Lung Cancer Incidence Trends in South Africa from 2011 to 2022 and Investigation of Non-Small Cell Lung Cancer Biomarkers Using Rule-Based Natural Language Processing

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

by

Matshediso Ivy Mohlala

(972315)

A dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Science in Epidemiology (Public Health Informatics)

Supervisors:

Dr. Mazvita Muchengeti

Dr. Okechinyere Achilonu

Ms. Carole Metekoua

October 2025

Declaration

I, Matshediso Ivy Mohlala, hereby declare that this research report is entirely my own work. This research report is submitted in partial fulfilment of the requirements for the degree of Master of Science in Epidemiology (Public Health Informatics) at the University of the Witwatersrand, Johannesburg. Wherever the work of others has been used, it is appropriately acknowledged through citations and references.

I further declare that this dissertation has not been previously submitted for a degree at this or any other institution and that it is free from any form of plagiarism.

Signature: 

Date: October 2025

[**Matshediso Mohlala]**

Dedication

This is dedicated to my beloved children, Masechaba, Bontle, and Leago Mohlala. Your smiles, innocent laughter, and your love kept me going. To my mother, Margaret Mohlala, your love, sacrifices, teachings, and constant guidance have shaped my academic journey. To my dearest sister, Dipuo Mohlala, your support and willingness to take care of my children during moments when my studies consumed me have been a great source of strength. To my husband, Odirile Sepato, your constant support, understanding, and encouragement have been invaluable in keeping me focused on this academic path.

In the loving memory of my late grandmother, Mary Matlala, and my late father, Mputukane Mohlala. Although you are no longer here, the lessons you taught me and the cherished memories we shared continue to guide me. I will always carry you in my heart.

Presentations from the study

1. Lung cancer incidence trends in South Africa: 2011-2022: Poster Presentation at the World Congress Epidemiology held in Cape Town on 24-27 September 2024

2. Racial disparities in incident Lung Cancer in South Africa: Poster Presentation at International Association of Cancer Registries held in Beijing China on 05-07 November 2024

Abstract

Background: Lung cancer is the leading cancer in men and second in women worldwide. Biomarkers play a critical role in lung cancer diagnosis, prognosis, and treatment, guiding therapeutic decisions and enabling personalised medicine. This study aimed to evaluate lung cancer trends from 2011 to 2022 using data from the pathology-based National Cancer Registry of South Africa (NCR-SA). Additionally, the study employed rule-based Natural Language Processing (NLP) techniques to automate the extraction of biomarkers from pathology reports and analysed their prevalence rates.

Methods: We conducted a cross-sectional study; using pathologically confirmed lung cancer data from NCR-SA. Using the Segi world standard population and mid-year population estimates from Statistics South Africa (StatsSA), we calculated the Age-Standardised Incidence Rates (ASIR). We computed annual percentage change (APC) and 95% confidence intervals (CI) to analyse lung cancer trends using Joinpoint regression. Rule-based Natural Language Processing (NLP) was used to extract biomarkers from pathology reports of Non-Small Cell Lung Cancer (NSCLC) cases. The algorithm's performance was evaluated using precision, recall, and F1 score on manually annotated pathology reports. The prevalence of these biomarkers was analysed.

Results: Lung cancer remains a major health concern in South Africa (SA), with 54,405 cases diagnosed between 2011 and 2022, comprising 46,661 NSCLC cases (85.8%) and 4,607 SCLC cases (8.5%). Males accounted for 61.7% of cases. White patients (48.0%) had the highest proportion, followed by Black (28.0%), Coloureds (18.0%), and Indian/Asian (4.0%) patients. Private hospitals diagnosed 64.0% of cases. Lung cancer incidence increased by 2.1%, with notable declines in 2017–2020, followed by a rise in 2020–2022. Among 46,160 NSCLC cases, adenocarcinoma (58.4%) was the most common, followed by squamous cell carcinoma (23.6%), unclassified NSCLC (12.5%), and large cell carcinoma (5.5%). Biomarker extraction showed high precision, with 22 biomarkers identified, including 14 diagnostics and 8 predictive biomarkers. Diagnostic biomarkers appeared in 92.2% of reports, while predictive biomarkers were reported in only 7.8%, reflecting gaps in molecular testing. TTF1 (28.0%), CK7 (22.0%), and Napsin A (11.4%) were the most tested diagnostic biomarkers. Predictive biomarkers, including PD-L1 (1.8%) and EGFR (1.1%), had lower testing rates, particularly in public hospitals, where EGFR and PD-L1 testing were nearly absent compared to private hospitals.

Conclusion: Lung cancer incidence in SA declined from 2015 to 2020, possibly due to smoking regulations. However, racial disparities persist, with Whites having the highest lung cancer incidence. These disparities likely reflect differences in diagnostic access, varying smoking prevalence rates among racial groups, and socioeconomic differences in the populations that use private healthcare. Diagnostic biomarkers are widely used in clinical practice, whereas predictive biomarkers have lower positivity rates, indicating limited availability. Expanding biomarker testing is essential for early detection, treatment optimisation, and broader adoption of personalised immunotherapy in SA.

Acknowledgements

First and foremost, I give thanks to God for His guidance, strength, and blessings throughout this journey. His grace has sustained me, giving me the perseverance and determination to reach this stage. I express my sincere gratitude to my supervisors, Dr. Mazvita Muchengeti, Dr. Okechinyere Achilonou, and Ms. Carole Metekuoa, for their guidance, encouragement, and support. Dr. Muchengeti, your mentorship has played a significant role in shaping my academic growth. Your belief in me, along with your guidance and support, has provided valuable opportunities to expand my knowledge in cancer epidemiology. Your impact on my journey has been profound, and I am truly grateful. Dr. Achilonou, your support, quick responses, and willingness to assist at every step made a world of difference. You were always just a call away, ready to offer guidance whenever I needed it. Your dedication and kindness have been invaluable, and I deeply appreciate it. Ms. Metekuoa, your patience, encouragement, and constant support kept me going throughout this journey. Your presence as a dependable source of help and reassurance meant more than words can express, and I am truly thankful.

I also extend my heartfelt thanks to the National Cancer Registry for granting me the time and resources to pursue my studies. I am also grateful for the opportunity to present my work at both international and local conferences, which has allowed me to engage with experts, gain new insights, and contribute to discussions on cancer research.

My appreciation also goes to the Wits School of Public Health for providing a strong academic foundation and an engaging learning environment. The diverse range of courses challenged my thinking, broadened my understanding of public health, and prepared me for the complexities of this field.

Contents

List of Figures

Tables

Abbreviations

APC	Annual Percent Change
ASIR	Age-Standardised Incidence Rate
LDCT	Low Dose Computed Tomography
SCC	Squamous Cell Carcinoma
LCC	Large Cell Carcinoma
NSCLC	Non-Small Cell Lung Cancer
SCLC	Small Cell Lung Cancer
TB	Tuberculosis
NCR-SA	National Cancer Registry of South Africa
NHLS	National Health Laboratory Services
NICD	National Institute for Communicable Diseases
NLP	Natural Language Processing
SA	South Africa
U. S	United States
STATS SA	Statistics South Africa
WHO	World Health Organization
SES	Socioeconomic Status
WSP	World Standard Population
Napsin A	Napsin A Aspartic Peptidase
TTF1	Thyroid Transcription Factor-1
CK7	Cytokeratin 7
CK20	Cytokeratin 20

CD56	Neural Cell Adhesion Molecule 1
CK5/6	Cytokeratin 5/6
P63	Tumour Protein P63
CEA	Carcinoembryonic Antigen
P40	Np63 Isoform (P40)
BerEP4	Epithelial-Specific Antigen
GATA3	GATA Binding Protein 3
BRAF	B-Raf Proto-Oncogene Serine/Threonine Kinase
PD-L1	Programmed Death-Ligand 1
EGFR	Epidermal Growth Factor Receptor
MET	MET Proto-Oncogene, Receptor Tyrosine Kinase
ALK	Anaplastic Lymphoma Kinase
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha
ROS1	ROS Proto-Oncogene 1, Receptor Tyrosine Kinase

Introduction

This chapter provides the background of the study and defines its overall objectives. It introduces the research problem, outlines the study's aims, and its justification. It discusses lung cancer incidence, risk factors, the importance of screening and awareness, and available treatment options both globally and in South Africa. Additionally, the chapter highlights the role of biomarkers in NSCLC diagnosis and treatment.

1.1 Background

1.1.1 Lung Cancer Classifications

Lung cancer originates within the lung tissue or airways [1]. It is classified into two primary types: Small Cell Lung Cancer (SCLC) and Non-Small Cell Lung Cancer (NSCLC). SCLC accounts for roughly 15% of lung cancer diagnoses, whereas NSCLC represents approximately 85% of all cases [2]. NSCLC is further classified into three major histological subtypes: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Among these, adenocarcinoma is the most frequently diagnosed type, contributing to about 40% of lung cancer cases. Squamous cell carcinoma and large cell carcinoma make up 25%-30% and 5%-10% of cases, respectively [3-4]. SCLC is an aggressive form of lung cancer known for its rapid growth and early spread to other parts of the body [5-6]. Unlike SCLC, NSCLC generally grows more slowly and is less likely to have early metastasis [3]. While it is important to understand the characteristics of lung cancer, exploring its associated risk factors will show how environmental and individual factors contribute to its development.

1.1.2 The Aetiology of Lung Cancer

Various factors contribute to the development of lung cancer, with tobacco smoking being the primary risk factor and responsible for about 90% of cases [7] and responsible for about 90% of cases because tobacco smoke contains over 70 carcinogens that cause DNA mutations in lung cells, resulting in uncontrolled cell growth [8]. Non-smokers can still develop lung cancer due to exposure to second-hand smoke, family history of cancer, genetic predisposition, exposure to occupational hazards like asbestos, previous lung diseases, lifestyle choices such as poor diet and physical inactivity, and prolonged exposure to radon gas [9]. In non-smokers, lung cancer accounts for

approximately 15–25% of all lung cancer cases [10]. Exposure to carcinogens in the workplace is associated with 5-10% of lung cancer cases worldwide, with asbestos making individuals 3.6 times more likely to develop the disease [7, 11]. Radon exposure is another significant risk factor for lung cancer, particularly among non-smokers. It is the second biggest risk factor in smokers, and the biggest in non-smokers [10]. Having a family history of lung cancer can increase the risk of developing the disease by 1.7 times. The risk is increased by approximately 50% among first-degree relatives [12-13]. Exposure to second-hand smoke can increase the likelihood of developing lung cancer by 20-24%. Of this percentage, it is estimated that 16-24% of cases occur in people who have never smoked [14]. Recognising the significance of these risk factors highlights the importance of advocacy in raising awareness about lung cancer and the importance of screening for early detection.

1.1.3 Lung Cancer Advocacy and Screening

Given that most lung cancer cases are associated with behavioural and environmental factors, it is important to raise awareness among various communities regarding the risks associated with these factors. Such awareness can be essential in preventing the development and late presentation of lung cancer in the future. Lung cancer is usually diagnosed during its advanced stages, with a survival rate of 5% [2, 15]. Late diagnosis of lung cancer is often the result of limited knowledge of symptoms and limited access to healthcare services [16-17]. As a result, many people present with advanced lung cancer [18]. However, public health interventions and awareness campaigns have been successful in increasing awareness and promoting early detection. These initiatives have led to an increase in chest X-rays and a decrease in the number of people presenting with advanced lung cancer [19-20]. In addition to raising awareness and understanding the impact of behavioural and environmental factors on lung cancer, it is important to implement efficient screening initiatives to help in early detection. Low-dose computed tomography (LDCT) has been shown to be a better method for detecting peripheral lung cancer in people with a history of tobacco use compared to chest radiography and is able to identify lung cancer at earlier stages when it is more treatable [21-22]. Using LDCT to screen individuals who are at risk for lung cancer has been proven to improve overall lung cancer prognosis and lower the mortality rate of lung cancer by 20% [23-24]. With the

importance of early detection in improving lung cancer prognosis, there is a growing need for effective treatment options and the use of biomarkers to personalise therapies for better patient outcomes.

1.1.4 Treatment Approaches

The primary conventional approaches for lung cancer treatment include surgical procedures, radiation therapy, and chemotherapy [25]. These procedures have limitations; surgery struggles with complete tumour removal, chemotherapy encounters drug resistance and side effects, and radiation therapy has accuracy and accessibility limitations [26-27]. Variations in the epidemiology of lung cancer and genetic diversity among patients result in differences in the specific treatments administered to individuals [26]. Individuals diagnosed with early-stage lung cancer typically receive surgical treatment, whereas those with advanced cancer are treated with chemotherapy or radiation [28].

Traditional treatment methods often lack specificity, as most chemotherapy drugs impact both cancerous and healthy tissues, leading to undesirable side effects [29]. However, advancements in lung cancer research have led to the discovery of molecular drivers [26]. These drivers are measurable indicators that reflect the biological processes of a patient's tumour or signal responses to treatment [30]. They are classified into three types: prognostic markers, which provide information about disease outcomes; predictive markers, which help identify individuals more likely to respond to particular treatments; and diagnostic markers, which differentiate between various lung cancer subtypes [30]. These molecular markers can potentially enhance early detection and improve patient management in NSCLC [30-31].

1.1.5 Problem Statement

Lung cancer is the leading cause of cancer-related deaths in SA, with individuals diagnosed with NSCLC having a survival rate of less than 10% [32]. This highlights the need to bridge existing gaps in understanding, management, and early detection strategies for NSCLC. The World Health Organization (WHO) recommends annual LDCT screening for individuals aged 55-74 with a history of smoking, including both current and former smokers [32]. However, this guideline does not account for non-smokers or individuals who develop lung cancer due to other risk factors and primarily focuses on early detection.

Once lung cancer is diagnosed, timely treatment is essential. Biomarkers found in pathology reports can help guide treatment decisions, ensuring more targeted and effective therapies. Automating the extraction of this information is necessary for improving efficiency in research settings. Currently, there is limited literature on lung cancer trends in SA, and even less data on lung cancer biomarker prevalence and distribution patterns in African populations. While biomarker testing has become standard practice in high-income countries, the implementation, usage patterns, and clinical outcomes associated with biomarker testing remain poorly characterised across the African continent. Understanding these trends and how they have evolved over time is important for gaining a clearer picture of the disease and guiding future healthcare planning and interventions.

1.1.6 Justification

With the development of a lung cancer policy in SA, it is important to examine lung cancer incidence trends over the years and provide an overview of the disease burden. However, this task is limited by challenges of underreporting within South Africa's pathology-based National Cancer Registry (NCR-SA). Because the NCR-SA depends largely on pathology reports, cases diagnosed clinically without histological confirmation, as well as cases from private laboratories that did not consistently report in earlier years, were often missed. This reliance on pathology alone has been shown to underestimate cancer incidence in SA, including for lung cancer, particularly when compared with active population-based surveillance methods [33-34].

The use of rule-based NLP for text extraction can be a valuable tool for obtaining the data needed for this study. Currently, no research has been conducted on applying rule-based NLP to extract NSCLC biomarkers from unstructured pathology reports in SA and based on the literature review done, similar NLP applications for NSCLC biomarker extraction have not been documented elsewhere on the African continent. Investigating the effectiveness of these methods within the South African healthcare context can help bridge the existing knowledge gap. The rule-based NLP approach is not the most preferred due to its labour-intensive nature and limited reusability. However, it delivers reliable results and remains a valuable option for extracting information from complex structured templates [34-35].

This approach has a proven track record of reliability in healthcare diagnostics, helping to address concerns about trust in artificial intelligence models. By incorporating human expertise into rule-based models, physicians can have greater confidence in the accuracy and dependability of the extracted information [35]. Incorporating rule-based NLP techniques into our research will enable the extraction and application of valuable information from pathology reports. This study has the potential to inform evidence-based policymaking and serve as a foundation for future research.

1.1.7 Research Question

What are the trends in pathology-diagnosed lung cancer in South Africa from 2011 to 2022, and how can a rule-based NLP algorithm be developed and validated to extract NSCLC biomarker information from pathology reports for estimating biomarker prevalence?

1.1.8 Aim

This study primarily aimed to analyse pathology-diagnosed lung cancer trends in South Africa from 2011 to 2022. As secondary aims, the study also aimed to develop and validate a rule-based natural language processing algorithm to extract NSCLC biomarker information from unstructured pathology reports and use the resulting data to estimate the prevalence of commonly tested NSCLC biomarkers.

1.1.8 Objectives

1.1.8.1 Primary Objective

i. To describe pathology-diagnosed lung cancer trends in South Africa over the study period (2011–2022).

1.1.8.2 Secondary Objectives

i. To develop and evaluate the performance of a rule-based NLP algorithm for extracting NSCLC biomarkers from the South African National Cancer Registry's pathology reports.
ii. To estimate the prevalence of extracted NSCLC biomarkers in the South African population.

2. Literature Review

2.1 Lung Cancer Epidemiology

2.1.1 Distribution of Lung Cancer Incidence

Lung cancer ranks among the most diagnosed cancers worldwide, contributing to approximately 12.4% of all cancer cases. The estimated age-standardised incidence rate (ASIR) is 23.6 per 100,000 persons globally [2, 37]. The highest incidence of lung cancer is observed in Asia, accounting for 58.5% of global cases, followed by Europe at 22.4% and North America at 12.1% [31]. In contrast, the incidence is lowest in Western and Central Africa [39]. Among men, lung cancer remains the most diagnosed cancer [39], with ASIR of 32.1 and 16.2 per 100, 000 in females [37]. It is the leading cause of cancer-related mortality, responsible for around 1.8 million deaths annually, thereby accounting for 18.7% of all cancer related deaths. The age-standardised mortality rate is estimated at 16.8 per 100,000 [3, 37]. Although lung cancer is widely diagnosed, its incidence varies significantly across different regions.

2.1.2 Lung Cancer Mortality

In developed countries, lung cancer mortality rates have declined in recent years, mainly due to reduced smoking rates. However, the tobacco epidemic persists in low and middle-income countries, eading to a consistent rise in lung cancer incidence [40]. Compared to developed countries, emerging economies such as Brazil, Russia, India, China, and South Africa (SA) have lower lung cancer incidence but higher mortality rates [41]. This disparity may be due to unequal access to healthcare, delayed diagnosis and treatment, environmental pollution, and socio-cultural barriers [42]. Studies conducted in the U.S indicate that lung cancer mortality is highest in black males compared to other racial groups [2]. A similar trend is observed in Africa, where males have a higher lung cancer incidence than females across all regions [43]. Overall, Africa's lung cancer incidence and mortality rates are relatively low compared to the rest of the world. This could be due to limited reliable cancer registries, which may underestimate the continent's true lung cancer burden [42-43].

2.1.3 Lung Cancer underestimation

Limited epidemiology reporting further contributes to the perception that Africa's lung cancer rates are lower than other regions [42]. The lower rates may also be explained by the fact that lung cancer and pulmonary tuberculosis (TB) share common symptoms such as chronic cough, weight loss, and haemoptysis. They also present similar radiological features, including masses, nodules, and cavitary lesions, which can lead to misdiagnosis or delayed diagnosis of lung cancer in TB-endemic regions like sub-Saharan Africa [44]. A study in Zimbabwe found that 53% of lung cancer patients had been previously misdiagnosed with TB, leading to a median diagnostic delay of over eight months, with 77% diagnosed at stage IV disease [45]. Given these diagnostic challenges, the implementation of reliable biomarkers for lung cancer detection has become increasingly important to improve early diagnosis and patient outcomes.

2.2 The Role of Biomarkers in Lung Cancer

In the early stages of NSCLC, patients often remain asymptomatic [46]. Therefore, making regular screening essential for timely diagnosis, appropriate treatment, and effective disease management. Over the past decade, lung cancer treatment has improved significantly with the introduction of immunotherapy and targeted drugs like tyrosine kinase inhibitors [46]. While these treatments have greatly advanced NSCLC management, access remains a major challenge in low- and middle-income countries due to limited molecular testing, high costs, and availability issues [46]. The lack of participation in clinical trials and inadequate funding further limit awareness and access to these therapies [46].

The role of biomarkers is to provide measurable information about a patient's tumour and, in some cases, predict how the body may respond to treatment [30]. Diagnostic biomarkers help identify or confirm the presence of a disease and can distinguish between subtypes of the same condition [30]. Predictive biomarkers determine which patients are most likely to benefit from a specific therapy, while prognostic biomarkers provide information on the likely course of the disease, including the chances of recurrence or progression[30].Using biomarkers for lung cancer diagnosis can help improve early detection and provide a more affordable and effective approach to treatment.

Approximately 50-60% of NSCLC patients have at least one biomarker [47-48]. The figure below (*Figure 1*) summarises the different types of lung cancer biomarkers. As understanding of the molecular biology of lung cancer has advanced, researchers have identified several promising biomarkers. These discoveries have made it possible to target specific patient groups who can benefit from personalised molecular therapies [2, 42]. This has significantly improved lung cancer treatment, with targeted therapies and emerging immunotherapy strategies proving to be highly effective [49]. For NSCLC, specific genetic changes such as Epidermal growth factor receptor (EGFR) mutations, Anaplastic lymphoma kinase (ALK) gene rearrangements, and other oncogenic driver mutations like Kirsten rat sarcoma viral oncogene homolog (KRAS) and B-Raf proto-oncogene serine/threonine kinase (BRAF) serve as predictive biomarkers to guide treatment decisions [30]. The presence of these biomarkers in NSCLC patients helps inform treatment choices and enables personalised therapy [30].

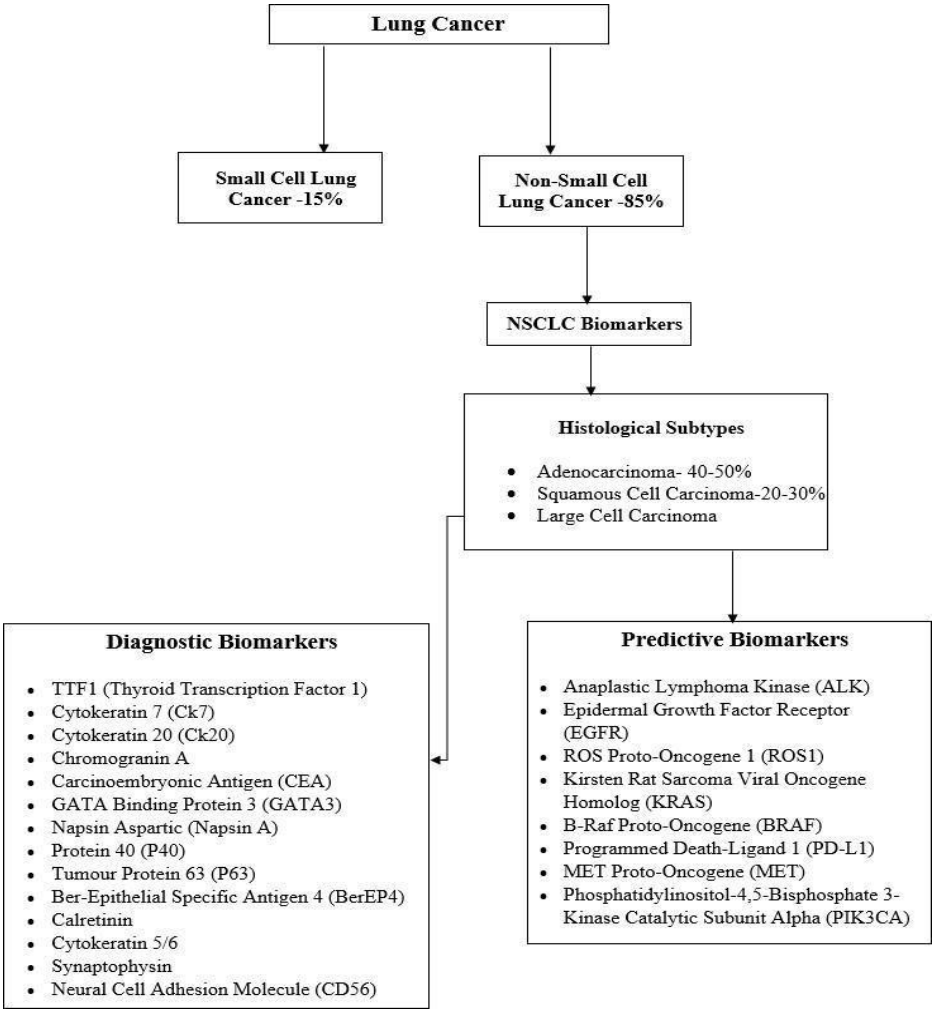


***Figure 1**: Lung cancer flow diagram and NSCLC biomarkers*

2.3 Advances in NSCLC Treatment

A range of biomarkers and chromosomal rearrangements found in NSCLC patients can now be targeted with U.S. Food and Drug Administration approved therapies [50]. In individuals with metastatic NSCLC, targeted therapies have led to better outcomes and lower toxicity compared to conventional treatment [50]. Effective treatment of advanced-stage NSCLC depends on the histological subtype, presence of targeted mutations, and the patient's clinical condition and comorbidities [4, 51]. The primary treatment for non-metastatic NSCLC typically includes surgical resection, along with adjuvant or neoadjuvant chemotherapy. For inoperable cases, immunotherapy is also a treatment option for stage III NSCLC [51]. Although targeted therapies for NSCLC are available, low- to middle-income countries continue to face major challenges in ensuring access to quality healthcare for patients. Understanding the specific burden and epidemiology of lung cancer in these regions is therefore crucial for addressing these treatment gaps.

2.4 Lung Cancer Statistics in South Africa

Lung cancer is the seventh most common cancer among females in SA, with an ASIR of 3.19 per 100,000. The incidence peaks in the 75-79 age group at 30.27 per 100,000 [43]. Among males, lung cancer is the third most common cancer, with an ASIR of 7.24 per 100,000, more than twice the rate observed in females. The highest incidence is observed in the 75-79 age group at 59.27 per 100,000 [52]. Compared to international data, its incidence in SA may be underreported [52-53] due to limitations in SA's pathology-based cancer registry system, which primarily relies on pathologically confirmed diagnoses and may not capture cases that are diagnosed clinically or radiologically [34].

It is the leading cause of cancer-related deaths in the country, with mortality rates higher than those in other African countries [43, 53]. Mortality rates due to lung cancer varied significantly between men and women in SA. Between 1995 and 2008, lung cancer accounted for 61,418 deaths, with men representing 70.5% of these cases [55]. During this period, mortality rates among men showed a declining trend, potentially reflecting improvements in early detection methods or reductions in smoking prevalence. In contrast, mortality rates among women increased by 26.4% throughout the same time period, a pattern possibly attributable to rising tobacco use among the female population [55].

2.5 Disparities in Biomarker Testing and Cancer Treatment in SA

In SA, diagnostic biomarker testing is available in private and public healthcare sectors. However, since treatment for patients with oncogenic drivers is not funded in the public sector, access to these therapies remains limited, reducing the practical benefit of testing in this setting [53]. In addition to the treatments available in the public sector, the private sector offers specialized treatments, advanced pathological examinations, and improved radiotherapy techniques that are not accessible in public healthcare facilities [56]. As a result, patients with access to private healthcare are more likely to receive better treatment than those in the public sector. SA’s healthcare system is divided into two main sectors: the private sector, which serves approximately 15.5% of the population, primarily those with medical aid, and the public sector, which caters to the remaining 84.5%.

The public healthcare system is often under-resourced and lacks the specialized treatments and advanced services available in the private sector [56]. Relying solely on biomarker testing data from

the private sector for research is insufficient, as it represents only 15.5% of the population. To gain a better understanding of disease patterns and treatment needs, it is important to include data from public hospitals. This will offer a more accurate picture of biomarker prevalence and assist policymakers in developing treatment strategies that address the needs of the broader population. However, specific data comparing biomarker-testing rates between public and private facilities is not available in published SA literature.

This data gap limits our understanding of biomarker patterns across the population and represents an important research need for developing effective precision medicine policies. This information is stored as unstructured medical text in electronic health records [58] but is not consistently extracted or used in research to guide treatment decisions in the public sector due to its labour-intensive nature [59]. Given these challenges with extracting biomarker information from unstructured medical records, computational approaches such as Natural Language Processing (NLP) offer potential solutions for mining this valuable clinical data.

2.6 Rule-based Natural Language Processing in Medical Research

Natural Language Processing broadly refers to the creation and application of computational algorithms designed to enable the interpretation and processing of human language by computers [60-61]. Within this field, Rule-based NLP refers to a structured approach that uses a set of predefined linguistic rules specifically created to extract information from unstructured text [62].

These manually created rules serve as instructions for the computer system to identify and process specific patterns and structures within natural language [62]. This approach has found significant application in healthcare, particularly in medical research. Data collection from clinical records is significant to medical research studies, and the systematic extraction of information from these documents is essential in cancer research [63-64]. Pathology reports are important in cancer research because they contain valuable information [58]. However, extracting textual data from these reports is often a time-consuming manual process. While some developed countries have transitioned to structured reporting formats, in many settings, including our NHLS pathology reports, remain in free-text format, making extraction difficult and labour-intensive. Automated approaches, such as NLP

algorithms, can address this challenge, particularly in resource-constrained environments where structured reporting systems have not yet been implemented.

These algorithms are designed to offer a potential alternative to costly and labour-intensive manual processing [58]. The effectiveness of this approach depends largely on how well the pre-defined rules can adapt to different contexts [65]. This approach can achieve high accuracy and effectiveness in tasks with clearly defined rules. Similar studies using automated NLP methods have reported high precision and recall. For example, Chen et al. achieved 95–97% precision with 97% sensitivity in extracting breast cancer data [58], while Hammami et al. reported 98% precision and sensitivity in cancer morphology classification [35]. In SA, NLP applications have likewise demonstrated strong performance, achieving 83–100% precision in extracting breast cancer parameters [66] and precision of 1.00 with 98–100% recall in extracting prostate cancer Gleason scores from unstructured pathology reports [67]. However, it may face challenges when dealing with more complex and variable tasks [64].

3 Submissible Draft Manuscript

Introduction

This section presents a draft of the manuscript prepared in accordance with the BioMed Central

Cancer journal guidelines. The background, methods, and results included here address primary objective 2: "To describe pathology-diagnosed lung cancer incidence trends in South Africa over the study period."

Manuscript Title: Racial disparities in incident Lung Cancer in South Africa 2011-2022

M. Mohlala (1, 2); O. Achilonu (2); M. Muchengeti (1,2,3); C. Metekoua (1,4);

[1]National Cancer Registry, National Institute for Communicable Diseases, Johannesburg, South Africa;

[2]University of the Witwatersrand, School of Public Health, Johannesburg, South Africa;

[3]DSI-NRF Centre of Excellence in Epidemiological Modelling and Analysis, Stellenbosch University, Stellenbosch, South Africa, [4]Graduate School for Health Sciences, University of Bern, Bern, Switzerland

Corresponding author

Name: Matshediso Mohlala

Email Address: matshediso.mohlala@gmail.com

3.1 Abstract

Background: Lung cancer is the leading cancer in men and second in women worldwide. Biomarkers play a critical role in lung cancer diagnosis, prognosis, and treatment, guiding therapeutic decisions and enabling personalised medicine. This study aimed to evaluate lung cancer trends from 2011 to 2022 using data from the pathology-based National Cancer Registry of South Africa (NCR-SA). Additionally, the study employed rule-based Natural Language Processing (NLP) techniques to automate the extraction of biomarkers from pathology reports and analysed their prevalence rates.

Methods: We conducted a cross-sectional study; using pathologically confirmed lung cancer data from NCR-SA. Using the Segi world standard population and mid-year population estimates from Statistics South Africa (StatsSA), we calculated the Age-Standardised Incidence Rates (ASIR). We computed annual percentage change (APC) and 95% confidence intervals (CI) to analyse lung cancer trends using Joinpoint regression. Rule-based Natural Language Processing (NLP) was used to extract biomarkers from pathology reports of Non-Small Cell Lung Cancer (NSCLC) cases. The algorithm's performance was evaluated using precision, recall, and F1 score on manually annotated pathology reports. The prevalence of these biomarkers was analysed.

Results: Lung cancer remains a major health concern in South Africa (SA), with 54,405 cases diagnosed between 2011 and 2022, comprising 46,661 NSCLC cases (85.8%) and 4,607 SCLC cases (8.5%). Males accounted for 61.7% of cases. White patients (48.0%) had the highest proportion, followed by Black (28.0%), Coloureds (18.0%), and Indian/Asian (4.0%) patients. Private hospitals diagnosed 64.0% of cases. Lung cancer incidence increased by 2.1%, with notable declines in 2017–2020, followed by a rise in 2020–2022. Among 46,160 NSCLC cases, adenocarcinoma (58.4%) was the most common, followed by squamous cell carcinoma (23.6%), unclassified NSCLC (12.5%), and large cell carcinoma (5.5%). Biomarker extraction showed high precision, with 22 biomarkers identified, including 14 diagnostics and 8 predictive biomarkers. Diagnostic biomarkers appeared in 92.2% of reports, while predictive biomarkers were reported in only 7.8%, reflecting gaps in molecular testing. TTF1 (28.0%), CK7 (22.0%), and Napsin A (11.4%) were the most tested diagnostic biomarkers. Predictive biomarkers, including PD-L1 (1.8%) and

EGFR (1.1%), had lower testing rates, particularly in public hospitals, where EGFR and PD-L1 testing were nearly absent compared to private hospitals.

Conclusion: Lung cancer incidence in SA declined from 2015 to 2020, possibly due to smoking regulations. However, racial disparities persist, with Whites having the highest lung cancer incidence. These disparities likely reflect differences in diagnostic access, varying smoking prevalence rates among racial groups, and socioeconomic differences in the populations that use private healthcare. Diagnostic biomarkers are widely used in clinical practice, whereas predictive biomarkers have lower positivity rates, indicating limited availability. Expanding biomarker testing is essential for early detection, treatment optimisation, and broader adoption of personalised immunotherapy in SA.

Keywords: Lung cancer, incidence, Age-Standardised Incidence Rate, racial disparities, smoking prevalence, pathology-based registry, South Africa.

3.2 Background

Lung cancer is the leading cause of cancer death globally, accounting for 18.7% of all cancer-related deaths [1]. It is the most common cancer in males and the second most common in females, with age-standardised incidence rates (ASIR) of 32.1 and 16.2 per 100,000, respectively [1]. The age-standardised mortality rate is higher in males (24.8 per 100,000) than in females (9.8 per 100,000) [1]. In South Africa (SA), lung cancer ranks as the second most common cancer in men and third in women, with 2022 ASIRs of 27.6 and 11.2 per 100,000, respectively [2] Tobacco smoking remains the leading risk factor, contributing to 90% of cases [3]. Global smoking prevalence is higher in males (32.6%) than females [4-5], contributing to higher male incidence rates [6]. Among sub-Saharan African countries, SA, Sierra Leone, and Zambia report the highest smoking prevalence [7]. To address this, SA implemented the Tobacco Products Control Act in 1993. Despite this, lung cancer disparities persist, particularly in detection and survival, which remain worse for Black populations compared to their White counterparts, likely due to multiple health risks at diagnosis, such as poorer access to healthcare, delayed care-seeking, and unhealthy lifestyle behaviours. [8-9]. However, evidence suggests that early detection can mitigate survival disparities across racial groups [10].

Socioeconomic status (SES) also influences outcomes, with higher survival rates observed in affluent areas [11]. In addition to racial disparities, it has been shown that SES impacts on survival in that those residing in higher SES areas, have improved survival rates compared to those in the lowest SES areas [9]. As an upper-middle-income country with diverse racial groups and urban-rural divides, SA faces significant healthcare disparities [10-12]. The public healthcare sector serves 84.5% of the population but is under-resourced compared to the private sector, [13] leading to advanced-stage diagnoses in public-sector patients [10]. Despite the critical role of early detection in improving survival, SA lacks formal lung cancer screening programmes [14]. Consequently, survival rates remain below 10%. The World Health Organization recommends Low Dose Computed Tomography screening for high-risk individuals aged 55–74, [14] however, SA has not implemented this. Even if adopted, the recommendation excludes non-smokers and those at risk from other factors. While one study analysed lung cancer mortality trends in SA from 1995 to 2008

[15], no studies have investigated lung cancer incidence trends in the country, highlighting a significant knowledge gap. This study aimed to address this by evaluating annual percentage changes in lung cancer incidence from 2011 to 2022 using pathology-based data from the National Cancer Registry of South Africa (NCR-SA). The findings will inform policymakers in resource allocation for lung cancer prevention, diagnosis, and management.

3.3 Methodology

3.3.1 Study setting and design

We conducted a retrospective cross-sectional study of individuals pathologically diagnosed with lung cancer from 2011 to 2022. The National Health Laboratory Service (NHLS) is a network of laboratories that provides diagnostic services to 80% of the South African population [16]. Demographic and diagnostic information for individuals diagnosed with lung cancer was routinely collected from all public hospitals across SA by NHLS and consolidated into a centralized data warehouse. Pathology reports indicating cancer malignancies were extracted from the warehouse and transferred to the NCR-SA cancer coding application. Data from private hospitals was submitted directly to the NCR-SA and integrated into the same coding application. Coders at the NCR-SA used the International Classification of Diseases for Oncology, Third Edition (ICD-O-3), to code the pathology reports. Each report was coded for topography (C34), representing the origin of the malignancy and morphologies, identifying specific lung cancer types.

3.3.2 Study population

Our study included all individuals diagnosed with lung cancer between 2011 and 2022 from both public and private hospitals, excluding cases clinically diagnosed or suspected cancers. Demographic and clinical characteristics of individuals were extracted from the NCR-SA database. The race variable was inconsistently collected in the laboratories; therefore, multiple imputation was performed to address the missing data and enhance its completeness for analysis, the methods are detailed here [17]. Mid-year population estimates, including information on race, sex, age-groups, and years from 2011 to 2022, were sourced from Statistics South Africa (StatsSA), the primary custodian of population data in the country.

3.3.3 Data management

Cases diagnosed between 2011 and 2022 with a topography code of C34, indicating cancers originating in the lung, were included in the analysis. Data including diagnostic year, cancer type, ICD-03 codes, age, sex, race, and laboratory type was obtained from the NCR-SA database. Lung cancer cases were identified using specified ICD-O-3 morphology codes. The dataset was further classified into 16 age groups, each defined by a 5-year interval ranging from 0–4 years to 75 years and older [18]. The annual lung cancer cases were calculated for each age group throughout the study period. Furthermore, the NCR-SA dataset was aggregated by year, sex, race, and age group to calculate the number of lung cancer cases at each stratification level. The mid-year population data, which included race, sex, age, and year-wise population estimates, was merged with the aggregated NCR-SA data to incorporate the corresponding population estimates per stratification. The Segi-Doll standard population dataset was used for standardisation, consisting of age groups and their respective weights [18]. The Segi–Doll standard population dataset is a reference population structure developed for international comparisons of cancer incidence and mortality. It was first proposed by Segi and later modified by Doll et.al, and it provides age-specific weights for 5-year age groups up to 75 years and older [18-20]. These weights are widely used in epidemiology to calculate age-standardised rates (ASRs), ensuring that differences between populations are not due to variations in age structure but rather reflect real differences in disease burden [19-20]. To ensure compatibility with the age groupings in the NCR-SA dataset, the weights for patients aged 75 and above were grouped into a single age category (75+)**.** The resulting data included three datasets, which were the number of lung cancer cases categorised by year, the population stratified by year and sex, and the population stratified by year and race.

3.3.4 Data analysis

The analysis was conducted using Python version 3.9 within the Spyder integrated development environment to perform descriptive statistics. Demographic and clinical factors, including sex, population group, and histological subtypes, were summarized using frequencies and percentages to provide an overview of the data. The Chi-Squared test was used to assess whether there was a

statistically significant association between each of the following categorical variables: sex, race, hospital, and age group and lung cancer in the study.

For trend analysis, each dataset was sequentially imported into the Joinpoint Regression Analysis software (version 5.3.0). This software was used to examine lung cancer incidence trends across the study period. Age-Standardised Incidence Rates (ASIR) were calculated and reported as the number of cases per 100,000 individuals at risk using the Segi-Doll world standard population as a reference [18]. The Joinpoint software identified different time periods where statistically significant changes occurred in the incidence trends. Key metrics derived from the analysis included crude rates, ASIRs and APC, along with their corresponding 95% confidence intervals (CI).

3.4 Results

3.4.1 Socio-Demographic Characteristics

There were 54,405 reported cases of lung cancer between 2011 and 2022. Of all the diagnosed lung cancers, 85.8% (n=46,661) were NSCLCs, 8.5% (n=4,607) were SCLCs and 5.8% (n=3,137) were unclassified types. There were 61.7% (n = 33,542) males and 38.3% (n = 20,845) females. A total of 3,881 (64.1%) patients were diagnosed in private hospitals, while 19,524 (35.9%) were diagnosed in public hospitals. Among different population groups, 28.3% (n=14,980) were Black, 18.3% (n=9,682) were Coloureds, 4.3% (n=2,254) were Indian/Asian, and 49.1% (n=25,970) were Whites (*Table 1*).

The distribution of sex (p=0.003), race (p=0.001) and hospital (p=0.001) differs significantly among lung cancer histological types. The median age at diagnosis was 65.0 years with an Interquartile Range of (57.0, 73.0). Age-specific incidence rates increase with age, rising from the 45–49 age group and a peaking at 70-74 years, followed by a subsequent decrease in individuals aged 75 and above across all racial groups. (*Figure 2*).

Table 2: *Demographic and Clinical Characteristics of Individuals Pathologically Diagnosed with Lung Cancer in South Africa (2011–2022)*

Characteristic	Non-small cell lung cancers n (%)	Small cell lung cancers n (%)	Unclassified n (%)	Total N (%)	p-value

	46,661(85.5)	4,607(8.5)	3,137(6.7)	54,405	
Sex					0.003
Male	28,899 (62.0)	2,752 (60.0)	1,891 (60.0)	33,542 (62.0)	
Female	17,744 (38.0)	1,855 (40.0)	1,246 (40.0)	20,845 (38.0)	
Missing	18(<1.0)	0(0.0)	0 (0.0)	18(0.0)	
Race					<0.001
Black	13,288 (28.0)	765 (17.0)	927 (30.0)	14,980 (28.0)	
Coloured	8,245 (18.0)	947 (21.0)	490 (16.0)	9,682 (18.0)	
Indian/Asian	1,975 (4.0)	133 (3.0)	146 (5.0)	2,254 (4.0)	
White	21,842 (47.0)	2,615 (57.0)	1,513 (48.0)	25,970 (48.0)	
Missing	1,311(3.0)	147(3.0)	61(2.0)	1,519(3.0)	
Hospital					<0.001
Private	29,791 (64.0)	3,154 (68.0)	1,936 (62.0)	34,881 (64.0)	
Public	16,870 (36.0)	1,453 (32.0)	1,201 (38.0)	19,524 (36.0)	
Median age (Q1, Q3)	65.0 (57.0, 73.0)	66.0 (59.0, 73.0)	62.0 (54.0, 71.0)	65.0 (57.0, 73.0)	
Age group					<0.001
<25	37 (0.1)	1 (0.0)	54 (2.0)	92 (0.0)	
25–34	313 (1.0)	17 (0.0)	83 (3.0)	413(1.0)	
35–44	1,715 (4.0)	98 (2.0)	184 (6.0)	997 (4.0)	
45–54	6,387 (14.0)	549 (12.0)	458 (15.0)	7,394 (14.0)	
55–64	13,778 (30.0)	1,329 (29.0)	1,024 (33.0)	16,131 (30.0)	
65–74	14,796 (32.0)	1,756 (38.0)	843 (27.0)	17,395 (32.0)	

75–84	8,132 (17.0)	741 (16.0)	410 (13.0)	9,283 (17.0)
85+	1,503 (3.0)	116 (3.0)	81 (3.0)	1,700 (3.0)

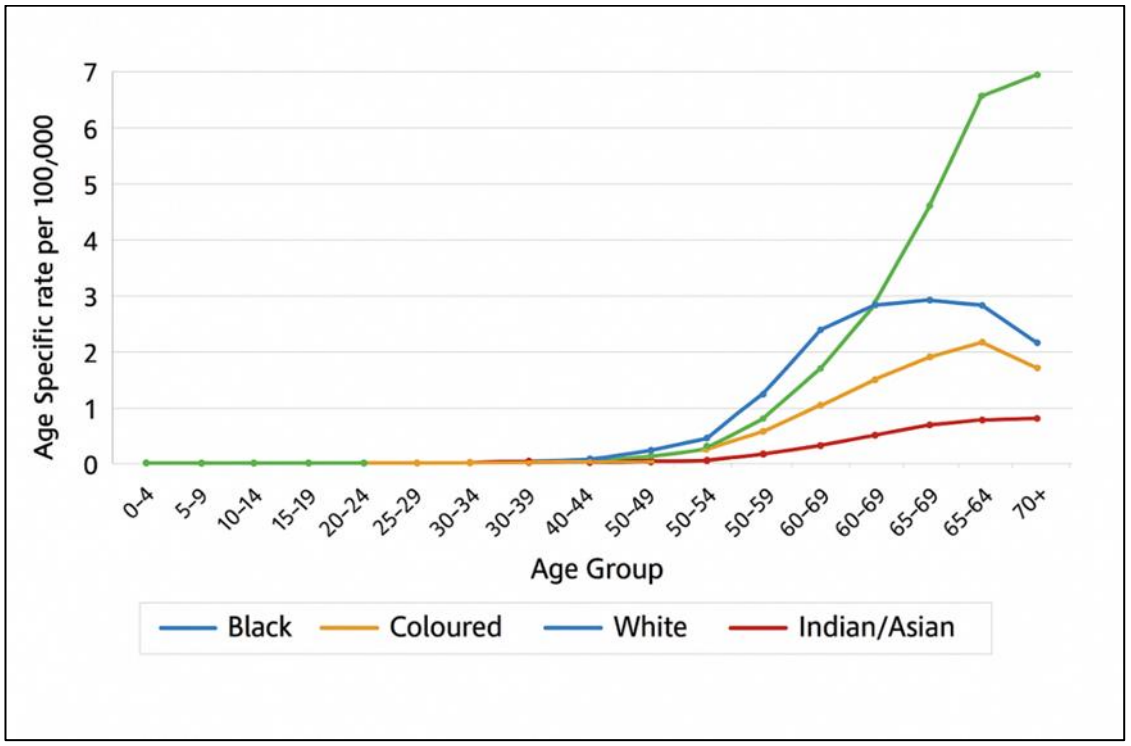


Figure 2: *Age-Specific Lung Cancer Incidence Rates by Race in South Africa (2011–2022)*

3.4.2 Lung cancer trends

The overall ASIR was 5.36 per 100,000 persons, with a 2.1% increase (95% CI: -2.8 to -1.3) observed over the study period. Between 2011 and 2017, there was a 3.4% increase in ASIR (95% CI: 2.0–5.5) for females and a 0.1% increase (95% CI: -1.3–2.7) for males. This was followed by an 11.5% decrease (95% CI: -14.3 to -7.6) in females and a 13.4% decrease (95% CI: -16.8 to -9.1) in males between 2017 and 2020. Subsequently, ASIR increased again by 7.0% (95% CI: 1.1–12.7) in females and 4.8% (95% CI: -3.4 to 12.0) in males (*Figure 3*).

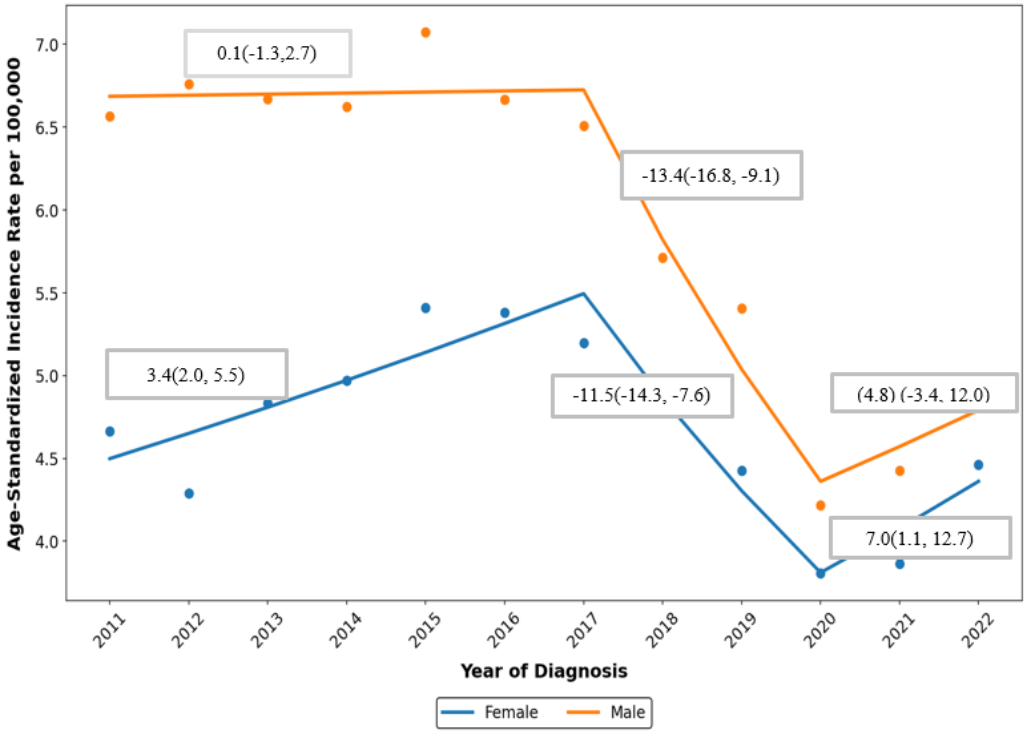


Figure 3: *Age-Standardised Lung Cancer Incidence Rates by Sex (2011–2022)*

The ASIR increased by 11.3% in Whites, 10.5% in Coloureds, 6.3% in Indians, and 2.7% in Blacks per 100,000 (Figure 4). Among Blacks, ASIR increased by 1.8% between 2011 and 2016, followed by an 11.1% decrease from 2016 to 2020, and a subsequent 5.5% increase between 2020 and 2022. The Coloured population had an ASIR increase of 0.3% (CI: 0.2, 5.0) in 2011-2017, a 16.6% (CI: -6.7, 19.7) decrease in 2017-2020, and a subsequent 7.9% (CI: -6.7, 20.0) increase in 2020-2022. Among Whites, ASIR increased by 2.5% (CI: 1.5, 3.8) from 2011 to 2017, followed by a 10.0% (CI: -11.8, 7.5) decrease in 2017-2020 and a 3.9% (CI: 0.14, 7.2) increase in 2020-2022. Among Indians, ASIR increased by 8.7% (CI: 0.5, 40.2) in 2011-2015 and decreased by 6.7% (CI: -17.3, -3.6) in 2015-2022 (*Table 2*).

Table 2: *Annual Percent Change (APC) in Pathologically Confirmed Lung Cancer by Sex and Race (2011–2022)*

Race	Trend 1		Trend 2		Trend 3	
	Calendar Period	APC (95% CI)	Calendar Period	APC (95% CI)	Calendar Period	APC (95% CI)
Black	2011–2016	1.8 (–0.7, 5.9)	2016–2020	–11.1 (–15.8, –7.8) *	2020–2022	5.5 (–2.6, 13.4)
Coloured	2011–2017	0.3 (–2.0, 5.0)	2017–2020	–16.6 (–6.7, 19.7) *	2020–2022	7.9 (–6.7, 20.0)
Indian	2011–2015	8.7 (0.5, 40.2) *	2015–2022	–6.7 (–17.3, –3.6) *	–	–
White	2011–2017	2.5 (1.5, 3.8) *	2017–2020	–10.0 (–11.8, –7.5) *	2020–2022	3.9 (0.14, 7.9) *
Females	2011–2017	3.4 (2.0, 5.5) *	2017–2020	–11.5 (–14.2, –7.6) *	2020–2022	7.0 (1.1, 12.7)
Males	2011–2017	0.1 (–1.3, 2.7)	2017–2020	–13.4 (–16.7, –8.9) *	2020–2022	4.8 (–3.3, 11.9)
Full Range	2011–2022		–2.1 (–2.8, –1.3) *		–	

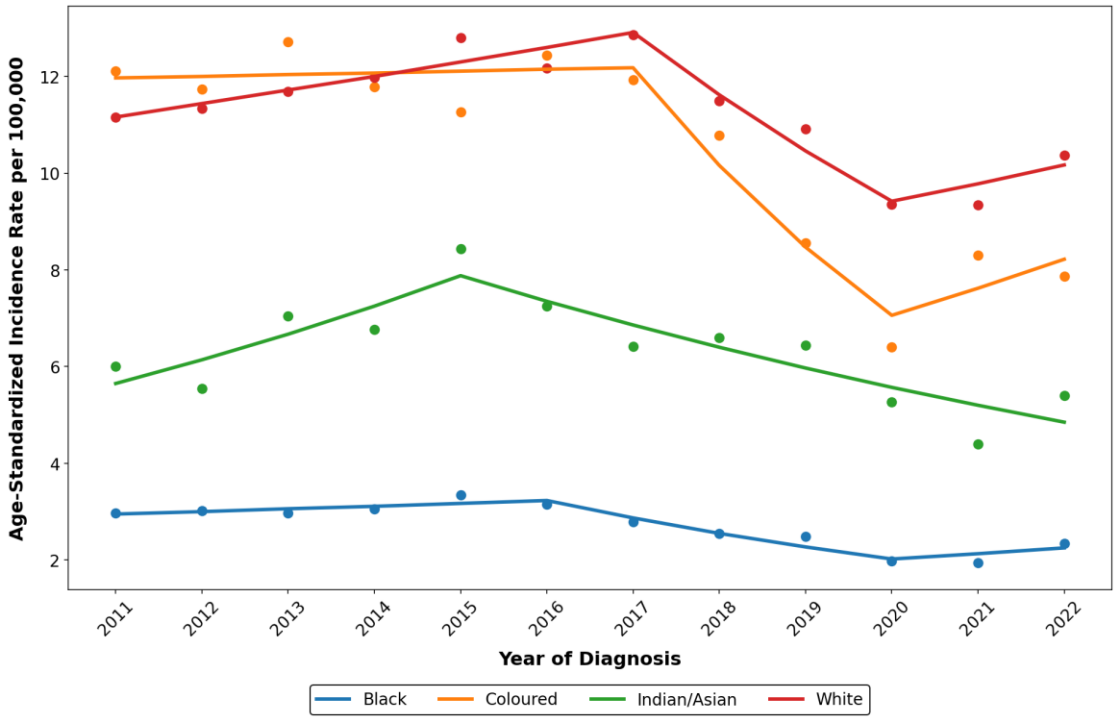


Figure 4: *Age-Standardised Lung Cancer Incidence Rates by Race in South Africa (2011–2022)*

3.5 Discussion

The findings of this study highlight significant disparities in lung cancer diagnosis across demographic groups and healthcare sectors, providing evidence in informing public health policies for targeted interventions. The study found that NSCLC (85.8%) was the dominant subtype when compared to SCLs (8.5%) and unspecified subtypes (5.8%) (Table 1). The predominance of NSCLCs in our dataset is consistent with global patterns [4]. The higher proportion of lung cancer cases observed in males (61.7%) compared to females (38.3%) aligns with existing literature, which reports a significantly higher smoking prevalence among males than females [21].

These findings are further supported by the Global Adult Tobacco Survey in South Africa, which showed significant variations in tobacco use across the population. The survey found that 25.8% of adults were current tobacco smokers and from those 41.2% were males and 11.2% were females [22]. Additionally, 18% of adults were exposed to tobacco smoke in their homes and 11.2% at their workplace [22]. The racial distribution of lung cancer cases also shows disparities, with Whites accounting for the highest proportion (49.1%), followed by Blacks (28.3%), Coloureds (18.3%) and Asians/Indians (4.3%). Given that tobacco smoking is a major risk factor for lung cancer, these findings contrast with smoking prevalence patterns in SA, where Coloureds have the highest smoking rate (40.1%), followed by Indians/Asians (22.1%), Whites (15.3%), and Black Africans (15.1%) [23].

This pattern may reflect differences in healthcare access and socioeconomic status rather than true variations in disease burden. Coloured and Black populations, despite having higher smoking rates, may experience diagnostic delays due to their dependence on resource-constrained public healthcare services [24]. The higher proportion of cases diagnosed in private hospitals (64.1%) compared to public facilities (35.9%) further suggests inequities in healthcare access. Private healthcare facilities, which are predominantly used by wealthier populations, offer more advanced diagnostic tools, shorter waiting times, and better access to specialist care, leading to earlier and more frequent lung cancer diagnoses [47]. In contrast, individuals relying on public healthcare often experience delayed diagnoses due to resource constraints, long waiting periods, and limited access to specialized oncology services [24]. This may contribute to underreporting of lung cancer cases in

the public sector and a higher proportion of late-stage diagnoses among economically disadvantaged populations [25-26].

Between 2011 and 2022, lung cancer incidence in SA showed an overall decline (Table 2, Figure 4). This trend may be influenced by multiple factors, including the impact of tobacco control policies and disparities in healthcare access across racial and socioeconomic groups. The 1993 Tobacco Products Control Act, which introduced regulations on advertising, health warnings, and smoke-free environments, may have contributed to the gradual decline in lung cancer incidence [27].

However, disparities in early detection likely played a role in observed racial differences. Whites, with greater economic means who are more likely to use private healthcare, may have higher reported lung cancer incidence due to better diagnostic availability, while underdiagnosis in public healthcare settings could contribute to lower reported rates among other racial groups. This is supported by studies indicating that Black patients are more frequently diagnosed at later stages, likely due to limited access to screening and diagnostic services [26].

The age-specific incidence rates of lung cancer peaked at 70–74 years (*Figure 1*). The distribution of cases in this study shows that the highest proportion occurred among individuals aged 55–74 years, with 32.0% in the 65–74 age group and 29.6% in the 55–64 age group. This aligns with smoking prevalence patterns, where the highest smoking rates among males are observed in the 45–64 and 25–44 age groups respectively [21].

Given the period between smoking exposure and lung cancer development, the peak incidence in older age groups likely reflects historical smoking behaviours. The observed sex-specific differences (Table 2, Figure 3) in lung cancer incidence trends may reflect variations in smoking patterns as reflected in other studies that in many high-income countries, lung cancer incidence has decreased in males, attributed to earlier smoking cessation campaigns and declining smoking prevalence [26]. The significant increase in female incidence between 2011 and 2017 aligns with global trends, where women initiated smoking later than men, leading to a delayed rise in lung cancer cases (Figure 3) [28]. The observed decline in lung cancer incidence from 2017 to 2020 may be attributed to the effects of tobacco control policies implemented in previous years [27]. Moreover, the pronounced decrease in 2019 could be partially attributed to healthcare service

disruptions during the COVID-19 pandemic, which may have resulted in diagnostic delays rather than an actual reduction in lung cancer cases.

3.6 Conclusion

In conclusion, this study provides important evidence into the epidemiology of lung cancer in SA. The observed incidence trends and demographic patterns provide valuable information for public health planning and resource allocation by identifying populations experiencing unequal disease burden, highlighting healthcare access inequities between sectors, and establishing baseline data for monitoring the effectiveness of tobacco control policies.

The disparity between the private and public healthcare sector diagnoses demonstrates these inequities, with private facilities diagnosing more cases compared to the populations they serve. These disparities highlight the need to strengthen diagnostic capacity in public facilities. Moreover, the significant differences in incidence across demographic groups, particularly the increasing burden among females and variations across racial groups, highlight the importance of developing targeted, population-specific intervention strategies to improve early detection and access to diagnostic services.

To further enhance understanding of the lung cancer burden in SA, it is essential to strengthen and expand pathology-based cancer registration systems to capture cases diagnosed through alternative means. In this context, the rule-based NLP algorithm showed reliable performance in extracting NSCLC biomarker information, providing a practical tool to support large-scale biomarker analysis from unstructured pathology reports. Future research should focus on identifying the most effective strategies for reducing disparities and improving outcomes for all South Africans affected by lung cancer.

Bibliography

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71

[2] International Agency for Research on Cancer. Estimated age-standardized incidence rates (world) in 2020, world, south africa, both sexes, ages 0–84 (excl. nmsc). https://gco.iarc.fr/today, 2020. Accessed March 2025.

[3] Krishna Chaitanya Thandra, Adam Barsouk, Kalyan Saginala, John Sukumar Aluru, and Alexander Barsouk. Epidemiology of lung cancer. Contemporary Oncology/Wsp´ol-czesna Onkologia, 25(1):45–52, 2021.

[4] Yanting Zhang, Salvatore Vaccarella, Eileen Morgan, Mengmeng Li, Jaione Etxeberria, Eric Chokunonga, Shyam Shunker Manraj, Bakarou Kamate, Abidemi Omonisi, and Freddie Bray. Global variations in lung cancer incidence by histological subtype in 2020: a population-based study. The Lancet Oncology, 24(11):1206–1218, 2023.

[5] Xiaochen Dai, Emmanuela Gakidou, and Alan D Lopez. Evolution of the global smoking epidemic over the past half century: strengthening the evidence base for policy action. Tobacco control, 31 (2):129–137, 2022.

[6] Lauren May, Kathryn Shows, Patrick Nana-Sinkam, Howard Li, and Joseph W Landry. Sex differences in lung cancer. Cancers, 15(12):3111, 2023.

[7] Rachel Brathwaite, Juliet Addo, Liam Smeeth, and Karen Lock. A systematic review of tobacco smoking prevalence and description of tobacco control strategies in sub-saharan african countries; 2007 to 2014. PloS one, 10(7):e0132401, 2015.

[8] Samir Soneji, Nichole T Tanner, Gerard A Silvestri, Christopher S Lathan, and William Black. Racial and ethnic disparities in early-stage lung cancer survival. Chest, 152(3):587–597, 2017.

[9] Stacey L Tannenbaum, Tulay Koru-Sengul, Wei Zhao, Feng Miao, and Margaret M Byrne. Survival disparities in non-small cell lung cancer by race, ethnicity, and socioeconomic status. The Cancer Journal, 20(4):237–245, 2014

[10] Ronwyn Van Eeden, Mariza Tunmer, Agneta Geldenhuys, Simon Nayler, and Bernardo Leon Rapoport. Lung cancer in south africa. Journal of Thoracic Oncology, 15(1):22–28, 2020

[11] Carissa C Jones, Sarah Fletcher Mercaldo, Jeffrey D Blume, Angela S Wenzlaff, Ann G Schwartz, Heidi Chen, Stephen A Deppen, William S Bush, Dana C Crawford, Stephen J Chanock, et al. Racial disparities in lung cancer survival: the contribution of stage, treatment, and ancestry. Journal of Thoracic oncology, 13(10):1464–1473, 2018.

[12] Buhle Lubuzo, Themba Ginindza, and Khumbulani Hlongwana. Exploring barriers to lung cancer patient access, diagnosis, referral and treatment in kwazulu-natal, south africa: The health providers' perspectives. Translational lung cancer research, 8(4):380, 2019.

[13] Andy Gray and Yousuf Vawda. Health legislation and policy. South African health review, 2018 (1):1–11, 2018.

[14] Coenraad FN Koegelenberg, Shane Dorfman, Ivan Schewitz, Guy A Richards, Shaun Maasdorp, Clifford Smith, Keertan Dheda, et al. Recommendations for lung cancer screening in southern africa. Journal of Thoracic Disease, 11(9):3696, 2019.

[15] Cornelius Nattey, Jimi Fadahun, Briamoh Bello, Danuta Kielkowski, and Gill Nelson. Trends in lung cancer mortality in south africa 1995–2008. Occupational and Environmental Medicine, 68 (Suppl 1):A111–A111, 2011.

[16] National Health Laboratory Service. Diagnostic services. https://www.nhls.ac.za/ diagnostic-services/, 2025. Accessed 12 February 2025.

[17] Wenlong Carl Chen, Patricia Kellett, Mike Greyling, Elvira Singh, and Mazvita Sengayi-Muchengeti. The use of surnames to impute missing ethnicity data in the south african national cancer registry database. 2022

[18] Freddie Bray and Jacques Ferlay. Age standardization. Unpublished or online document, 2014. International Agency for Research on Cancer (IARC).

[19]. Segi M. Cancer mortality for selected sites in 24 countries (1950–1957). Sendai: Tohoku University School of Public Health; 1960.

[20]. Doll R, Payne P, Waterhouse JAH. Cancer incidence in five continents: a technical report. Berlin: Springer-Verlag; 1966.

[21] Catherine O Egbe, Portia Nevhungoni, and Mukhethwa Londani. Disparities in tobacco smoking by sex and employment status: Results from south africa's first global adult tobacco survey. Social Sciences, 13(10):550, 2024.

[22] National Department of Health (NDoH), South African Medical Research Council (SAMRC), Statistics South Africa (Stats SA), and World Health Organization (WHO). Global adult tobacco survey: South africa report. Technical report, National Department of Health, Pretoria, South Africa, 2021.

[23] Priscilla Reddy, Khangelani Zuma, Olive Shisana, K Jonas, and Ronel Sewpaul. Prevalence of tobacco use among adults in south africa: results from the first south african national health and nutrition examination survey. South African Medical Journal, 105(8):648–655, 2015.

[24] Ayhan Arısoy. Natural language processing algorithms and performance comparison. Yalva¸c Akademi Dergisi, 9(2):106–121.

[25] Ethelwynn L Stellenberg. Accessibility, affordability and use of health services in an urban area in south africa. Curationis, 38(1):1–7, 2015.

[26] Jennifer Richmond, Megan Hollister Murray, Cato M Milder, Jeffrey D Blume, and Melinda C Aldrich. Racial disparities in lung cancer stage of diagnosis among adults living in the South Eastern United States. Chest, 163(5):1314–1327, 2023.

[27] Government Gazette No. 14916. Tobacco products control act 83 of 1993, 1993.

[28] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6):394–424, 2018.

4 Extended Methodology

Introduction

This chapter details the methodologies used to achieve Objectives 2 and 3. It describes the rule-based NLP approach for automating biomarker extraction and explains how the extracted biomarkers are analysed to determine their prevalence and testing rates.

4.1 Rule-Based Natural Language Processing

4.1.1 Data Preparation

Pathology reports for individuals diagnosed with lung cancer during the study period were received from the NCR-SA database. The dataset contained a mix of structured and unstructured data. Demographic information, ICD10 morphology and topography codes were populated in structured columns, while a column named "reports," which contained pathology reports for each patient, was in free text format. To ensure the dataset included only lung cancer cases, the data was filtered using the topography code C34, which represents the lung according to the ICD-10 coding manual. The dataset was further filtered using morphology codes from ICD-10 to include only NSCLCs. The pathology reports were pre-processed by removing special characters to ensure consistency and standardization. This process involved removing extra spaces, special characters, and converting all text to lowercase (Figure 5).

The variables were standardised and categorised to ensure comparability. Sex was reported as male and female. Age was calculated at diagnosis and grouped into categories (<25, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, and 85+ years). Race was grouped into four standard categories (Black, Coloured, Indian/Asian, and White) as routinely captured by the NCR-SA. Hospital sector was categorised as public (NHLS laboratories) or private (reporting private laboratories). Histological subtypes of NSCLC were derived from ICD-O-3 morphology codes and classified into adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. A small number of less common NSCLC morphologies (e.g., sarcomatoid carcinoma, adenosquamous carcinoma, mucoepidermoid carcinoma) were combined into an "Unclassified NSCLC" category to ensure stable reporting, given their small case numbers. Missing values were retained in all descriptive tables to provide transparency about data completeness.

4.1.2 Gold Standard Dataset

To create a gold standard dataset that was used for evaluation, an extensive literature review was conducted to identify specific NSCLC biomarkers for extraction from pathology reports. A subset of 300 pathology reports was randomly selected, and each report was manually annotated for biomarker presence and corresponding result texts. Each extracted result was classified into one of three categories: positive (indicating the biomarker was detected), negative (indicating absence), or not reported (when the biomarker was not mentioned). The individual biomarkers from the literature were listed in columns. For each pathology report, if a listed biomarker was tested, the corresponding cell was annotated with its result as **"positive"** or **"negative"**; if it was not tested, it was annotated as **"not reported".**

4.1.3 Rules creation

Using these annotated reports, regular expression-based rules that would be able to recognize various biomarkers along with their associated result texts from the pathology reports were developed. The result texts were classified into three groups: positive, for texts indicating a positive result; negative, for texts representing a negative result; and not reported, for cases where the pathology report did not mention the specific biomarker. The rule-based algorithm was developed iteratively rather than in a single step. Initial rules were tested on the annotated subset of 300 reports and outputs were compared against the gold standard annotations. False positives (cases where biomarkers were incorrectly identified) and false negatives (cases where true biomarkers were missed) were reviewed. Based on these findings, the rules were refined, expanded, and retested in repeated cycles until satisfactory extraction performance was achieved across all biomarkers.

4.1.4 Rule Application

These pre-defined rules were applied to the subset of pathology reports to extract biomarkers along with their corresponding result texts from each report within the report column. The algorithm scanned the text in each report, identifying patterns that matched the specified rule conditions and automatically populating columns with the extracted biomarkers and their respective result texts. To

maintain consistency, each column containing biomarkers extracted by the algorithm was named using the biomarker's name followed by "_extracted".

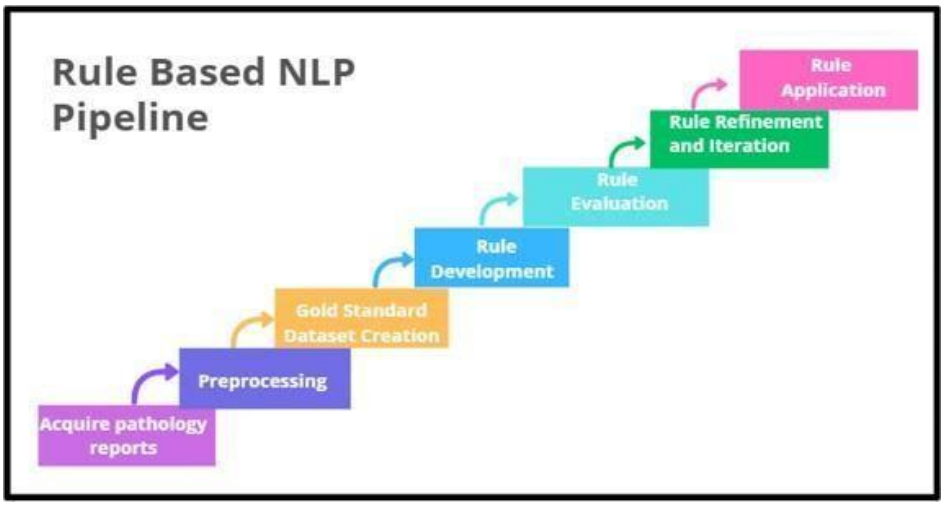


Figure 5: *Rule Based NLP pipeline*

4.1.5 Model Evaluation

Validation was conducted by comparing the algorithm's predictions against the manually annotated gold standard subset. For each biomarker, true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were calculated, and confusion matrices were generated to visualize performance. These values were then used to derive precision, recall, and F1-scores, ensuring that the algorithm's accuracy was thoroughly quantified before applying it to the full dataset. Precision measured the proportion of correctly identified biomarkers among all biomarkers predicted by the model, while recall evaluated the proportion of correctly identified biomarkers relative to the actual biomarkers present in the data. The F1-score, calculated as the harmonic mean of precision and recall, assess the overall performance of the algorithm [68-69]. Metrics were calculated as follows [69]:

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \qquad F1\ Score = \frac{Precision\ X\ Recall}{Precision+Recall}$$

4.1.6 Rule Application

Once the algorithm achieved satisfactory performance, it was applied to the entire dataset to extract biomarkers and their corresponding result texts from pathology reports. The extracted biomarkers were populated in their respective columns and classified as positive, negative, or not reported. The

prevalence rate for each biomarker was calculated as the total number of positive extractions divided by the sum of positive, negative, and not reported cases, assuming that reports labelled as “not reported” were negative. Testing rates for different hospital types were determined for each biomarker, calculated as the proportion of reports with a positive or negative result for that biomarker relative to the total number of reports, including positive, negative, and not reported cases.

5 Extended Results

Introduction

This chapter presents the results of the rule-based NLP extraction from pathology reports. It includes an evaluation of the algorithm's performance using precision, recall, and F1-score metrics. It also provides an analysis of the extracted biomarkers, focusing on their prevalence and testing rates.

5.1 Demographics Characteristics

A total of 46,160 individuals were diagnosed with NSCLC during the study period (Table 3). Adenocarcinoma was the predominant subtype, representing 58.4% (n=26,956) of cases, followed by squamous cell carcinoma at 23.6% (n=10,896), unclassified NSCLCs at 14.6% (n=3,929), and large cell carcinomas at 5.5% (n=2,551). This distribution shows that adenocarcinoma is the dominant form of NSCLC in SA, consistent with global trends.

Most patients (64.0%, n=29,409) were diagnosed in private hospitals compared with 36.0% (n=16,751) in public facilities, suggesting possible differences in access to diagnostic services between the public and private sectors. Males accounted for majority of cases 64.5% (n=17,434), reflecting historical smoking patterns and females for 35.2% (n=9,469) of case.

Among the racial groups, 29.0% (n=13,162) were Black, 18.8% (n = 8,122) were Coloured, 4.0% (n=1,948) were Asian/Indian, and 40.0% (n=21,627) were White. The Median age at diagnosis was 65.0 (57.0, 73.0), with most cases observed in individuals aged 65-74 at 32.0 %

(n=14,594) and those aged 54-64 at 30.0 % % (n=14,594) and those aged 54-64 at 30.0%

(n=13,620). This pattern highlights that NSCLC remains largely a disease of older adults. There were statistically significant differences in the distribution across NSCLC subtypes ($p < 0.001$) for sex, race, hospital type, and age groups.

Table 3: *Demographic and Clinical Characteristics of Non-Small Cell Lung Cancer Patients by Hospital and Cancer Subtype*

Characteristics	Unclassified NSCLC N (%)	Adenocarcinomas N (%)	Squamous Cell N (%)	Large Cell N (%)	Total N (%)	p-value
	5,757(12.5)	26,956(58.4)	10,896(23.6)	2,551(5.5)	46,160(100)	
Sex						<0.001
Male	3,748 (65.0)	15,207 (56.0)	7,848 (72.0)	1,785 (70.0)	28,588 (62.0)	
Female	2,008 (35.0)	11,737 (44.0)	3,045 (28.0)	764 (30.0)	17,554 (38.0)	
Missing	1(<1.0)	12(<1.0)	3(<1.0)	2(<1.0)	18(<1.0)	
Race						<0.001
Black	1,631 (28.0)	6,956 (26.0)	3,642 (33.0)	933 (37.0)	13,162 (29.0)	
Coloured	1,382 (24.0)	4,350 (16.0)	1,975 (18.0)	415 (16.0)	8,122 (18.0)	
Indian/Asian	156 (3.0)	1,360 (5.0)	306 (3.0)	126 (5.0)	1,948 (4.0)	
White	2,442 (42.0)	5 13,484 (50.0)	4,687 (43.0)	1,014 (40.0)	21,627 (4.0)	
Missing	146(3.0)	806 (3.0)	286(3.0)	63 (2.0)	1,301 (3.0)	
Hospital						<0.001
Public	3,351 (58.0)	7,131 (26.0)	5,004 (46.0)	1,265 (50.0)	16,751 (36.0)	
Private	2,406 (42.0)	19,825 (74.0)	5,892 (54.0)	1,286 (50.0)	29,409 (64.0)	
Median age (Q1, Q3)	64.0 (55.0, 71.0)	65.0 (57.0, 73.0)	66.0 (59.0, 73.0)	62.0 (55.0, 71.0)	65.0 (57.0, 73.0)	
Age Group						<0.001
<25	8 (0.0)	17 (0.0)	11 (0.0)	3(0.0)	39 (0.0)	
25-34	44 (1.0)	188 (1.0)	54 (0.0)	27(1.0)	313(1.0)	
35-44	244 (4.0)	1,113 (4.0)	237 (2.0)	102(4.0)	1,696(4.0)	
45-54	955(17.0)	3,716 (14.0)	1,150 (11.0)	468(18.0)	6,329(14)	
55-64	1,826 (32.0)	7,660 (28.0)	3,304 (30.0)	830(33)	13,620 (30)	
65-74	1,693(29.0)	8,488 (31.0)	3,724(34.0)	689(27.0)	14,594 (32)	
75-84	769 (13)	4,849 (18)	2,077 (19.0)	376(15.0)	8,071 (170)	
85+	93 (3.0)	925(3)	339 (3.0)	56 (2.0)	1,498 (3%)	

5.2 Model Evaluation

The biomarker extraction model achieved extraction of 22 different biomarkers from pathology reports (Figure 6). Performance differed across biomarkers, with some achieving perfect scores while others showed reduced extraction capabilities. Eighteen biomarkers achieved perfect precision scores of 1.000, indicating accurate identification when biomarker mentions were detected

in pathology reports. The remaining four biomarkers showed strong performance: NAPSIN A (0.983), Thyroid Transcription Factor-1(TTF1) (0.993), Tumour protein p63 (P63) (0.950), and Np63 isoform (P40) (0.889). This high precision across the panel suggests minimal false positive extraction errors during the automated extraction.

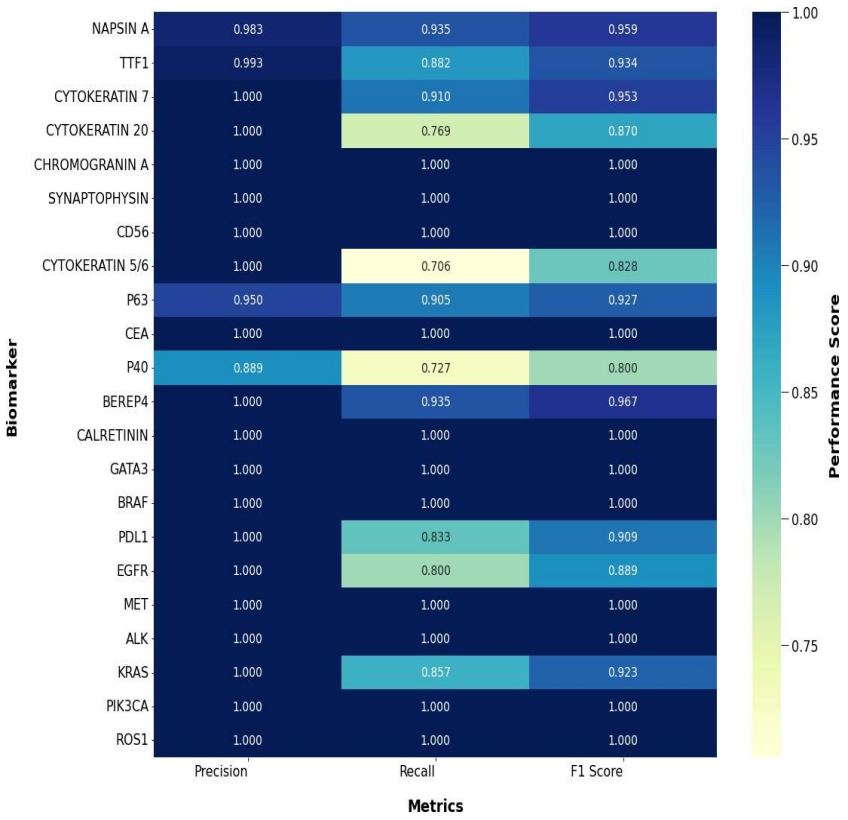


Figure 6: *Performance Metrics for Lung Cancer Biomarker Extraction*

Recall performance showed results that are more diverse across the biomarker set. Eleven biomarkers achieved perfect recall scores of 1.000, including Chromogranin A, Synaptophysin, Cluster of Differentiation 56 (CD56), Carcinoembryonic Antigen (CEA), Calretinin, GATA Binding Protein 3 (GATA3), B-Raf Proto-Oncogene (BRAF), MET Proto-Oncogene Receptor Tyrosine Kinase (MET), Anaplastic Lymphoma Kinase (ALK), Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA), and ROS Proto-Oncogene 1 Receptor Tyrosine Kinase (ROS1). Lower recall was observed for several biomarkers: Cytokeratin 5/6 (CK5/6) (0.706), P40 (0.727), Cytokeratin 20 (CK20) (0.769), Epidermal Growth Factor Receptor (EGFR) (0.800), and Programmed Death-Ligand 1 (PDL1) (0.833). The remaining biomarkers achieved recall scores above 0.880, reflecting reliable detection of biomarker information from pathology reports.

5.3 Prevalence of NSCLC Biomarkers

The 22 extracted biomarkers comprised 14 diagnostic biomarkers and 8 predictive biomarkers. Of the total of 42,909 pathology reports, 49,625 biomarker tests were conducted with an average of 1.2 tests per report. Diagnostic biomarker testing accounted for 95.9% (n=47,576) of all tests, while predictive biomarker testing accounted for 4.1% (n=2,049) of all tests.

TTF1 was the most frequently tested biomarker, performed in 28.0% of reports (n=12,017), followed by CK7 in 22.0% of reports (n=9,461) and CK20 in 16.3% of reports (n=7,015). These biomarkers had a prevalence rate of (15.8, 16.2, 1, 2), respectively. A huge proportion of the reports did not report or test these biomarkers, suggesting that biomarkers are either no tested or reported on the pathology reports (Table 4).

Table 4: *Summary of Biomarker Prevalence*

Biomarker	Total Tests	Positive	Negative	Not Reported	Prevalence Rate (%)
Diagnostic	**N = 47,576(95.9%)**				
Napsin A	4,893 (11.4)	2,758	2,135	38,016	6.4
TTF1	12,017 (28.0)	6,777	5,240	30,892	15.8
Cytokeratin 7	9,461 (22.0)	6,961	2,500	33,448	16.2
Cytokeratin 20	7,015 (16.3)	528	6,487	35,894	1.2
Chromogranin A	1,472 (3.4)	204	1,268	41,437	0.5
Synaptophysin	1,472 (3.4)	317	1,522	41,070	0.7
CD56	866 (2.0)	205	661	42,043	0.5
Cytokeratin 5/6	3,093 (7.2)	1,019	2,074	39,816	2.4
P63	5,111 (11.9)	2,312	2,799	37,798	5.4
CEA	1,124 (2.6)	733	391	41,785	1.7
P40	1,043 (2.4)	359	684	41,866	0.8
Berep4	1,044 (2.4)	693	351	41,865	1.6
Calretinin	1,887 (4.4)	119	1768	41,022	0.3
GATA3	933 (2.2)	172	761	41,976	0.4
Predictive	**N = 2,049(4.1%)**				
BRAF	23 (0.1)	7	16	42,886	0.0
PDL1	773 (1.8)	505	268	42,136	1.2
EGFR	477 (1.1)	74	403	42,432	0.2
MET	131 (0.3)	34	97	42,778	0.1
ALK	194 (0.5)	12	182	42,715	0.0
KRAS	116 (0.4)	47	67	42,795	0.1
PIK3CA	166 (0.4)	3	163	42,743	0.0
ROS1	169(0.4)	3	166	42,740	0.0

Napsin A was tested in 11.4% (n=4,893) of pathology reports with a prevalence rate of 6.4%. p63 was reported in 11.9%(n=5,111) reports with a prevalence rate of 5.4%. CK5/6 was tested in 3,093 (7.2%) with a prevalence rate of 2.4%. GATA3 was the least diagnostic biomarker tested at in 2.2 %(n=933) of reports with a prevalence rate of 0.4%. PDL1 was the most frequently tested predictive biomarker, performed in 1.8% of reports (n=773) with a prevalence rate of 1.2%, followed by EGFR tested in 1.1% of reports (n=477) with a prevalence rate of 0.2%, and MET tested in 0.3% of reports (n=131) with a prevalence rate of 0.1% (Table 4).

ALK was tested in 0.5% of reports (n=194) with a prevalence rate of 0.0%, while KRAS was tested in 0.4% of reports (n=116) with a prevalence rate of 0.1%. PIK3CA and ROS1 showed similar testing frequencies at 0.4% of reports each (n=166 and n=169, respectively), both with prevalence rates of 0.0%. BRAF had the lowest testing frequency at 0.1% of reports (n=23) with a prevalence

rate of 0.0%. Overall, these predictive biomarkers showed much lower testing frequencies compared to diagnostic biomarkers, with most tested in less than 0.5% of reports.

5.4 Testing Rates

The biomarker testing rates across private and public hospitals are summarised in Table 5 below. TTF1 was the most frequently tested biomarker in both hospital sectors, with similar rates observed in private hospitals (27.9%) and public hospitals (28.3%). CK7 was the second most commonly tested biomarker, showing higher testing rates in private hospitals (23.2%) compared to public hospitals (20.0%). Cytokeratin 20 was more frequently tested in private hospitals (17.0%) than public hospitals (15.3%). P40 showed a disparity with 3.7% testing in private hospitals compared to only 0.2% in public hospitals.

Several showed comparable testing rates between hospital types. Napsin A was tested in 11.7% of private hospital reports and 10.9% of public hospital reports. CD56 and GATA3 maintained consistent low-level testing across both sectors, with rates of approximately 2% in each hospital type.

Table 5: *Biomarker Testing Rates by Health Sector*

Biomarker	Private (%)	Public (%)
CD56	2.07	1.94
CEA	2.56	2.72
Calretinin	4.19	4.75
Chromogranin A	3.17	3.89
CK20	16.98	15.25
CK7	23.21	20.03
EGFR	1.74	0.02
GATA3	2.14	2.24
KRAS	0.42	0.00
MET	0.48	0.00
Napsin A	11.69	10.90
P63	10.70	14.03
ROS1	0.62	0.00
Synaptophysin	4.11	4.59
TTF1	27.87	28.25
ALK	0.68	0.05
BerEp4	2.19	2.86
BRAF	0.08	0.00
CK 5/6	6.95	7.65
P40	3.69	0.23
PDL1	2.70	0.23
PIK3CA	0.61	0.00

By comparison, certain biomarkers showed different testing patterns across sectors. P63 testing occurred in 14.0% of public hospital reports and 10.7% of private hospital reports. CK5/6 was tested in 7.7% of public hospital reports and 7.0% of private hospital reports. Other biomarkers included Synaptophysin (4.6% public, 4.1% private), Calretinin (4.8% public, 4.2% private), Chromogranin

A (3.9% public, 3.2% private), BerEP4 (2.9% public, 2.2% private), and CEA (2.7% public, 2.6% private).

Other biomarkers were tested predominantly in private hospitals: KRAS (0.4% private, 0.0% public), MET (0.5% private, 0.0% public), ROS1 (0.6% private, 0.0% public), BRAF (0.08% private, 0.0% public), and PIK3CA (0.6% private, 0.0% public). ALK testing occurred in 0.7% of private hospital reports and 0.05% of public hospital reports.

6. Discussion and Conclusion

Introduction

This chapter discusses and synthesizes the extended results of the study. It examines findings on NSCLC characteristics, biomarker prevalence, and the performance of the natural language processing method for biomarker extraction. The chapter analyses these results in the context of existing literature. The conclusion summarizes key findings.

6.1 Discussion

This study found that adenocarcinoma (58%) was the most common NSCLC subtype, followed by squamous cell carcinoma (23.4%), unclassified NSCLC (13.1%), and large cell carcinoma (5.5%) (Table 3). These findings align with global trends, where adenocarcinoma has become the most prevalent subtype of lung cancer, replacing squamous cell carcinoma as the leading subtype, which is known to be associated with smoking [70]. Therefore, this shift is probably due to the declining smoking rates in men [15]. Males had a higher proportion (61.9%) of NSCLC while 38.1% were diagnosed in females. SCC was predominantly diagnosed in males (72.0%), while adenocarcinoma was more evenly distributed, with 56.4% of cases occurring in men. This distribution reflects global patterns, where SCC is predominantly found in men due to historically higher smoking rates, whereas adenocarcinoma is more common in women, particularly non-smokers [37]. Research suggests that 20% of women with lung cancer have never smoked, highlighting additional risk factors such as genetic predisposition, environmental exposures, and hormonal influences [71]. The study further highlights significant racial disparities in healthcare access in SA.

Despite being a minority population group, White patients accounted for 48.3% of NSCLC cases, whereas Black (29.3%), Coloured (18.1%), and Indian (4.3%) patients made up a smaller proportion of diagnoses. These findings may suggest unequal access to diagnostic services in the different population groups. These inequities are further reflected in healthcare sector disparities. Private hospitals, which serve only 15% of the population [57], diagnosed 63.7% of cases, while public hospitals accounted for just 36.3%. The disparity is even more apparent in adenocarcinoma diagnoses, where 73.5% of cases were identified in private hospitals compared to 26.5% in public

hospitals. This imbalance reflects structural inequities within South Africa's healthcare system [56]. Patients in the public sector often present with advanced disease, with several factors such as limited health education, poor healthcare accessibility, and inefficiencies in referral systems contributing to delayed diagnoses [53].

In this study, we developed a rule-based NLP model to extract NSCLC biomarkers from pathology reports, achieving high precision across multiple biomarkers. Compared to conventional manual extraction methods, our automated approach demonstrates several well-documented benefits. It removes the need for labour-intensive, time-consuming manual abstraction that is prone to human error [35], while also providing automated analysis functions that support clinical workflow optimisation and quality monitoring [59]. Manual abstraction of pathology reports demands extensive human resources and is often impractical for large research projects [72]. In contrast, NLP methods can efficiently process thousands of reports with consistency and scalability. CK7, CK20, Chromogranin A, Synaptophysin, CD56, CK5/6, CEA, BerEP4, Calretinin, GATA3, BRAF, PD-L1, EGFR, MET, ALK, KRAS, PIK3CA, and ROS1 all achieved a precision score of 1.000, showing high reliability in biomarker identification. These findings align with previous studies showing that rule-based NLP methods can perform well in extraction tasks, even when applied to unstructured pathology reports, particularly when domain-specific rules are carefully designed [66]. However, certain biomarkers showed slightly lower performance metrics. Napsin A, TTF1, P63, and P40, despite high precision (0.983, 0.993, 0.950, and 0.889, respectively), had lower recall values, suggesting that certain instances were missed. P40 had a recall of 0.727, which is consistent with previous research highlighting inconsistencies in how this text is documented across reports, making it more difficult to extract using predefined linguistic rules. Additionally, a portion of our pathology reports were written in Afrikaans, introducing further complexity in standardizing extraction rules and potentially contributing to variations in recall. Studies evaluating NLP models found that rule-based approaches often struggled with complex and inconsistent sentence structures and variability in text documentation and as a result there is need for more advanced, context-aware NLP techniques to improve extraction performance across diverse clinical narratives [73, 74]. While structured mentions (e.g., "EGFR positive") are easily identified, more complex sentence structures

(e.g., “The specimen tested negative for the presence of EGFR mutations, indicating that targeted therapy may not be effective in this case”) can lead to false negatives, impacting recall scores. Addressing this issue requires context-aware approaches that improve the model’s ability to detect biomarker mentions across different linguistic styles [74].

The strong performance of our model in EGFR, ALK, and PD-L1 extraction further validates the role of NLP in biomarker-driven NSCLC classification. These findings provide a foundation for further research into how these biomarkers guide targeted therapies. The effectiveness of NLP-driven biomarker extraction in enhancing precision medicine is further supported by studies showing that structured extraction methods improve the identification of targetable genetic alterations for personalised treatment strategies [75]. Based on our validation results demonstrating precision scores of 0.88-1.00, recall of 0.73-1.00, and F1 scores of 0.80-1.00 when compared against the manually annotated data (Figure 6), our NLP model reliably extracted NSCLC biomarkers from pathology reports.

Given this validated performance, it is important to examine biomarker prevalence in our dataset to assess potential clinical applications. The prevalence of diagnostic biomarkers such as TTF-1 (15.8%), CK7 (16.2%), and Napsin A (6.4%) in the samples that were tested aligns with their established role in NSCLC subtyping and classification [76]. These prevalence rates are consistent with the proportion of adenocarcinoma cases in our dataset, supporting the known association between these markers and this NSCLC subtype (Table 3). The testing frequencies for TTF1, CK7 and Napsin A (Table 5) observed across both private and public hospitals indicate frequent use of these diagnostic markers. This is consistent with findings that a full panel of diagnostic biomarkers, including CK7, CK20, CK5/6, P63, p40, TTF1, and Napsin A, is routinely available in both private and public South African hospitals [53]. While the testing of diagnostic biomarkers is readily available and routinely performed in both private and public healthcare settings, access to predictive biomarker testing is far more limited [53].

The distribution of diagnostic and predictive biomarkers indicates that diagnostic markers are tested more frequently, whereas predictive biomarkers have a disproportionately high proportion of reports indicating that the biomarkers were either not mentioned or tested, which may suggest limited

testing rates (Figure 7). The lower prevalence of predictive biomarkers such as PD-L1 and EGFR when compared to diagnostic biomarkers in our findings is likely a consequence of their low testing rates and this may indicate gaps in biomarker testing or the absence of standardised testing guidelines in our country (Figure 8). This indicates that many NSCLC patients may not be receiving molecular profiling necessary for targeted therapy and immunotherapy. This disparity is particularly evident in the public sector, where testing for oncogenic drivers is rarely conducted due to the lack of funding for targeted therapies [53].

Testing rates in private hospitals were consistently higher across both diagnostic and predictive biomarkers (TTF1: 25.80% vs. 19.10%, EGFR: 3.80% vs. 0.05%**),** suggesting that patients in the public sector may have limited access to molecular profiling, influencing their treatment options and clinical outcomes (Table 5). The low testing rates for emerging predictive biomarkers such as KRAS (0.52%), BRAF (0.57%), and PIK3CA (0.33%) are not unexpected given that even well-established predictive biomarkers like EGFR and PD-L1 remain underutilized in routine clinical practice. As a result, the low prevalence of these biomarkers in this study may reflect testing limitations rather than their true incidence in the population.

It is evident that private hospitals dominate in biomarker testing, while public hospitals contribute a significantly smaller proportion (Figure 9). However, although such testing is available in the private sector, it is largely restricted to patients covered by top-tier medical aid schemes, limiting access for many individuals who may benefit from targeted therapies [53]. This inequality highlights the socioeconomic barriers to precision medicine in SA, where access to advanced diagnostic tools remains largely dependent on financial means rather than clinical need. Addressing these inequities will require policy-driven efforts to expand biomarker testing in the public sector, ensuring that all NSCLC patients, regardless of socioeconomic status, have access to comprehensive molecular profiling and targeted treatment options.

6.2 Conclusion and Recommendation

This study examined demographic patterns, biomarker extraction and validation, and disparities in NSCLC diagnosis and molecular testing in SA. It highlights significant disparities in NSCLC diagnosis and molecular testing in the country, with racial and socioeconomic factors influencing

access to care. White patients were disproportionately represented in NSCLC cases, suggesting barriers to early diagnosis among Black, Coloured, and Indian patients, particularly in the public healthcare sector. Expanding diagnostic services, increasing awareness campaigns, and implementing targeted screening programs could help reduce these disparities [21].

The development of an NLP algorithm for biomarker extraction provides a methodological foundation for large-scale pathology report analysis, addressing the scalability limitations of manual extraction approaches. Our findings on biomarker testing in SA can directly inform the development of cost-effective lung cancer diagnostic panels. Our data shows that CK7, TTF1, Napsin A, P63, and PDL1 were the most prevalent in the local context. These patterns provide a foundation for creating a SA-specific panel of biomarkers. Research by Blanco-Prieto et al. showed that a targeted panel approach to biomarker testing could achieve high sensitivity (97%) for lung cancer detection, particularly for early-stage disease [78]. However, our biomarker prevalence calculations are based only on cases where testing was performed, which may introduce bias on those who were tested. Patients selected for biomarker testing likely represent cases with higher clinical suspicion or specific tumour characteristics, potentially resulting in prevalence estimates that do not reflect the broader NSCLC population.

We therefore recommend implementing a diagnostic panel that combines these five key biomarkers (CK7, TTF1, Napsin A, P63, and PDL1) to improve both diagnostic accuracy and cost-efficiency compared to individual testing methods. The demographic patterns identified in our study could guide prioritization strategies, ensuring resources are directed toward patients most likely to benefit from comprehensive testing. This panel-based approach represents a practical solution to resource constraints while addressing disparities in molecular testing access between private and public sectors, potentially improving early detection rates in historically underserved populations across SA.

Bibliography

[1] K Farzam, VS Tivakaran, and Prolonging Drugs QT. Statpearls publishing. Treasure Island (FL: 2021 Jan.[Internet] PMID: 30335315.[Google Scholar], 2023.

[2] Sean Blandin Knight, Phil A Crosbie, Haval Balata, Jakub Chudziak, Tracy Hussell, and Caroline Dive.Progress and prospects of early detection in lung cancer. Open biology, 7(9):170070, 2017.

[3] Cecilia Zappa and Shaker A Mousa. Non-small cell lung cancer: current treatment and future advances.Translational lung cancer research, 5(3):288, 2016.

[4] Narjust Duma, Rafael Santana-Davila, and Julian R Molina. Non–small cell lung cancer: epidemiology,screening, diagnosis, and treatment. In Mayo Clinic Proceedings, volume 94, pages 1623–1640.Elsevier, 2019.

[5] Ekaterina A Semenova, Remco Nagel, and Anton Berns. Origins, genetic landscape, and emerging therapies of small cell lung cancer. Genes & development, 29(14):1447–1462, 2015.

[6] Charles M Rudin, Elisabeth Brambilla, Corinne Faivre-Finn, and Julien Sage. Small-cell lung cancer.Nature Reviews Disease Primers, 7(1):3, 2021.

[7] Krishna Chaitanya Thandra, Adam Barsouk, Kalyan Saginala, John Sukumar Aluru, and Alexander Barsouk. Epidemiology of lung cancer. Contemporary Oncology/Wsp´ol-czesna Onkologia,25(1):45–52, 202

[8] Tang MS, Lee HW, Weng MW, Wang HT, Hu Y, Chen LC, Park SH, Chan HW, Xu J, Wu XR, Wang H. DNA damage, DNA repair and carcinogenicity: Tobacco smoke versus electronic cigarette aerosol. Mutation Research/Reviews in Mutation Research. 2022 Jan 1;789:108409.

[9] Maria Lorenzo-Gonzalez, Alberto Ruano-Ravina, Maria Torres-Duran, Mariano Provencio-Pulla, Karl Kelsey, Isaura Parente-Lamelas, Iria Vidal-Garcia, Virginia Leiro-Fernandez, Cristina Martinez, Jesus Hernandez, et al. Lung cancer risk and do-it-yourself activities. a neglected risk factor for lung cancer. Environmental research, 179:108812, 2019.

[10] Mariona Riudavets, Marta Garcia de Herreros, Benjamin Besse, and Laura Mezquita. Radon and lung cancer: current trends and future perspectives. Cancers, 14(13):3142, 2022.

[11] Steven B Markowitz, Stephen M Levin, Albert Miller, and Alfredo Morabia. Asbestos, asbestosis,smoking, and lung cancer. new findings from the north american insulator cohort. American journal of respiratory and critical care medicine, 188(1):90–96, 2013.

[12] Jolanta Lissowska, Lenka Foretova, Joanna Dabek, David Zaridze, Neonila Szeszenia-Dabrowska, Peter Rudnai, Eleonora Fabianova, Adrian Cassidy, Dana Mates, Vladimir Bencko, et al. Family history and lung cancer risk: international multicentre case–control study in eastern and central Europe and meta-analyses. Cancer causes & control, 21:1091–1104, 2010

[13] Michele L Cot´e, Mei Liu, Stefano Bonassi, Monica Neri, Ann G Schwartz, David C Christiani, Margaret R Spitz, Joshua E Muscat, Gad Rennert, Katja K Aben, et al. Increased risk of lung cancer in individuals with a family history of the disease: a pooled analysis from the international lung cancer consortium. European journal of cancer, 48(13):1957–1968, 2012

[14] Viviane Teixeira L de Alencar, Amanda B Figueiredo, Marcelo Corassa, Kenneth J Gollob, and Vladmir C Cordeiro de Lima. Lung cancer in never smokers: Tumor immunology and challenges for immunotherapy. Frontiers in Immunology, 13:984349, 2022

[15] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2020. CA: A Cancer Journal for Clinicians, 70(1):7–30, 2020

[16] Lisa Carter-Harris. Lung cancer stigma as a barrier to medical help-seeking behavior: Practice implications. Journal of the American Association of Nurse Practitioners, 27(5):240–245, 201

[17] C Friedemann Smith, KL Whitaker, K Winstanley, and J Wardle. Smokers are less likely than non-smokers to seek help for a lung cancer ‘alarm’symptom. Thorax, 71(7):659–661, 2016.

[18] Mohamad M Saab, Serena FitzGerald, Brendan Noonan, Caroline Kilty, Abigail Collins, A´ine Lyng, Una Kennedy, Maidy O’Brien, and Josephine Hegarty. Promoting lung cancer awareness, help-seeking and early detection: a systematic review of interventions. Health promotion international, 36(6):1656–1671, 2021.

[19] E Power and J Wardle. Change in public awareness of symptoms and perceived barriers to seeing a doctor following be clear on cancer campaigns in england. British Journal of Cancer, 112(1): S22–S26,2015.

[20] L Ironmonger, E Ohuma, N Ormiston-Smith, Carolynn Gildea, CS Thomson, and MD Peake. An evaluation of the impact of large-scale interventions to raise public awareness of a lung cancer symptom. British journal of cancer, 112(1):207–216, 2015.

[21] Masahiro Kaneko, Kenji Eguchi, Hironobu Ohmatsu, Ryutaro Kakinuma, Tuguo Naruke, Keiichi Suemasu, and Noriyuki Moriyama. Peripheral lung cancer: screening and detection with low-dose spiral ct versus radiography. Radiology, 201(3):798–802, 1996

[22] Claudia I Henschke. Early lung cancer action project: overall design and findings from baseline screening. Cancer, 89(S11):2474–2482, 2000.

[23] National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine, 365(5):395–409, 2011.

[24] Alexandre Sadate, Bob V Occean, Jean-Paul Beregi, Aymeric Hamard, Takieddine Addala, Helene de Forges, Pascale Fabbro-Peray, and Julien Frandon. Systematic review and meta-analysis on the impact of lung cancer screening by low-dose computed tomography. European Journal of Cancer 134:107–114, 2020.

[25] Chih-Yang Huang, Da-Tong Ju, Chih-Fen Chang, P Muralidhar Reddy, and Bharath Kumar Velmurugan. A review on the effects of current chemotherapy drugs and natural agents in treating non–small cell lung cancer. Biomedicine, 7(4):23, 2017

[26] Amanda McIntyre and Apar Kishor Ganti. Lung cancer—a global perspective. Journal of surgical oncology, 115(5):550–554, 2017

[27] Zafar A, Khatoon S, Khan MJ, Abu J, Naeem A. Advancements and limitations in traditional anti-cancer therapies: a comprehensive review of surgery, chemotherapy, radiation therapy, and hormonal therapy. Discover oncology. 2025 Apr 24;16(1):607.

[28] KM Monirul Islam, Xiaqing Jiang, Trisari Anggondowati, Ge Lin, and Apar Kishor Ganti. Comorbidity and survival in lung cancer patients. Cancer Epidemiology, Biomarkers & Prevention, 24(7):1079–1085, 2015.

[29] Parvarish Sharma, Meenu Mehta, Daljeet Singh Dhanjal, Simran Kaur, Gaurav Gupta, Harjeet Singh, Lakshmi Thangavelu, S Rajeshkumar, Murtaza Tambuwala, Hamid A Bakshi, et al. Emerging trends in the novel drug delivery approaches for the treatment of lung cancer. Chemico-biological interactions, 309:108720, 2019.

[30] Šutić M, Vukić A, Baranašić J, Försti A, Džubur F, Samaržija M, Jakopović M, Brčić L, Knežević J. Diagnostic, predictive, and prognostic biomarkers in non-small cell lung cancer (NSCLC) management. Journal of personalized medicine. 2021 Oct 27;11(11):1102.

[31] Peter J Mazzone, Catherine Rufatto Sears, Doug A Arenberg, Mina Gaga, Michael K Gould, Pierre P Massion, Vish S Nair, Charles A Powell, Gerard A Silvestri, Anil Vachani, et al. Evaluating molecular biomarkers for the early detection of lung cancer: when is a biomarker ready for clinical use? an official american thoracic society policy statement. American journal of respiratory and critical care medicine, 196(7):e15–e29, 2017.

[32] Coenraad FN Koegelenberg, Shane Dorfman, Ivan Schewitz, Guy A Richards, Shaun Maasdorp,Clifford Smith, Keertan Dheda, et al. Recommendations for lung cancer screening in southern africa.Journal of Thoracic Disease, 11(9):3696, 2019

[33] Singh E, Underwood JM, Nattey C, Babb C, Sengayi M, Kellett P. South African National Cancer Registry: Effect of withheld data from private health systems on cancer incidence estimates. S Afr Med J. 2015;105(2):107-109. PubMed

[34] Somdyala NIM, Mbuthini L, Müller B, Sithole N, Ncinitwa A, Bradshaw D. Active case-finding method improves completeness and accuracy of data reported to the rural Eastern Cape Cancer Registry in South Africa. ecancermedicalscience. 2021;15:1251

[35] Chen, Y., Hao, L., Zou, V.Z. et al. Automated medical chart review for breast cancer outcomes research: a novel natural language processing extraction system. BMC Med Res Methodol 22, 136 (2022). https://doi.org/10.1186/s12874-022-01583-z

[36] Agnieszka Mykowiecka, Mal-gorzata Marciniak, and Anna Kup´s´c. Rule-based information extraction from patients' clinical data. Journal of biomedical informatics, 42(5):923–936, 2009.

[37] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71 (3):209–249, 2021.

[38] Joanna Domagala-Kulawik and Anna Trojnar. Lung cancer in women in 21th century. Journal of Thoracic Disease, 12(8):4398, 2020.

[39] Rabab Gaafar. Lung cancer in africa: challenges and perspectives. Journal of Thoracic Oncology, 12(1):S115–S116, 2017.

[40] Amanda Leiter, Rajwanth R Veluswamy, and Juan P Wisnivesky. The global burden of lung cancer: current status and future trends. Nature reviews Clinical oncology, 20(9):624–639, 2023.

[41] Julie A Barta, Charles A Powell, and Juan P Wisnivesky. Global epidemiology of lung cancer. Annals of global health, 85(1):8, 2019.

[42] Heiko Becher and Volker Winkler. Lung cancer mortality in sub-saharan africa. International Journal of Cancer, 129(6):1537, 2011.

[43] Abhinav Urman, Shilpa Josyula, Ariella Rosenberg, David Lounsbury, Thomas Rohan, and H Dean Hosgood. Burden of lung cancer and associated risk factors in africa by region. Journal of Pulmonary Respiratory Medicine, 6(340):1–5, 2016

[44] Morteza Keikha and Behzad Nazari Esfahani. The relationship between tuberculosis and lung cancer. Advanced Biomedical Research, 7(1):58, 2018.

[45] Tawanda A Mazhindu, N Ndlovu, E Matsikidze, S Chibonda, and W Kadzatsa. Pulmonary tuberculosis: A comorbidity or misdiagnosis of primary lung cancer in africa? In Proceedings, volume 100, page 2.MDPI, 2024.[46] Malini Surapaneni and Dipesh Uprety. Lung cancer management in low and middle-income countries-current challenges and potential solutions. International Journal of Cancer Care and Delivery, 3(1), 2023.

[47] LV Sequist, RS Heist, AT Shaw, P Fidias, R Rosovsky, JS Temel, IT Lennes, S Digumarthy, BA Waltman, E Bast, et al. Implementing multiplexed genotyping of non-small-cell lung cancers into routine clinical practice. Annals of oncology, 22(12):2616–2624, 2011.

[48] MG Kris, BE Johnson, DJ Kwiatkowski, AJ Iafrate, II Wistuba, SL Aronson, JA Engelman, Y Shyr, FR Khuri, CM Rudin, et al. Identification of driver mutations in tumor specimens from 1,000 patients with lung adenocarcinoma: The nci's lung cancer mutation consortium (lcmc). Journal of Clinical Oncology, 29(18 suppl):CRA7506–CRA7506, 2011.

[49] Pamela Villalobos and Ignacio I Wistuba. Lung cancer biomarkers. Hematology/oncology clinics of North America, 31(1):13, 2017.

[50] Tristan Lee, Jeffrey M Clarke, Deepali Jain, Sendhilnathan Ramalingam, and Vishal Vashistha. Precision treatment for metastatic non-small cell lung cancer: A conceptual overview. Cleveland Clinic Journal of Medicine, 88(2):117–127, 2021

[51] D Planchard, S Popat, K Kerr, S Novello, EF Smit, C Faivre-Finn, TS Mok, M Reck, PE Van Schil, MD Hellmann, et al. Correction to:"metastatic non-small cell lung cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up". Annals of Oncology, 30(5): 863–870, 2019.

[52] National Cancer Registry. Cancer in south africa. 2022 full report., 2022

[53] Ronwyn Van Eeden, Mariza Tunmer, Agneta Geldenhuys, Simon Nayler, and Bernardo Leon Rapoport. Lung cancer in south africa. Journal of Thoracic Oncology, 15(1):22–28, 2020

[54] Coenraad FN Koegelenberg, Shane Dorfman, Ivan Schewitz, Guy A Richards, Shaun Maasdorp, Clifford Smith, Keertan Dheda, et al. Recommendations for lung cancer screening in southern africa. Journal of Thoracic Disease, 11(9):3696, 2019.

[55] Cornelius Nattey, Jimi Fadahun, Briamoh Bello, Danuta Kielkowski, and Gill Nelson. Trends in lung cancer mortality in south africa 1995–2008. Occupational and Environmental Medicine, 68 (Suppl 1):A111–A111, 2011.

[56] Winnie T Maphumulo and Busisiwe R Bhengu. Challenges of quality improvement in the healthcare of south africa post-apartheid: A critical review. Curationis, 42(1):1–9, 2019.

[57] Andy Gray and Yousuf Vawda. Health legislation and policy. South African health review, 2018 (1):1–11, 2018.

[58] Linda Hammami, Alessia Paglialonga, Giancarlo Pruneri, Michele Torresani, Milena Sant, Carlo Bono, Enrico Gianluca Caiani, and Paolo Baili. Automated classification of cancer morphology from italian pathology reports using natural language processing techniques: A rule-based approach. Journal of Biomedical Informatics, 116:103712, 2021

[59] Han HW, Yang SY, Song G, Sa S, Chung GE, Seo JY, Jin EH, Kim H, An D. Natural language processing for assessing quality indicators in free-text colonoscopy and pathology reports: development and usability study. JMIR Medical Informatics. 2022 Apr 15;10(4):e35257.

[60] Ayhan Arısoy. Natural language processing algorithms and performance comparison. Yalva¸c Akademi Dergisi, 9(2):106–121.

[61] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: state of the art, current trends and challenges. Multimedia tools and applications, 82(3):3713–3744, 2023

[62] Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. Natural language processing for smart healthcare. IEEE Reviews in Biomedical Engineering, 17:4–18, 2022.

[63] Amy Harris Nordo, Hugh P Levaux, Lauren B Becnel, Jose Galvez, Prasanna Rao, Komathi Stem, Era Prakash, and Rebecca Daniels Kush. Use of ehrs data for clinical research: historical progress and current applications. Learning health systems, 3(1):e10076, 2019

[64] Agnieszka Mykowiecka, Mal-gorzata Marciniak, and Anna Kup´s´c. Rule-based information extraction from patients' clinical data. Journal of biomedical informatics, 42(5):923–936, 2009.

[65] Chengtai Li, Yiming Zhang, Ying Weng, Boding Wang, and Zhenzhu Li. Natural language processing applications for computer-aided diagnosis in oncology. Diagnostics, 13(2):286, 2023.

[66] Achilonu OJ, Singh E, Nimako G, Eijkemans RM, Musenge E. Rule-based information extraction from free-text pathology reports reveals trends in South African female breast cancer molecular subtypes and Ki67 expression. BioMed Research International. 2022;2022(1):6157861.

[67] Cassim N, Mapundu M, Olago V, Celik T, George JA, Glencross DK. Using text mining techniques to extract prostate cancer predictive information (Gleason score) from semi-structured narrative laboratory reports in the Gauteng province, South Africa. BMC Medical Informatics and Decision Making. 2021 Nov 25;21(1):330.

[68] Fabio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: pretrained bert models for brazilian portuguese. In Brazilian conference on intelligent systems, pages 403–417. Springer, 2020.

[69] S Sathyanarayanan and B Roopashri Tantri. Confusion matrix-based performance evaluation metrics. African Journal of Biomedical Research, pages 4023–4031, 2024

[70] David J Myers and Jason M Wallen. Lung adenocarcinoma. In StatPearls [Internet]. StatPearls Publishing, 2023

[71] Seth Kligerman and Charles White. Epidemiology of lung cancer in women: risk factors, survival, and screening. American journal of roentgenology, 196(2):287–295, 201

[72] Munzone E, Marra A, Comotto F, Guercio L, Sangalli CA, Lo Cascio M, Pagan E, Sangalli D, Bigoni I, Porta FM, D'Ercole M. Development and validation of a natural language processing algorithm for extracting clinical and pathological features of breast cancer from pathology reports. JCO Clinical Cancer Informatics. 2024 Aug;8:e2400034.

[73] Bernhard Waltl, Georg Bonczek, and Florian Matthes. Rule-based information extraction: Advantages, limitations, and perspectives. Jusletter IT (02 2018), 4, 2018.

[74] Venet Osmani. Natural language processing of clinical notes on chronic diseases: Systematic review. (No Title), 2019.

[75] Hyojin Kim and Jin-Haeng Chung. Biomarker testing of cytology specimens in personalized medicine for lung cancer patients. Journal of Pathology and Translational Medicine, 56(6): 326–333, 2022.

[76] Yunfan Ma, Mengying Fan, Liang Dai, Xiaozheng Kang, Yiqiang Liu, Yu Sun, Wanpu Yan, Zhen Liang, Hongchao Xiong, and Keneng Chen. The expression of ttf-1 and napsin a in early-stage lung adenocarcinoma correlates with the results of surgical treatment. Tumor Biology, 36:8085–8092, 2015

[78] Sonia Blanco-Prieto, Loretta De Chiara, Mar Rodr´ıguez-Girondo, Lorena V´azquez-Iglesias, Francisco Javier Rodr´ıguez-Berrocal, Alberto Fern´andez-Villar, Mar´ıa Isabel Botana-Rial, and Mar´ıa P´aez de La Cadena. Highly sensitive marker panel for guidance in lung cancer rapid diagnostic units. Scientific Reports, 7(1):41151, 2017

6 Appendices

6.1 Plagiarism Declaration

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

FACULTY OF HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

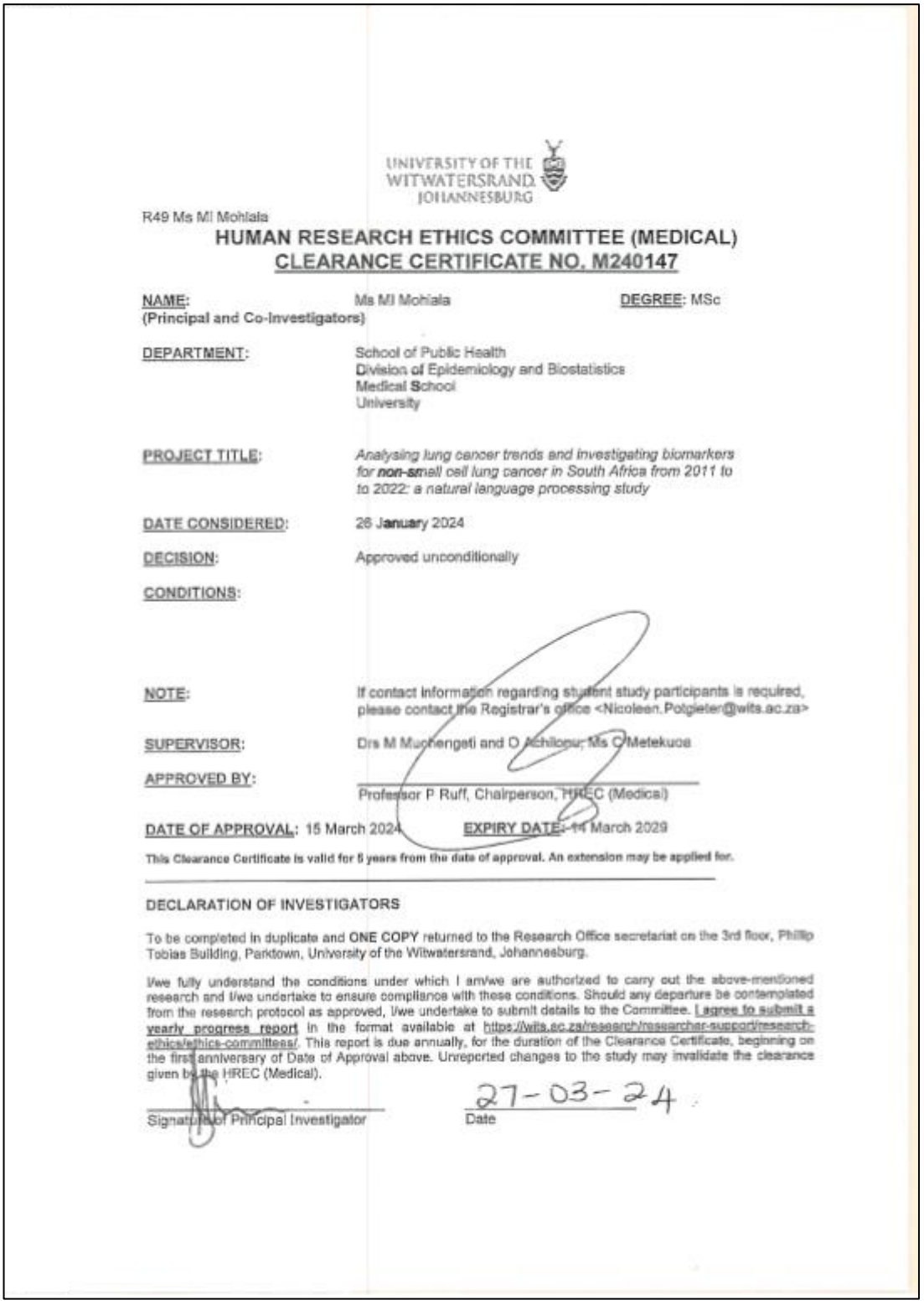
I MATSHEDISO IVY MOHLALA (Student number: 972315) am a student registered for the degree of MSc. Epidemiology(PHI) in the academic year 2023.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: _____ Date: 23 March 2025

6.2 Ethics Clearance Certificate



UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R49 Ms MI Mohlala

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M240147

NAME: Ms MI Mohlala **DEGREE:** MSc
(Principal and Co-Investigators)

DEPARTMENT: School of Public Health
Division of Epidemiology and Biostatistics
Medical School
University

PROJECT TITLE: *Analysing lung cancer trends and investigating biomarkers for non-small cell lung cancer in South Africa from 2011 to to 2022: a natural language processing study*

DATE CONSIDERED: 26 January 2024

DECISION: Approved unconditionally

CONDITIONS:

NOTE: If contact information regarding student study participants is required, please contact the Registrar's office <Nicoleen.Potgieter@wits.ac.za>

SUPERVISOR: Drs M Muchengeti and O Achilonu; Ms C Metekuce

APPROVED BY:
Professor P Ruff, Chairperson, HREC (Medical)

DATE OF APPROVAL: 15 March 2024 **EXPIRY DATE:** 14 March 2029

This Clearance Certificate is valid for 5 years from the date of approval. An extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office secretariat on the 3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to submit details to the Committee. I agree to submit a yearly progress report in the format available at https://wits.ac.za/research/researcher-support/research-ethics/ethics-committees/. This report is due annually, for the duration of the Clearance Certificate, beginning on the first anniversary of Date of Approval above. Unreported changes to the study may invalidate the clearance given by the HREC (Medical).

Signature of Principal Investigator

27-03-24
Date

6.3 Journal guidelines for authors

The full instructions are found here: Research article | BMC Cancer

Preparing your manuscript

The information below details the section headings that you should include in your manuscript and what information should be within each section.

Please note that your manuscript must include a 'Declarations' section including all of the subheadings (please see below for more information).

Title page

The title page should:

- present a title that includes, if appropriate, the study design e.g.:
  - "A versus B in the treatment of C: a randomized controlled trial", "X is a risk factor for Y: a case control study", "What is the impact of factor X on subject Y: A systematic review"
  - or for non-clinical or non-research studies a description of what the article reports
- list the full names and institutional addresses for all authors
  - if a collaboration group should be listed as an author, please list the Group name as an author. If you would like the names of the individual members of the Group to be searchable through their individual PubMed records, please include this information in the "Acknowledgements" section in accordance with the instructions below
  - Large Language Models (LLMs), such as ChatGPT, do not currently satisfy our authorship criteria. Notably an attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs. Use of an LLM should be properly documented in the Methods section (and if a Methods section is not available, in a suitable alternative part) of the manuscript. • indicate the corresponding author

Abstract

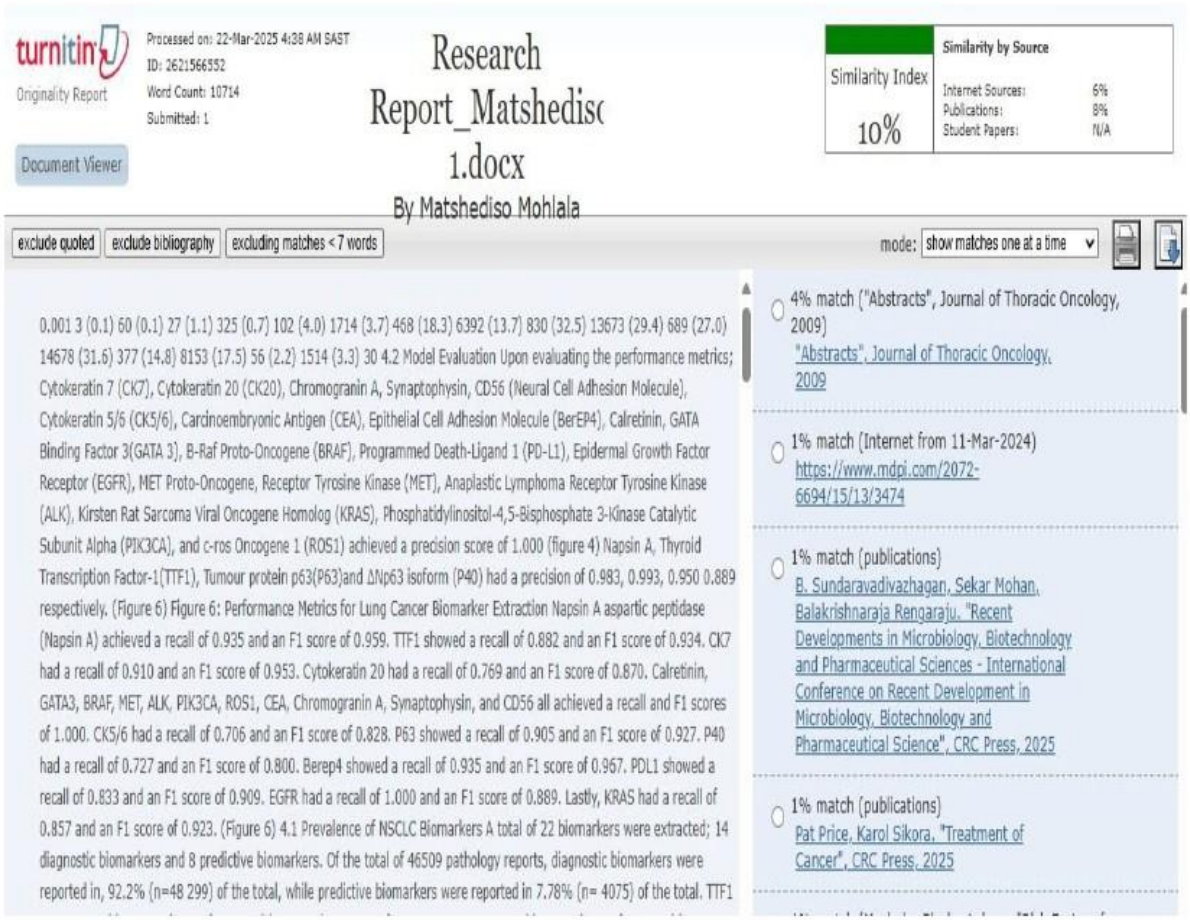
The Abstract should not exceed 350 words. Please minimize the use of abbreviations and do not cite references in the abstract. Reports of randomized controlled trials should follow the CONSORT extension for abstracts. The abstract must include the following separate sections:

- **Background:** the context and purpose of the study
- **Methods:** how the study was performed and statistical tests used
- **Results:** the main findings
- **Conclusions:** brief summary and potential implications
- **Trial registration:** If your article reports the results of a health care intervention on human participants, it must be registered in an appropriate registry and the registration number and date of registration should be stated in this section. If it was not registered prospectively (before enrollment of the first participant), you should include the words 'retrospectively registered'. See our editorial policies for more information on trial registration

Keywords

Three to ten keywords representing the main content of the article.

6.4 Turnitin Report



turnitin Originality Report

Processed on: 22-Mar-2025 4:38 AM SAST
ID: 2621566552
Word Count: 10714
Submitted: 1

Document Viewer

Research Report_Matshediso 1.docx

By Matshediso Mohlala

Similarity Index 10%

Similarity by Source
Internet Sources: 6%
Publications: 8%
Student Papers: N/A

exclude quoted | exclude bibliography | excluding matches < 7 words

mode: show matches one at a time

0.001 3 (0.1) 60 (0.1) 27 (1.1) 325 (0.7) 102 (4.0) 1714 (3.7) 468 (18.3) 6392 (13.7) 830 (32.5) 13673 (29.4) 689 (27.0) 14678 (31.6) 377 (14.8) 8153 (17.5) 56 (2.2) 1514 (3.3) 30 4.2 Model Evaluation Upon evaluating the performance metrics; Cytokeratin 7 (CK7), Cytokeratin 20 (CK20), Chromogranin A, Synaptophysin, CD56 (Neural Cell Adhesion Molecule), Cytokeratin 5/6 (CK5/6), Carcinoembryonic Antigen (CEA), Epithelial Cell Adhesion Molecule (BerEP4), Calretinin, GATA Binding Factor 3(GATA 3), B-Raf Proto-Oncogene (BRAF), Programmed Death-Ligand 1 (PD-L1), Epidermal Growth Factor Receptor (EGFR), MET Proto-Oncogene, Receptor Tyrosine Kinase (MET), Anaplastic Lymphoma Receptor Tyrosine Kinase (ALK), Kirsten Rat Sarcoma Viral Oncogene Homolog (KRAS), Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha (PIK3CA), and c-ros Oncogene 1 (ROS1) achieved a precision score of 1.000 (figure 4) Napsin A, Thyroid Transcription Factor-1(TTF1), Tumour protein p63(P63)and ΔNp63 isoform (P40) had a precision of 0.983, 0.993, 0.950 0.889 respectively. (Figure 6) Figure 6: Performance Metrics for Lung Cancer Biomarker Extraction Napsin A aspartic peptidase (Napsin A) achieved a recall of 0.935 and an F1 score of 0.959. TTF1 showed a recall of 0.882 and an F1 score of 0.934. CK7 had a recall of 0.910 and an F1 score of 0.953. Cytokeratin 20 had a recall of 0.769 and an F1 score of 0.870. Calretinin, GATA3, BRAF, MET, ALK, PIK3CA, ROS1, CEA, Chromogranin A, Synaptophysin, and CD56 all achieved a recall and F1 scores of 1.000. CK5/6 had a recall of 0.706 and an F1 score of 0.828. P63 showed a recall of 0.905 and an F1 score of 0.927. P40 had a recall of 0.727 and an F1 score of 0.800. Berep4 showed a recall of 0.935 and an F1 score of 0.967. PDL1 showed a recall of 0.833 and an F1 score of 0.909. EGFR had a recall of 1.000 and an F1 score of 0.889. Lastly, KRAS had a recall of 0.857 and an F1 score of 0.923. (Figure 6) 4.1 Prevalence of NSCLC Biomarkers A total of 22 biomarkers were extracted; 14 diagnostic biomarkers and 8 predictive biomarkers. Of the total of 46509 pathology reports, diagnostic biomarkers were reported in, 92.2% (n=48 299) of the total, while predictive biomarkers were reported in 7.78% (n= 4075) of the total. TTF1

4% match ("Abstracts", Journal of Thoracic Oncology, 2009)
"Abstracts", Journal of Thoracic Oncology, 2009

1% match (Internet from 11-Mar-2024)
https://www.mdpi.com/2072-6694/15/13/3474

1% match (publications)
B. Sundaravadivazhagan, Sekar Mohan, Balakrishnaraja Rengaraju. "Recent Developments in Microbiology, Biotechnology and Pharmaceutical Sciences - International Conference on Recent Development in Microbiology, Biotechnology and Pharmaceutical Science", CRC Press, 2025

1% match (publications)
Pat Price, Karol Sikora. "Treatment of Cancer", CRC Press, 2025

Signatures by Supervisors:

Dr. Mazvita Muchengeti: .....

Dr. Okechinyere Achilonu

Ms. Carole Metekoua: