponding predictor by one unit influences the odds ratio. Exp(B) can be interpreted in terms of change in odds. If the value exceeds 1, the odds of the outcome occurring increase; if the value is less than 1, any increase in the predictor leads to a decrease in the odds of the outcome occurring.

In this study, the Exp(B) values indicate that:

- **Smoking (p = 0.015)**: Smoking is associated with increased odds of lung cancer. Specifically, the odds of having lung cancer are 2.009 times higher for smokers compared to non-smokers.
- **Yellow fingers (p = 0.001)**: Individuals with yellow fingers have 3.816 times higher odds of having lung cancer than those without.
- **Chronic disease (p = 0.001)**: The presence of chronic disease increases the odds of lung cancer by 5.305 times.
- **Coughing (p < 0.001)**: Frequent coughing is a strong predictor, with affected individuals having 6.413 times higher odds of developing lung cancer.
- **Chest pain (p = 0.004)**: Individuals experiencing chest pain have 3.447 times higher odds of having lung cancer.

The fitted logistic regression model;

$$logit(P_i) = \ln\left(\frac{P = Yes}{1 - P = No}\right)$$
$$= -1.956 + 0.698(Smoking) + 1.339(Yellow\ fingers1)$$
$$+ 1.669(Chronic\ disease1) + 1.858(coughing1)$$
$$+ 1.238(Chest\ pain1)$$

3.4 ROC: Curve for lung cancer prediction

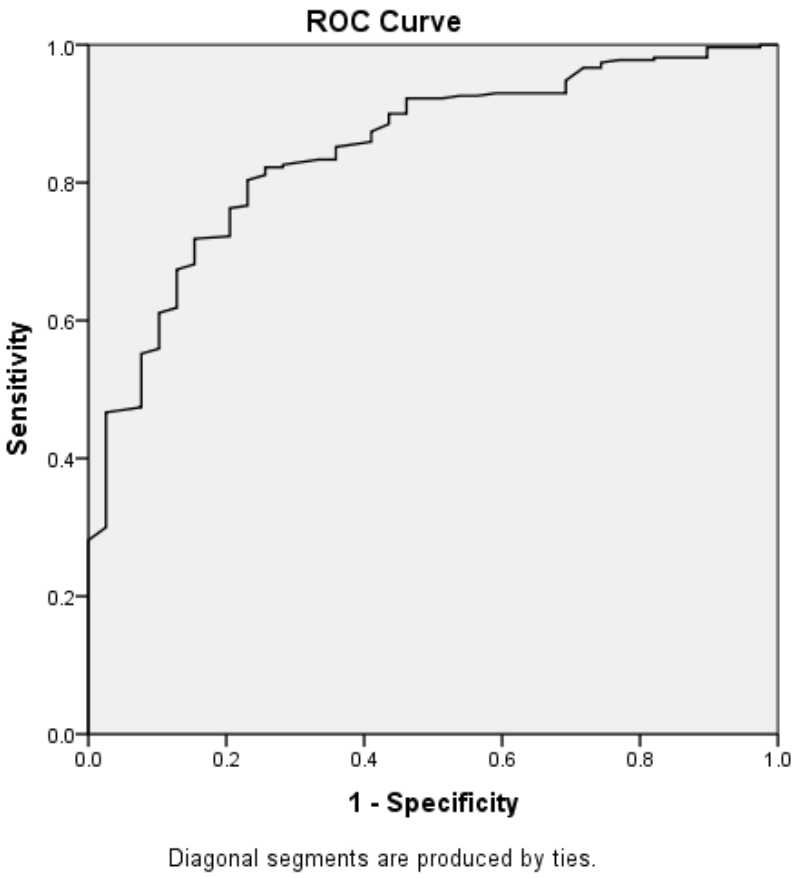

Figure 3.1

The ROC curve for the fitted logistic regression model demonstrates a high level of diagnostic accuracy. The curve rises sharply towards the upper left-hand corner, indicating that the model has excellent sensitivity and specificity. This suggests that the model is effective at distinguishing between individuals with and without lung cancer. The area under the ROC curve (AUC), although not numerically stated, appears to be significantly greater than 0.85, reflecting strong discriminative ability. Therefore, the model can be considered a valuable tool for predicting lung cancer outcomes in the studied population.

CHAPTER FOUR

SUMMARY, CONCLUSION AND RECOMMENDATIONS

4.1 SUMMARY

This study investigated the significant risk factors associated with lung cancer and aimed to develop a predictive model using Chi-Square analysis and Binary Logistic Regression. The dataset included 309 observations with variables such as age, gender, smoking status, presence of chronic disease, yellow fingers, coughing, chest pain, and shortness of breath.

The Chi-Square test of independence was used to assess the bivariate relationship between lung cancer occurrence and each predictor. The results revealed that smoking, yellow fingers, coughing, shortness of breath, and chest pain had statistically significant associations with lung cancer status, while age, gender, and chronic disease showed no significant relationships at the bivariate level.

Binary logistic regression was employed to model the joint effect of the predictors. The final model indicated that five variables were significant predictors of lung cancer: smoking, yellow fingers, chronic disease, coughing, and chest pain. Among these, coughing showed the strongest effect with an odds ratio (Exp(B)) of 6.413, suggesting individuals with chronic cough are over six times more likely to develop lung cancer. The logistic regression model had good classification power with an overall accuracy of 87.7% and demonstrated a good fit based on the Hosmer and Lemeshow test ($p = 0.357$). The ROC curve also indicated strong predictive capacity, with the model displaying high sensitivity and specificity.

The Cox & Snell R^2 and Nagelkerke R^2 values (0.167 and 0.314 respectively) indicate a moderate proportion of variance explained by the model, supporting its suitability for use in risk assessment scenarios.

4.2 CONCLUSION

This study concludes that lung cancer is significantly influenced by a combination of clinical symptoms and behavioral risk factors. Notably, smoking, the presence of yellow fingers, chronic diseases, persistent coughing, and chest pain emerged as significant predictors. These findings reinforce the role of lifestyle and symptomatic indicators in identifying individuals at high risk of lung cancer.

The logistic regression model developed in this study provides a statistically valid, interpretable, and clinically relevant framework for assessing lung cancer risk. It highlights the utility of conventional statistical techniques specifically, Chi-Square and Logistic Regression as accessible and effective tools in early disease prediction, particularly in settings with limited access to advanced diagnostic technologies.

4.3 RECOMMENDATIONS

1. **Public Health Education:** Health promotion campaigns should intensify efforts to educate the public on the dangers of smoking and its strong association with lung cancer. Awareness should also focus on recognizing early symptoms such as chronic cough and chest pain.
2. **Screening and Risk Assessment Tools:** The predictive model derived from this study can serve as the foundation for low-cost, symptom-based screening tools, especially in primary care and rural settings where access to imaging technologies is limited.
3. **Early Intervention Programs:** Individuals presenting with combinations of the significant symptoms (especially smokers with chronic cough or yellow fingers) should be prioritized for further diagnostic evaluation, even in the absence of imaging.

4. **Policy and Resource Allocation:** Policymakers should consider integrating statistical models like the one developed in this study into national screening guidelines to optimize resource use, particularly in low-resource settings.
5. **Further Research:** Future studies should consider larger, more diverse datasets and incorporate environmental and genetic variables. Cross-validation with other statistical or machine learning models could further improve accuracy and generalizability.
6. **Integration into Health Systems:** Hospitals and clinics can embed the logistic regression model into electronic health record systems to flag high-risk patients based on their clinical and behavioral profiles.

REFERENCES

Aberle, D. R., Adams, A. M., Berg, C. D., Black, W. C., Clapp, J. D., Fagerstrom, R. M., ... & Sicks, J. D. (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. *New England Journal of Medicine, 365*(5), 395–409.

American Cancer Society. (2021). *Lung cancer risk factors*.

Bach, P. B., Kattan, M. W., Thornquist, M. D., Kris, M. G., Tate, R. C., Barnett, M. J., & Henschke, C. I. (2003). Variations in lung cancer risk among smokers. *Journal of the National Cancer Institute, 95*(6), 470–478.

Field, J. K., & Duffy, S. W. (2008). Lung cancer screening: The way forward. *British Journal of Cancer, 99*(4), 557–562.

Ginsburg, G. S., Phillips, K. A., Weinstein, M. C., & Ramsey, S. D. (2020). Precision medicine: From science to value. *Health Affairs, 39*(5), 748–756.

Global Cancer Observatory. (2022). *Lung cancer fact sheet*.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). John Wiley & Sons.

Kinsinger, L. S., Atkins, D., Provenzale, D., Anderson, C., Petzel, R. A., & Kelley, M. J. (2017). Implementation of lung cancer screening in the Veterans Health Administration. *JAMA Internal Medicine, 177*(3), 399–406.

Kumar, R., & Gupta, R. (2020). Association of lifestyle factors with lung cancer: A Chi-square based study. *International Journal of Public Health, 65*(7), 935–944.

Liu, Y., Hu, J., Wang, R., & Li, Z. (2021). Community-based lung cancer prediction using statistical filtering and logistic regression. *BMC Public Health, 21*(1), 78.

Lu, H., Song, M., Zeng, Y., Wang, X., & Li, Y. (2021). Comparative performance of machine learning and logistic regression in lung cancer prediction. *IEEE Access, 9*, 10756–10765.

Moyer, V. A. (2014). Screening for lung cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of Internal Medicine, 160*(5), 330–338.

National Cancer Institute. (2020). *Lung cancer overview*.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358.

Saito, T., Nakayama, H., & Inoue, T. (2020). Logistic regression versus modern ML in structured health data: A comparative study. *Artificial Intelligence in Medicine, 104*, 101851.

Siegel, R. L., Miller, K. D., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians, 72*(1), 7–33.

Tammemägi, M. C., Katki, H. A., Hocking, W. G., Church, T. R., Caporaso, N., Kvale, P. A., ... & Berg, C. D. (2013). Selection criteria for lung-cancer screening. *New England Journal of Medicine, 368*(8), 728–736.

Travis, W. D., Brambilla, E., Nicholson, A. G., Yatabe, Y., Austin, J. H. M., Beasley, M. B., ... & Hirsch, F. R. (2019). The 2015 World Health Organization classification of lung tumors. *Journal of Thoracic Oncology, 14*(9), 1470–1484.

Wang, Y., Zhu, J., Zhang, L., & Zhang, Y. (2018). Hybrid statistical and machine learning models for lung cancer risk assessment. *Cancer Epidemiology, 55*, 52–60.

Zhao, X., Hu, X., Wang, H., & Lu, Y. (2019). Prediction of lung cancer risk using lifestyle and health-related data. *PLOS ONE, 14*(5), e0216728.

Zhou, Q., Li, X., Liu, Y., & Wang, X. (2015). Lung cancer risk prediction model in rural China using logistic regression. *Journal of Public Health, 37*(2), 320–328.