The Association between Serum Cancer Antigen 125 (CA 125) and Risk of Lung Cancer in Females: Assessing the Possibilities for Early Detection

By

Joanna (Asia) Przepiórkowski, B.PH (Honours)

In partial fulfillment of the requirements for the degree of
Master of Science in Applied Health Sciences
(Health Sciences)

Faculty of Applied Health Sciences
Brock University, St. Catharines, Ontario, Canada

ABSTRACT

Background: Few studies have closely examined the relationship between CA 125 and lung cancer. This study is expected to provide more understanding about CA 125 and its role as a potential predictor for lung cancer risk.

Objectives: To evaluate: i) the association between CA 125 and lung cancer; ii) if the associations differ by potential effect modifier (smoking status); and iii) if the association between CA 125 and lung cancer differs by lung cancer stage (early vs. advanced).

Methods: The present research was conducted using secondary data from the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) randomized controlled trial (RCT). The associations between explanatory variables and lung cancer were evaluated using multivariable logistic regression. Each multivariable logistic regression model was adjusted for age, education, current body mass index (BMI), family history of lung cancer, personal history of cancer, chronic obstructive pulmonary disease (COPD), average number of cigarettes smoked per day and number of years smoked.

Results: The study demonstrated that CA 125 is significantly and independently associated with lung cancer and that CA 125 is associated with early-stage lung cancer. It was found that an elevated CA 125 level was associated with a higher risk of lung cancer in individuals who smoked. Although the study demonstrated promising results, CA 125 did not have a large effect on the study's lung cancer risk prediction models.

Conclusion: CA 125 is not a strong enough predictor to be used as an indicator in lung cancer screening alone, however it may be useful in a panel of complimentary biomarkers. Future research is needed to explore whether a panel of complimentary biomarkers including CA 125 can improve lung cancer risk prediction.

KEY WORDS: Lung cancer, CA 125, public health, lung cancer risk prediction

ACKNOWLEDGEMENTS

I would like to begin by thanking my supervisor and most importantly, my mentor, Dr. Martin Tammemägi. I would not be where I am today without your continuous support and guidance. Thank you for always believing in me and for guiding me every step of the way. I am forever grateful for all the countless hours you have spent with me over Zoom reviewing my thesis and for always having the time to answer my questions, no matter how big or small. Thank you for taking me under your wing, and for allowing me to be one of your graduate students. I am forever grateful for the experience. You have always been and continue to be my inspiration for becoming a great Epidemiologist.

I would like to extend my gratitude to my committee members, Dr. Brent Faught and Dr. Joanne Crawford. Thank you for taking the time out of your busy schedules to help guide me through my thesis. I know it has been a long journey and I greatly appreciate all of the help and support along the way. To my External Examiner, Dr. Joseph Tota and my Thesis Defence Chair, Dr. Litsa Tsiani. Thank you for all your support and encouragement, I am so grateful to have had such a supportive and positive team during my thesis defence. I would also like to thank Dr. Deborah O'Leary, who always had the time to talk to me whenever I had a doubt in my mind, thank you for making me feel so welcome and supported in the Faculty of Applied Health Sciences.

Finally, I would like to thank my support system, Ala, Darek, Ela, Dorotka, and Łucja. You have never failed to stop believing in me and have always encouraged me every step of the way, no matter how challenging it was. I am forever grateful for the time you spent rooting for me and for always lending an ear whenever things got tough! Thank you from the bottom of my heart – *Dziękuję z całego serca.*

Table of Contents

LIST OF TABLES ..... ***vii***

LIST OF FIGURES ..... ***viii***

CHAPTER I - INTRODUCTION ***1***

1.2 Lung Cancer Mechanisms **1**

1.3 Literature Gaps **3**

1.4 Response to Gaps in Current Knowledge **3**

1.5 Study Aims, Objectives, Hypotheses **4**

1.6 Conclusion **6**

CHAPTER II - LITERATURE REVIEW ***7***

2.1 Introduction **7**

2.2 Lung Cancer Statistics **7**

2.2.1 Worldwide Statistics 7

2.2.2 Canadian Statistics 7

2.2.3 US Statistics 8

2.3 Lung Cancer Biology **9**

2.3.1 Etiology & Pathogenesis 9

2.3.2 Histology 10

2.4 Tumor Staging **12**

2.4.1 Tumour-Node-Metastasis Cancer Staging System 12

2.4.2 Stage Grouping 14

2.4.3 Tumour Grade 16

2.5 Clinical Features **16**

2.5.1 Signs and Symptoms 16

2.6 Diagnosis **17**

2.7 Treatment and Prognosis **17**

2.8 Lung Cancer Risk Factors **18**

2.8.1 Modifiable Risk Factors 18

2.8.2 Non-Modifiable Risk Factors 21

2.9 CA 125 ..... **25**

2.9.1 CA 125 and Ovarian Cancer Screening 26

2.9.1 CA 125 and Lung Cancer 29

CHAPTER III - METHODS ***31***

3.1 Introduction **31**

3.2 Source Data – PLCO Trial **31**

3.2.1 Study Sample Data 31

3.2.2 Inclusion & Exclusion Criteria 32

3.2.3 Randomization and Screening Process 33

3.2.4 Diagnostic and Therapeutic Follow-Up 35

3.3 - Modeling and Analysis 35
3.3.1 Statistical Approach 35
3.3.2 Data Preparation 36
3.3.3 Descriptive Statistics, Univariate Analysis and Exploratory Analysis 37
3.3.4 Model Building Strategy 40

3.4 - Model Evaluation 41
3.4.1 Fit Diagnostics 41
3.4.2 Influential Observations 42

CHAPTER IV - RESULTS 43

4.1 Data Preparation 43

4.2 Sample Characteristics 43

4.3 Predictor Associations with Lung Cancer 48
4.3.1 Sociodemographic 48
4.3.2 Medical History 48
4.3.3 Exposures 48

4.4 Key Study Findings 52

4.4.1 Summary of Tables 52
4.4.2 Summary of Predictor Variable Findings for Lung Cancer~CA 125 55
4.4.3 Smoking Status – Individuals who smoked vs. individuals who never smoked among total study sample 57
4.4.4 Early Stage vs. Advanced Stage Lung Cancer among Total Study Sample 61
4.4.5 Prediction Models 64

4.5 Assumption Checking 66
4.5.1 Independence of Errors 66
4.5.2 Assessment of Non-Linear Associations 66
4.5.3 Assessment of Collinearity 66
4.5.4 Lack of Influential Outliers 66
4.6 Fit Diagnostics 68

CHAPTER 5 - DISCUSSION 69

5.1 Main Findings 69
5.1.1 Smoking Status, CA 125 and Lung Cancer 70
5.1.2 Early-stage Lung Cancer and CA 125 71
5.1.3 Sex Differences among CA 125 71
5.1.4 Exploratory Analysis 72

5.2 Impact on Public Health 72

5.3 Limitations 73

5.4 Strengths 74

5.5 Future Direction for CA 125 Research 75

5.5 Conclusion 75

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
AJCC	American Joint Committee on Cancer
ASIR	Age Standardized Incidence Rate
BAP	Benzo[a]pyrene
BIC	Bayesian Information Criterion
BMI	Body Mass Index
CA 125	Cancer Antigen 125
CA 19-9	Carbohydrate antigen 19-9
CEA	Carcinoembryonic antigen (CEA)
CI	Confidence Interval
COPD	Chronic Obstructive Pulmonary Disease
CTFPHC	Canadian Task Force on Preventative Health Care
CV	Coefficient of Variation
EGFR	Epidermal Growth Factor Receptor
FP	Fractional Polynomials
GOF	Goodness-of-Fit
IARC	International Agency for Research in Cancer
LDCT	Low-dose Helical Computed Tomography
LN	Natural Log Transformed
MFP	Multivariable Fractional Polynomials
NCI	National Cancer Institute

NNK Nicotine-derive Nitrosoaminoketone

NSCLC Non-Small Cell Lung Cancers

OR Odds Ratio

PAHs Polycyclic Aromatic Hydrocarbons

PLCO Prostate, Lung, Colorectal and Ovarian Screening Trial

PPV Positive Predictive Value

RCT Randomized Controlled Trial

ROC Receiver Operating Characteristic Curve

RR Relative Risk

ScC Squamous Cell Carcinoma

SCLC Small Cell Lung Cancer

SHS Second-hand smoking

TNM Tumour-Node-Metastasis

UCLA University of California Los Angeles

UICC Union for International Cancer Control

UK United Kingdom

UKCTOCS UK Collaborative Trial of Ovarian Cancer Screening

US United States of America

USPSTF US Preventive Services Task Force

VIF Variance Inflation Factors

LIST OF TABLES

Table 1.	Tumour-Node-Metastasis (TNM) cancer staging, 7th edition	13
Table 2.	Stage grouping based on TNM cancer staging system	15
Table 3.	Tumour grading system	16
Table 4.	Candidate variables and potential confounders for evaluating associations with lung cancer	38-39
Table 5.	Characteristics of overall participants by lung cancer and univariate logistic associations with lung cancer	45-47
Table 6.	Univariate logistic regression for lung cancer (yes vs no) and CA 125 results by screening rounds (T0-T5)	50
Table 7.	Univariate logistic regression for lung cancer (yes vs no) and log transformed CA 125 levels by screening rounds (T0-T5)	51
Table 8.	Multivariable logistic regression for lung cancer and CA 125 dichotomous results in screening round T3	53
Table 9.	Multivariable logistic regression for lung cancer and LN CA 125 levels in screening round T3	54
Table 10.	Multivariable logistic regression odds ratios for lung cancer and predictors CA 125 dichotomous results and log transformed continuous CA 125 levels, by smoking status	58
Table 11	CA 125 (dichotomous and log transformed continuous) interaction with smoking status in multivariable logistic regression for lung cancer	60
Table 12	Multivariable logistic regression odds ratios for lung cancer and predictors CA 125 dichotomous results by lung cancer stage, early vs advanced	62
Table 13	Multivariable logistic regression odds ratios for lung cancer and predictors log transformed continuous CA 125 levels by lung cancer stage, early vs advanced	63
Table 14	Comparison of ROC curve of prediction models with and without CA 125	64
Table 15	Multivariable linear regression for log transformed continuous CA 125 levels and potential covariates in screening round T3	65
Table 16	Collinearity evaluation for model	67

LIST OF FIGURES

Figure 1.	Possible causal pathways	4
Figure 2.	Odds ratio for lung cancer by dichotomous CA 125 results (positive/negative) by screening rounds T0-T5	56
Figure 3.	Odds ratio for lung cancer by LN CA 125 levels by screening rounds T0-T5	56
Figure 4.	Odds ratio for lung cancer by dichotomous CA 125 results (positive/negative) by screening rounds T0-T5 among individuals who never smoked	59
Figure 5	Odds ratio for lung cancer by LN CA 125 levels by screening rounds T0-T5 among individuals who never smoked	59
Figure 6.	Odds ratio for lung cancer by dichotomous CA 125 results (positive/negative) by screening rounds T0-T5 by early-stage lung cancer	61
Figure 7	Odds ratio for lung cancer by LN CA 125 levels by screening rounds T0-T5 by early-stage lung cancer	62
Figure 8	Inspection of influential observations with Pearson standardized residual in T3 dichotomous CA 125 Results model	67
Figure 9	Inspection of influential observations with Pearson standardized residual in T3 log transformed CA 125 Levels model	68

CHAPTER I - INTRODUCTION

The burden of cancer incidence and mortality is growing quickly worldwide.[1] Lung cancer specifically, is the second most commonly diagnosed cancer in the world, with approximately 2.2 million cases, globally. Lung cancer is also the leading cause of cancer death with an estimated 1.8 million deaths worldwide, annually.[1] Comparable trends can be seen in the United States and Canada. Cancer is the second most commonly diagnosed disease in the United States and the second leading cause of death due to high rates of obesity and cardiovascular disease.[2,3] In Canada, it was estimated that 225,800 new cancers cases and 83,300 cancer deaths were expected in 2020.[4] Among Canadians, lung cancer is the most commonly diagnosed cancer and the leading cause of cancer death.[5] In Canada, the incidence rate and morality rate of lung cancer is 20% and 31% higher in males than females, respectively.[5]

Although the survival rate of lung cancer has been increasing, most lung cancer cases are diagnosed at an advanced stage, making it difficult to successfully treat and cure the disease. In Canada, 70% of lung cancers are diagnosed at a late stage (stage III or IV) annually.[5] Therefore, it is imperative to continue to produce research that may detect lung cancer early and reduce lung cancer morbidity and mortality.

1.2 Lung Cancer Mechanisms

One of the most well-known cause-and-effect oncological relationships is tobacco use and lung cancer. Approximately 15% of the Canadian population continues to smoke on a daily basis.[6] Research has shown that approximately 85% of lung cancer cases are attributed to cigarette smoking.[7] Despite decreased use of tobacco over the past few decades, efforts to control tobacco use

are still needed to further reduce the burden of lung cancer.[8] The effect of tobacco use also has not been comprehensively evaluated in association to biomarkers found within the body. However, it has been found that differences in tumour mutational patterns indicate that the carcinogenetic pathways leading to lung cancer differ between individuals who smoked and individuals who never smoked.[9] In Couraud and colleagues' review on current clinical and molecular aspects of lung cancer in individuals who never smoked, it was suggested that polymorphisms have an effect on one's lung cancer risk.[9] The Kirsten rat sarcoma viral oncogene (KRAS), tumor protein p53 and epidermal growth factor receptor (EGFR) mutations in particular, have been suggested to have differentiating oncogenic mechanisms in lung cancer.[10] KRAS and p53 mutations have been linked to tobacco carcinogens while EGFR mutations are found particularly in individuals who never smoked (although the relationship is unclear).[11]

Many cancer tumours produce by-products, which are released into the bloodstream.[12] Biomarkers are a measurable substance that may occur with the body due to an infection or a disease, such as cancer tumours and its byproduct.[13] A commonly measured biomarker found among ovarian cancer patients is cancer antigen 125, CA 125. CA 125 is a glycoprotein found in the epithelial surface of the reproductive tract (i.e. ovaries), digestive tract, respiratory tract and the eyes.[12] Most research conducted in relation to CA 125 has been centered on its use in ovarian cancer risk prediction.[14-17] However, there have been some studies that have shown CA 125 to be beneficial in assessing the risk of ovarian cancer when it is used in conjunction with risk prediction algorithms and other diagnostic tests (i.e. transvaginal ultrasound).[14,16] Although CA 125 has been found in the respiratory tract, few studies have closely examined the relationship between CA 125 and lung cancer. The main purpose of this study was to determine whether CA 125 is associated with lung

cancer. If an association exists, the next step is to identify whether smoking status is an effect modifier for this association, and to examine the association of CA 125 and lung cancer stratified by lung cancer stage. This study is expected to provide more understanding about CA 125 and its role as a potential predictor of lung cancer risk.

1.3 Literature Gaps

Previous studies suggest a possible relationship between CA 125 and lung cancer. However, currently there are no studies exploring the association between CA 125 and lung cancer in particular. The focus of prior literature was mainly on: the relationship between CA 125 and lung cancer in regards to survival prediction[18-20]; the evaluation of CA 125 against other cancer biomarkers for lung cancer[21-24]; or investigating the mechanistic role of CA 125 in relation to lung cancer.[25,26] Most studies had a low sample of ~100-600 participants.[18,19,24,26] These studies also had samples that were overrepresented by males, and underrepresented by females.[18,19,24-26] The varying topics on CA 125 and lung cancer warrant the need for further exploration of the relationship between CA 125 and lung cancer.[18,19,24-26]

1.4 Response to Gaps in Current Knowledge

The current study addresses the gaps in the literature by analyzing secondary data from the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial.[27,28] The PLCO cancer screening trial was a large American population-based randomized controlled trial that was sponsored by the U.S. National Cancer Institute (NCI) starting in 1993. This multi-center trial focused on determining the effects of various screening modalities on cancer-specific mortalities as well as secondary endpoints in males and females 55-74 years of age.[27,28] Further description of the trial can be found in Chapter 3. The PLCO cancer screening trial data had concurrently looked at the association between CA 125 and

ovarian cancer. To date, no studies have explored the relationship between CA 125 and lung cancer using the PLCO cancer screening trial data. Elevated levels of CA 125 might provide insights into carcinogenic mechanisms associated with lung cancer. This study provides the opportunity to describe this relationship, which in turn will further expand direction for future research.

1.5 Study Aims, Objectives, Hypotheses

Study Aim: To evaluate the association between cancer antigen 125 (CA 125) and lung cancer.

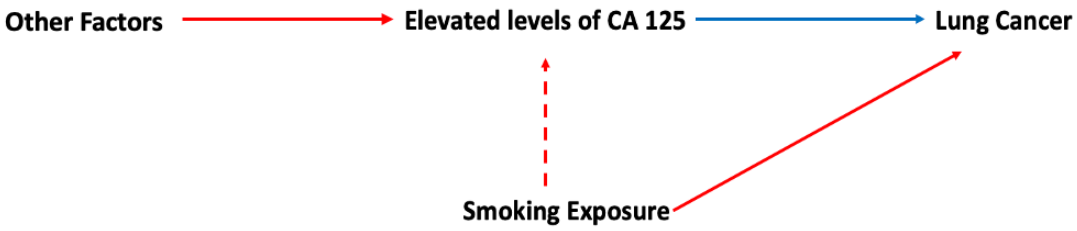

Figure 1. Possible Causal Pathways

Study Questions:

(1) Is there an association between CA 125 and lung cancer?

   a. Is there an association between CA 125 (dichotomous or continuous) and lung cancer?

(2) Does the association between CA 125 and lung cancer differ by potential effect modifier, smoking status?

   a. Does the association between CA 125 (dichotomous or continuous) and lung cancer differ by smoking status?

(3) Does the association between CA 125 and lung cancer differ by lung cancer stage?

a. Does the association between CA 125 (dichotomous or continuous) and lung cancer differ by lung cancer stage?

Study Objectives:

(1) Evaluate the association between CA 125 and lung cancer.

(2) Evaluate if the associations differ by potential effect modifier (smoking status).

(3) Evaluate if the association between CA 125 and lung cancer differs by lung cancer stage, early vs. advanced.

Hypotheses:

H0: CA 125 is not associated with lung cancer as estimated by odds ratio approaching the null value.

H1: CA 125 is associated with lung cancer as estimated by odds ratio away from the null value.

H0: The expected positive association between CA 125 and lung cancer will not be observed in individuals who smoked and not in individuals who never smoked.

H1: The expected positive association between CA 125 and lung cancer will be observed in individuals who smoked and in individuals who never smoked.

H0: The expected positive association between CA 125 and lung cancer will not be observed in early stage and not in advanced stage lung cancer.

H1: The expected positive association between CA 125 and lung cancer will be observed in early stage and in advanced stage lung cancer.

The independent variable of this study is CA 125 (dichotomous and continuous), and the dependent variable is lung cancer (yes/no). Logistic regression analyses will be used to examine the associations. Analyses will adjust for other appropriate covariates (risk factors for lung cancer).

1.6 Conclusion

Lung cancer is a disease that is often diagnosed at a later stage, when treatment is not curative, and prognosis is poor. In Canada, the most common cause of cancer death is lung cancer.[5] It was estimated that the pooled five-year survival rate for lung cancer patients, for all stages, was only 19%.[5] In 2020, lung cancer is expected to be the most common cause of cancer death among Canadians, accounting for 25% and 26% of all cancer deaths in males and females, respectively.[4] It is crucial to continue to produce research that could help further the understanding of this disease, and the relationships it has with various factors. Therefore, this study aims to contribute to the existing body of knowledge by evaluating the associations between CA 125, smoking status and lung cancer. The associations found in this study may expand understanding of the relationships between the CA 125 marker and lung cancer and determine if an association differs by smoking status and lung cancer stage. By developing an understanding of CA 125 and lung cancer, we may improve knowledge of lung carcinogenesis and provide insight into how early detection of lung cancer may be expanded integrating the use of biomarkers, such as CA 125.

CHAPTER II - LITERATURE REVIEW

2.1 Introduction

The current chapter will provide a basis and rationale for the present study. This chapter includes an overview of lung cancer biology, etiology, pathogenesis, and histology of lung cancer along with modifiable and non-modifiable risk factors for lung cancer. It will conclude with a review of the current literature and evidence pertaining to the relationship between CA 125 and lung cancer.

2.2 Lung Cancer Statistics

2.2.1 Worldwide Statistics

Lung cancer is the most common cancer and cause of cancer death worldwide.[29] In 2018, the International Agency for Research on Cancer (IARC) stated that worldwide, the crude age standardized incidence rate (ASIR) and mortality rate for lung cancer were 22.5 and 18.6 per 100,000, respectively.[30] Globally, lung cancer caused 2.09 million new cases and 1.76 million deaths.[29] The highest lung cancer incidences have been reported in South Korea, China, Turkey, Singapore and the Philippines.[31]Lung cancer mortality was reported to be highest in Eastern Europe, Western Asia, Northern Africa and Eastern and South Eastern Asia. [31]Worldwide, males (31.5 per 100,000) have a higher age standardized incidence rate (ASIR) compared to females (14.6 per 100,000) mostly due to varying tobacco usage over the past decades.[30] Lung cancer is a highly fatal disease as it has an overall mortality to incidence ratio of 0.87.[30]

2.2.2 Canadian Statistics

Lung cancer is the most common cancer in Canada.[3,5] The projected new cancer cases for lung cancer in 2020 among Canadians was 29,800 with about 21,200 deaths.[3] In Canada, the ASIR is higher

among males (64.8 per 100,000) compared to females (59.3 per 100,000).[29] The sex difference in rates could be reflective of smoking patterns, as males reached a peak in smoking use trend in the 1960s, while females smoking use peaked in the 1980s.[5] The lung cancer mortality rate in males began to level off in the late 1980s and has continued to decline.[5] The mortality rate for females increased until approximately 2006 and is now declining.[5] Despite this contrast, men continue to have a higher lung cancer mortality rate compared to females.[5] It was estimated that among Canadians, 98% of lung cancer cases occur in adults aged 50+.[5] Lung cancer mortality remains highest among those aged 70-80 years.[5] The most recent estimated 5-year net survival rate for lung cancer among Canadians was 19% overall, with 15% for males and 22% for females.[5]

2.2.3 US Statistics

Lung cancer is the second most common cancer and the leading cause of cancer death in both males and females in the US.[32] It was estimated that approximately 228,000 people will be newly diagnosed with lung cancer and about 135,000 lung cancer deaths will occur in the US in 2020.[32] It is estimated that a male's chance of being diagnosed with lung cancer in his lifetime is approximately 1 in 15, while for females it is 1 in 17.[32] Within the past few decades, the incidence rate of lung cancer among American males has decreased almost twice the decline of incidence of lung cancer in females. This may be explained by the differences in smoking uptake and cessation in the US, as seen in Canadian trends.[3] Similar to Canadian trends, Americans are most commonly diagnosed with lung cancer around the age of 65 years or older with an average age of lung cancer diagnosis at 70 years old.[32]

2.3 Lung Cancer Biology

2.3.1 Etiology & Pathogenesis

Lung cancer can be described as the proliferation of neoplastic cells within the respiratory epithelium.[33] Molecularly, cancer forms as a result of cell mutations that inhibit the function and control of healthy cells.[34] These mutations cause uncontrolled rapid cell division leading to the potential spread to neighboring tissue.[33] Three different types of genes can affect the formation of neoplastic, cancerous cells. These genes are proto-oncogenes, tumour-suppressor genes, and DNA repair genes.[35] Proto-oncogenes usually are involved with the control of cell growth and the division of cells. If altered, these genes become hyperactive and become cancer-causing oncogenes.[35] Similarly, tumour-suppressor genes are involved with cell proliferation, helping with the suppression of cellular replication.[35] When mutated, tumour-suppressor genes lose their inhibiting controls, which lead to uncontrolled cell division.[35] Lastly, DNA repair cells support the process of repairing DNA cells. When DNA repair cells become mutated, they lose their function, causing an accumulation of unrepaired and mutated cells, leading them to develop activation mutations in proto-oncogenes and deactivating mutations in tumour-suppressor genes, which in turn lead to cancer progression.[35] Due to the molecular complexity of genetic mutations, there is no one underlying cause or factor that leads to all cancer formation. Some of the most common mutated genes in lung cancer include EGFR, KRAS, TP53, MET proto-oncogene (MET), liver kinase B1 (LKB1), v-raf murine sarcoma viral oncogene homolog B (BRAF), catalytic subunit alpha gene (PIK3CA), anaplastic lymphoma kinase (ALK), receptor tyrosine kinase (ROS1), and rearranged during transfection (RET).[33,36]

2.3.2 Histology

Lung cancer is categorized into two main histological types, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). Through light microscopy and histochemical analysis, lung cancer is differentiated into four main classes: SCLC, and subtypes of NSCLC, which are adenocarcinomas, squamous cell carcinomas (ScC), and large-cell carcinomas.[37] The following section will go into detail on the four main histological classes of lung cancer.

2.3.2.1 Small-Cell Lung Cancers

Small-cell lung cancer represents approximately 13% of all new lung cancer cases or ~180,000 cases worldwide per year.[38] Over the past 30 years, incidence rates of SCLC have decreased due to the changes in smoking trends, predominately in industrialized countries such as Canada and the US.[38] The opposite has been found in Eastern Europe and Asia, where the incidence of SCLC has increased due to continued high prevalence of smoking.[38] Van Meerbeeck and colleagues describe SCLC as "a malignant epithelial tumour consisting of small cells with scant cytoplasm, ill-defined cell borders, finely granular nuclear chromatin, and absent or inconspicuous nucleoli".[35] In 90% of cases, only small cells are involved, while the remainder of cases contain large cell components.[39] SCLC stages most commonly are described as either limited or extensive.[32,38] Limited stage disease involves a tumour that is localized within the lung but has not yet spread as far as diagnostic evaluation can determine.[38] The extensive-stage disease has spread to the contra-lateral lung, distant lymph nodes and/or other organs in the body.[32,37] The main characteristics of SCLC are: they are aggressive tumours that metastasize early, have a high likelihood of spreading and multiplying, and have a high initial response towards chemotherapy.[38]

2.3.2.2 Non-Small Cell Lung Cancers

2.3.2.2.1 Adenocarcinomas

Adenocarcinomas account for approximately 40% of all lung NSCLC .[40] Adenocarcinomas are the most common type of NSCLC found in individuals who never smoked.[40] These tumours are predominantly diagnosed in females compared to males.[32] Adenocarcinomas form from glandular cells, which would have been secretory cells if they did not undergo mutation.[32] They are often peripheral masses found within the lungs that tend to grow slower than other lung tumours but metastasize faster than squamous-cell carcinomas.[32,37]

2.3.2.2.2 Squamous-Cell Carcinomas

Squamous-Cell Carcinomas (ScC) account for approximately 25-30% of all NSCLC cases.[32] Squamous cell tumours are small flat cells that are formed from reserve cells and replace those cells that have been damaged within the lung airways.[4] These cells develop and are usually found in the major lung airways and branches or within the bronchus.[32] ScC is a slow-growing cancer where the cells metastasize late, and symptoms develop slowly.[40] Almost all cases of ScC are caused by tobacco exposure.[41]

2.3.2.2.3 Large-Cell Carcinomas

Large-cell undifferentiated carcinomas represent approximately 10-15% of all NSCLC.[4] Among all subtypes, large-cell carcinomas represent the smallest proportion of cases compared to adenocarcinomas and ScC.[32] Large-cell undifferentiated carcinomas are aggressive and can spread quickly.[34] Although large-cell carcinomas can be found anywhere in the lung, it is predominately a peripherally located mass.[32] Due to the tumour's location, the process of diagnosis is easier than

central masses as it is not masked by the heart, esophagus and mediastinum, which are central and can obscure nodules.[32]

2.4 Tumor Staging

2.4.1 Tumour-Node-Metastasis Cancer Staging System

Accurate tumor classification and staging is an essential step in determining treatment options and prognosis. The Tumour-Node-Metastasis (TNM) tumour staging system, created by the American Joint Committee on Cancer (AJCC) and the Union for International Cancer Control (UICC), is typically used in identifying and characterizing non-small cell lung carcinoma (NSCLC).[42] This system describes NSCLC using three classifications, T- the size of the primary tumour in long axis, or the direct extent of the tumour into neighboring structures; N- the degree of spread to regional lymph nodes; M – the presence of metastases beyond regional lymph nodes.[43] Further information on the TNM system is found in Table 1.

Table 1. Tumour-Node-Metastasis (TNM) cancer staging system, 7th edition (American Joint Commission on Cancer, 2010)

Primary Tumour (T)	
TX	Primary tumour cannot be assessed, or tumour proven by the presence of malignant cells in sputum or bronchial washings but not visualized by imaging or bronchoscopy
T0	No evidence of primary tumour
Tis	Carcinoma *in situ*
T1	Tumour 3 cm or less in greatest dimension, surrounded by lung or visceral pleura, without bronchoscopic evidence of invasion more proximal than the lobar bronchus (i.e., not in the main bronchus)
T1a	Tumour 2 cm or less in greatest dimension
T1b	Tumour more than 2 cm but 3 cm or less in greatest dimension
T2	Tumour more than 3 cm but 3cm or less or tumour with any of the following features (T2 tumours with these features are classified T2a if 5cm or less); Involves main bronchus, 2cm or more distal to the carina; Invades visceral pleura (PL1 or PL2); Associated with atelectasis or obstructive pneumonitis that extends to the hilar region but does not involve the entire lung
T2a	Tumour more than 3 cm but 5 cm or less in greatest dimension
T2b	Tumour more than 5 cm but 7cm or less in greatest dimension
T3	Tumour more than 7cm or one that directly invades any of the following: parietal pleural (PL3) chest wall (including superior sulcus tumours), diaphragm, phrenic nerve, mediastinal pleura, parietal pericardium; or tumour in the main bronchus (less than 2 cm distal to the carina; or associated atelectasis or obstructive pneumonitis of the entire lung or separate tumour nodule(s) in the same lobe
T4	Tumour of any size that invades any of the following: mediastinum, heart, great vessels, trachea, recurrent laryngeal nerve, esophagus, vertebral body, carina, separate tumour nodule(s) in a different ipsilateral lobe
Regional Lymph Nodes (N)	
NX	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastases
N1	Metastasis in ipsilateral peribronchial and/or ipsilateral hilar lymph nodes and intrapulmonary nodes, including involvement by direct extension
N2	Metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s)
N3	Metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph node(s)
Distant Metastasis (M)	
M0	No distant metastasis
M1	Distant metastasis
M1a	Separate tumour nodule(s) in a contralateral lobe tumour with pleural nodules or malignant pleural (or pericardial) effusion
M1b	Distant metastasis

2.4.2 Stage Grouping

Once an NSCLC is confirmed, the TNM staging system is used. Tumours that have similar characteristics (such as the same prognosis and treatment options) are grouped and classified into one of the following stages: 0, I, II, III or IV, with lower stage numbers indicating a better prognosis.[42] Stages are often subdivided into an A or B to distinguish specific characteristics.[42] Further information on stage grouping is found in Table 2.

Table 2. Stage grouping for lung cancer (American Joint Commission on Cancer, 2010)

Stage	Tumour status	Nodal status	Metastases
Occult carcinoma	TX	N0	M0
Stage 0	Tis	N0	M0
Stage IA	T1a	N0	M0
	T1b	N0	M0
Stage IB	T2a	N0	M0
Stage IIA	T2b	N0	M0
	T1a	N1	M0
	T1b	N1	M0
	T2a	N1	M0
Stage IIB	T2b	N1	M0
	T3	N0	M0
Stage IIIA	T1a	N2	M0
	T1b	N2	M0
	T2a	N2	M0
	T2b	N2	M0
	T3	N1	M0
	T3	N2	M0
	T4	N0	M0
	T4	N1	M0
Stage IIIB	T1a	N3	M0
	T1b	N3	M0
	T2a	N3	M0
	T2b	N3	M0
	T3	N3	M0
	T4	N2	M0
	T4	N3	M0
Stage IV	Any T	Any N	M1a
	Any T	Any N	M1b

2.4.3 Tumour Grade

Samples of the tumour taken during a biopsy can be further analyzed to determine the degree of abnormality, which can show how quickly a tumour might grow and/or spread.[42] This procedure is called tumour grading and is assessed using microscopy.[42] If most cells and tissues of the tumour resemble healthy cells and tissues, the tumour is classified as "well-differentiated".[42] Well-differentiated tumours spread slowly, which indicate a better prognosis.[42] Tumours that have a high number of abnormal-looking cells and tissues are classified as "undifferentiated" or "poorly differentiated".[42] The grades are classified as the following: GX, G1, G2, G3, and G4.[42] Lower-grade numbers correspond to a better prognosis. A complete description of the tumour grading system can be found below in Table 3.

Table 3. Tumour grading system. (American Joint Commission on Cancer, 2010)

Grade	Description
GX	Grade cannot be assessed
G1	Well-differentiated (low grade)
G2	Moderately differentiated
G3	Poorly differentiated
G4	Undifferentiated (high grade)

2.5 Clinical Features

2.5.1 Signs and Symptoms

The symptom onset for most lung cancers is slowly progressing, often appearing when the cancer has already reached a later stage. Approximately 70% of patients already have locally advanced or metastasized lung cancer at time of diagnosis.[21] Only about 10% of cases of lung cancer are diagnosed in asymptomatic patients.[37] Those diagnosed with lung cancer often experience non-

specific systemic symptoms such as weight loss, fatigue, and anorexia.[34] Approximately 40% of patients present with systemic symptoms directly related to the first spread of the tumour, such as chest wall invasion, esophageal symptoms, phrenic nerve paralysis, pleural effusion, recurrent laryngeal nerve paralysis or superior vena cava obstruction.[37] One-third of lung cancer patients present with extra thoracic signs such as bone pain, confusion, headache, personality change, nausea, vomiting, seizures, or weakness.[37] A total of 10% of patients also experience paraneoplastic syndromes due to the biochemical secretion, such as hypercalcemia and antidiuretic hormone, which are produced and released by the tumour itself.[37] Due to the broad nature of these symptoms, some cases of lung cancer may be initially misdiagnosed as pneumonia or collapsed lung.[40]

2.6 Diagnosis

Despite clinical indications and symptoms, lung cancer cases need a histopathological diagnosis to be conclusive.[37] Patients who are suspected of having early-stage NSCLC are recommended to have tissue samples taken using surgical resection, thoracoscopy or needle biopsy.[37] For those patients with suspected SCLC or metastasized NSCLC, thoracentesis of a pleural effusion, excisional biopsy of an accessible node, bronchoscopy, or transthoracic needle aspiration can be used for pathology tissue sampling.[37] The accuracy of these various tests can vary depending on the location and size of the tumour.[37]

2.7 Treatment and Prognosis

Lung cancer treatment differs depending on the histologic type of tumour and cancer stage. SCLC has two kinds of stages, limited-stage SCLC and extensive-stage SCLC.[35] Approximately 30% of SCLC are limited stage.[35] The primary treatment option for this type of lung cancer is chemotherapy, although response rates vary depending on stage at diagnosis.[37] The five-year survival rate is 15-25%

and less than 5% for limited-stage and extensive-stage lung cancer, respectively.[37,38] For NSCLC, treatment options vary depending on stage. For patients with stage I-IIIA lung cancer, surgery can be performed in combination with pre- or post-operative chemotherapy, with the possibility of radiotherapy.[37] For stages IIB and IIIA lung cancer, where surgery is not possible, chemotherapy with or without radiotherapy is the first option.[37] For advanced stages of IIIB and IV lung cancer, the primary treatment is chemotherapy with or without radiotherapy. With stages IIIB and IV disease, lung cancer has often spread to the brain.[34] In this occurrence, some surgical resection may be possible to remove the brain tumour.[37] For NSCLC, the 5-year survival rates are as follows: 60-70% in stage I, 40-50% in stage II, 15-30% in stage III, 10-20% in stage IIIA/IIIB.[37] For stage IIIB/IV, there is only a 2-year survival of 10-15%.[37]

2.8 Lung Cancer Risk Factors

To gain a better understanding of the present research study, it is essential to understand the relationship between various risk factors and lung cancer. This section will describe two main types of risk factors that affect the incidence of lung cancer: modifiable and non-modifiable risk factors. Modifiable risk factors are lifestyle risk factors that can be changed. Non-modifiable risk factors are factors that cannot be changed.

2.8.1 Modifiable Risk Factors

2.8.1.1 Smoking

Since its peak in the mid-20th century, cigarette smoking (whether first-hand or second-hand) continues to be one of the main direct causes of lung cancer.[44] Approximately 80-90% of lung cancer cases worldwide are caused by smoking with a 10 to 20- fold increased odds of lung cancer in those who smoke compared to individuals who never smoked.[33,45,46] Although cigars and pipe smoking have

been found to have a causal relationship with lung cancer as well, the relationship is less prominent due to the difference in frequency and inhalation of this type of smoke.[45] In 2017, cigarette smoking prevalence (those presently smoking every day or occasionally) in Canada was 15% or 4.6 million individuals.[6] This was an increase from 13% (3.9 million individuals who smoked) in 2015.[6] Males continue to have a higher prevalence of smoking with 17% (2.5 million) of the Canadian population being individuals who smoked compared to 13% (2.1 million) of females of the Canadian population being individuals who smoked.[6] In the US, 34.2 million adults were reported to currently smoke cigarettes in 2018.[47] In recent data, the current smoking rate in the US has declined from 20.9% in 2005 to 13.7% in 2018.[47] Similarly, to Canada, the US has more males who currently smoke compared to females who currently smoke.[47] Approximately 15.6% of the US male population currently smoke while 12% of the female population currently smoke.[2] Several studies have also found that females were found to be less successful in smoking cessation attempts compared to male counterparts.[48,49] It is estimated that approximately 20% of worldwide cancer deaths would be eliminated if smoking cessation attempts were successful.[33]

2.8.1.1.1 Individuals who never smoked

Lung cancer is often found to be associated with individuals who smoked, yet there is a large amount of literature identifying a relationship between individuals who never smoked and adenocarcinomas, predominately among females.[9,50,51] An individual who never smoked is defined as smoking less than 100 cigarettes in their lifetime.[29] It has been estimated that 10-25% of lung cancers worldwide occur in individuals who never smoked.[9] Females, who never smoked or used tobacco, are approximately 2.5 times more likely to be diagnosed with lung cancer compared to males who never smoked or used tobacco.[50,52] It has also been predicted that lung cancer in non-smoking females may

have a hormonal element that may possibly have an interaction with factors such as indoor air pollution, unhealthy cooking practices, indoor heating fumes, exposure to ionizing radiation, hereditary risks and a history of respiratory infections.[9,52,53] Individuals who never smoked are often not considered at risk for lung cancer causing diagnoses to be delayed and more likely to be found at later stages of the disease.[52]

2.8.1.1.2 Cigarette Smoke

Cigarette smoke contains thousands of chemicals with 62 carcinogens identified by the IARC.[30] One of the most potent carcinogens included in cigarettes is polycyclic aromatic hydrocarbons (PAHs) such as benzo[a]pyrene (BAP) and nicotine-derive nitrosoaminoketone (NNK).[11,33] Additionally, there are organic and inorganic compounds included in the makeup of cigarette smoke such as benzene, vinyl chloride, arsenic, chromium, and radioactive matter such as radon, and polonium.[33] These carcinogens can become carcinogenic metabolites that disrupt normal cell production, causing severe mutations resulting in genomic instability, genetic changes, and dysfunctional DNA repair.[11,33]

2.8.1.1.3 Smoking Measures

Smoking measures such as smoking duration and smoking intensity are found to have varying impacts on the relationship between smoking and lung cancer. To account for these differences, Doll and Peto created a model that quantified how changes in duration and intensity affect lung cancer risk.[45,54] Through this model, a linear relationship was found between cigarettes smoked per day (smoking intensity) and lung cancer risk; that is, double the number of cigarettes smoked per day, double the risk of lung cancer.[55] Of note, one's susceptibility to lung cancer is not only affected by smoking duration and intensity but also by one's genetic predisposition and exposure to environmental factors.[33]

2.8.1.1.4 Second-Hand Smoking

Second-hand smoking (SHS) has been continuously linked to an increased risk of lung cancer. It has been found that individuals who never smoked have a 25-29% increased risk of lung cancer when married to an individual who smoked.[45] A similar trend has been found among the workplace, where secondhand smoke exposure causes approximately a 17% increased lung cancer risk.[45] Further evidence found that SHS had a dose-response relationship between smoking and lung cancer. Even at low dosages of second-hand smoke, an increased risk of lung cancer was consistently found, demonstrating no threshold effect.[21]

2.8.1.4 Physical Activity

Physically active individuals have been shown to have a lower risk of lung cancer compared to those who are physically inactive.[21,45] Many studies have shown an inverse association between physical activity and lung cancer risk.[21,45] Despite not knowing the exact mechanism that causes this decrease in lung cancer risk through physical activity, a 13-30% reduction in lung cancer risk has been found among those who were moderate or highly engaged in physical activity.[21,45]

2.8.2 Non-Modifiable Risk Factors

2.8.2.1 Sex

Since the 1980s, lung cancer has become the leading cause of cancer mortality, surpassing breast cancer death among females.[56] Twice as many females die from lung cancer compared to breast cancer.[56] As mentioned previously in this chapter, there is a gender gap regarding lung cancer and mortality due to the 20-year gap between peak prevalence of smoking among males (who peaked in the 1960s) and females (who peaked in the 1980s).[33] There is also controversy linked to whether females are more susceptible to the carcinogens within cigarette smoke compared to males.[33] A study

by Ryberg and colleagues noted that despite lower intensity cigarette smoke exposure, female participants were still more susceptible to carcinogens and in turn, more likely to develop lung cancer compared to males.[57] Moreover, female hormones have been seen to have a mechanistic effect on lung cancer. Literature has identified that females using hormonal therapies such as estrogen and progestin had a higher likelihood of developing lung cancer.[58] For females, using hormone replacement therapy for 10 years or longer, resulted in a 50% increased risk of lung cancer.[58] Notably, females experiencing early menopause (40 years or younger) showed an inverse effect and had a decreased risk of lung cancer.[58]

2.8.2.2 Age

Due to the era of the baby boomers, the average age of most populations in developed countries is increasing, causing large numbers of older adults to be at risk for cancer, as generally cancer incidence increases with age.[33] Older age has been associated with cancer development due to biological factors such as DNA damage over time and the shortening of telomeres.[3] It has been reported that about 10% of lung cancer cases occur in patients less than 55 years of age; whereas, 53% of lung cancer cases occur in individuals 55 to 74 years of age whilst 37% occur in those over the age of 75 years.[3] Along with biological factors, many older adults have underlying co-morbidities, which are thought to play an additional role in older adults' susceptibility to cancer compared to younger populations.[3]

2.8.2.3 Race/Ethnicity

Racial and ethnic disparities continue to be seen within lung cancer incidence rates, especially among the US population.[33] In 2015, Siegel and colleagues reviewed cancer statistics within the US, and compared and contrasted variations between race/ethnicity and lung cancer among the

American population.[56] It was found that Non-Hispanic Black males had the highest lung cancer incidence of 87.9 per 100,000 while Asian/Pacific Islander and Hispanic males had the lowest incidence with a rate of 45.2 and 40.6 per 100,000, respectively.[59] Non-Hispanic White males and American Indian/Alaskan Native males were found to have moderate incidence rates with 75.9 and 71.9 per 100,000, respectively. In contrast, Non-Hispanic White, American Indian/Alaskan Native and Non-Hispanic Black females were reported to have the highest lung cancer incidence rates of 57.6, 55.9 and 50.1 per 100,000, respectively.[59] These rates are almost twice that of Asian/Pacific Islander and Hispanic females where the lung cancer incidence rates are 27.9 and 25.2 per 100,000 respectively. It is unclear why these discrepancies exist, therefore, there is a large need to further explore lung cancer and race/ethnic disparities through research.

2.8.2.4 Socioeconomic Status

Research has shown that lung cancer incidence is highly impacted by factors surrounding socioeconomic status. Individuals in a higher income bracket have been found to have greater access to resources such as health information and healthcare, in turn leading to less disparities in mortality and improved survival.[59] Individuals in a lower income bracket have a greater prevalence of cigarette smoking uptake and a lower likelihood to participate in smoking cessation practices compared to those in a higher income bracket.[59] It has also been found that approximately 27.9% of individuals below the poverty threshold in the US smoke cigarettes.[60] In Siegel et al.'s article looking at US cancer statistics, the authors found that 32.1% of individuals with less than a high school education smoked cigarettes while only 9.1% of college graduates smoked cigarettes.[56] Torre and colleagues also examined US cancer statistics and found that individuals with less than a high school education had a lung cancer incidence rate of 166.6 per 100,000 while college graduates had a lung cancer incidence

rate of 57.6 per 100,000.[60] In de Groot and colleagues article reviewing the epidemiology of lung cancer, the authors found that some studies suggested an association between low socioeconomic status and lung cancer regardless of cigarette smoking status and exposure.[3] It was suggested that regardless of cigarette smoking, individuals with low socioeconomic status were more likely to be exposed to environmental risk factors such as dangerous housing (due to mold or poor plumbing), housing near industrial properties (i.e. pollution) and occupational exposures that in turn impact their risk of lung cancer.[61]

2.8.2.5 Familial Aggregation

Through various family linkage analysis and genomic studies, it has been consistently shown that individuals with relatives who have lung cancer have a higher risk of the disease themselves.[62] Although genetic heritability has often been seen as a cause of familial aggregation, it also is represented as the accumulation of shared exposures and habits (such as smoking) among a family.[63] In a study by Lissowska and colleagues[71], family history and lung cancer risk were explored. The study found that those who were associated with having a first-degree relative with lung cancer had a significant risk (OR=1.72, 95% CI 1.56-1.88) compared to those with no family history.[64] Similarly, a study conducted by Brenner and colleagues found that the ratio of observed incident cases of lung cancer among first-degree relatives had a 1.9 risk (95% CI 1.6-2.4) compared to the expected frequency of lung cancer.[65] Overall, if a family history of lung cancer is present, an individual's risk of lung cancer increases, especially with multiple affected relatives.[63]

2.8.2.6 Previous History of Lung Disease

Numerous lung diseases are potential triggers in the formation of lung cancer. Acquired lung diseases such as chronic obstructive pulmonary disease (COPD), adult pneumonia, and tuberculosis

have all been suggested to influence lung carcinogenesis due to their inflammatory processes on tissues.[65] In a meta-analysis conducted by Brenner and colleagues[72], previous lung diseases were explored as potential risk factors for the development of lung cancer. In the study, individuals with prior history of COPD (chronic bronchitis and emphysema) had RRs of 2.22 (95% CI: 1.66-2.97), 1.52 (95% CI: 1.25-1.84), and 2.04 (95% CI: 1.72-2.41), respectively (after adjusting for smoking).[65] For those individuals who had a previous history of pneumonia or a prior history of tuberculosis, the RR was 1.43 (95% CI 1.22-1.68) and 1.76 (95% CI 1.49-2.08), respectively.[65] Although each of these lung diseases have diverse pathologies, each lead to a common outcome, inflammation. Therefore, it is essential to consider each when exploring potential risk factors for lung cancer.

2.9 CA 125

CA 125 is "a high-molecular-weight glycoprotein recognized by a monoclonal antibody, which was raised using an ovarian cancer cell line as an immunogen".[66] It was first discovered in 1981 by Dr. Robert C. Bast, when testing for different proteins on ovaries in pursuit of creating an antibody for females with ovarian cancer.[66] Dr. Bast was successful in his pursuit during his 125th attempt where he discovered an antigen that could be used as a tumour marker for ovarian cancer, which was named CA 125.[66] Unfortunately, regardless of this discovery, the use of CA 125 as a biomarker came with limitations.[66] In 1989, Dr. Bast and colleagues performed a meta-analysis examining studies that used CA 125 as a biomarker.[66] It was found that for a patient with stage I ovarian cancer, only ~50% of patients would be identified.[16] For patients with stage II-IV, the identification proportion increased to around 85-94%, causing CA 125 to be a useful marker primarily in detecting late stages of ovarian cancer, which diminished the purpose of the biomarker.[16] Similarly, while examining ovarian cancer by histological types, there were discrepancies in the identification of elevated CA 125 levels among

the various types.[16] For patients with serous ovarian cancer, approximately 80% of patients would have elevated levels of CA 125, compared to patients with mucinous ovarian cancer, which only 70% of patients would have elevated levels of CA 125.[16] Furthermore, elevated levels of CA 125 have been found among patients with benign gynecological and non-gynecological conditions such as pregnancy, endometriosis, COPD, congestive heart failure, colitis, appendicitis and more.[67-69] Therefore, it is difficult to distinguish whether CA 125 accurately measures the risk of ovarian cancer, when its levels can be affected by various causes. With these limitations, it has been suggested that for ovarian cancer screening, CA 125 should be incorporated as an additional screening marker along with more robust and sophisticated evidence-based cancer screening practices.

2.9.1 CA 125 and Ovarian Cancer Screening

Following the discovery of CA 125, two screening trials investigating CA 125 for early detection of ovarian cancer were initiated in the 1990s.[16] The first trial was conducted in the United Kingdom (UK) and assessed the performance of using CA 125 for ovarian cancer screening. This trial followed 22,000 post-menopausal females ages >45 who were followed up every 3 months. The second trial occurred in Sweden, which had two annual screens, and followed 5,550 females ages >40.[70] The trials had a combination of the CA 125 blood test, followed by an imaging test for females with a positive blood test.[70,71] This screening procedure provided a combined false positive rate of 0.1-0.2% with a specificity of 99.8%.[70,71] The number of false positive surgeries per true positive surgeries were found to be less than five in the UK trial and two in the Swedish trial.[70,71]The two trials provided promising proportions of specificity and positive predictive value (PPV), however, a concern remained over the presumed low sensitivity for early-stage ovarian cancer.

One of the first major studies that looked at the relationship between CA 125 screening and ovarian cancer was the PLCO cancer screening trial. The PLCO cancer screening trial looked at post-menopausal females at average risk with no personal history of ovarian cancer and examined whether screening would detect their risk of ovarian cancer.[72] A total of 78,237 females took part in the trial, and of these females, ~39,000 received no screening.[72] In the intervention arm, 28,803 received initial CA 125 screening, 28,519 received initial transvaginal ultrasound, 28,816 received at least one test and 28,506 received both tests.[72] Results demonstrated that 1.4% of the entire female population had a positive CA 125 test and 4.7% had an abnormal ultrasound result.[72] Of the females with abnormal tests, 541 received surgical intervention to detect 29 cases of cancer. Therefore, to identify 1 cancer, 19 females had to undergo surgery who otherwise would not have had surgery and were at risk for unnecessary complications.[72] The study concluded that the effect of screening on ovarian cancer mortality in the PLCO cancer screening trial cohort required longer follow-up and that changes were needed if this biomarker was to be used in the future for this purpose.[72]

In order to maximize the value of this insight on the differential longitudinal behaviour of CA 125, formal statistical algorithms have been applied in conjunction with CA 125 and ultrasound testing. Lu and colleagues[82]created a more sophisticated screening trial, implementing an ovarian cancer risk algorithm (ROCA) in conjunction with the CA 125 test and transvaginal ultrasound. The study placed females into various risk categories (normal risk, intermediate risk and high risk) based on their CA 125 result and calculated an individual's risk overtime using the ROCA algorithm.[14] If the ROCA screen came back normal, the individual would be screened in one year, if an individual had a result of intermediate risk, they would be screened in 3 months, and if an individual had a result of high risk, the individual would be sent for an immediate ultrasound and examined by a gynecological

oncologist.[14] It was found that only 6% of individuals received an intermediate risk, and less than 1% received a high-risk result and were triaged to immediate ultrasound, causing screening to be beneficial and not overly burdensome.[14] Of the 10 individuals who received a high-risk result and received surgery, 4 had invasive ovarian cancer, 2 had low malignant/borderline ovarian cancer and 1 had endometrial cancer.[14] Therefore, only 3 out of the 10 individuals had benign disease, showing a more successful outcome from screening compared to the PLCO cancer screening trial.[14] Of the 10 individuals with a high-risk result, no individual had physical symptoms based on screening and many had undergone screening for several years. It was found that only after the 3rd, 6th or 8th test, a slight change in CA 125 levels was found, many of the test results could have been deemed as "normal" levels of CA 125, when the algorithm was triggered and signified that these individuals should be closely screened. It was noted that it is not the exact level of CA 125 that was important but a change in the level of CA 125 over time.[14]

A more recent ambitious trial in the United Kingdom called the UKCTOCS (UK Collaborative Trial of Ovarian Cancer Screening) used the ROCA screening algorithm and tried to determine whether this method could increase survival rates of ovarian cancer.[15] Approximately 200,000 postmenopausal females ages>50 were randomized into a no screening arm (n=100,000) and a screening arm (n=100,000).[15] Among the screening arm, ~50,000 received annual ultrasound screening and ~50,000 received a multimodal screening.[15] This multimodal screening used the annual CA 125 test results and interpreted the pattern over time while using the ROCA algorithm, similarly to Lu and colleagues[83] ovarian cancer screening trial. Although there was an improvement in mortality, statistically there was no difference between the screening group and non-screening group.[15]

The benefits and harms of the UKCTOCS and PLCO trials have been thoroughly reviewed by the US Preventive Services Task Force (USPSTF) and the Canadian Task Force on Preventive Health Care (CTFPHC), which both have recommended against ovarian cancer screening in asymptomatic females, due to the lack of efficacy and statistically significance found within these trials.[73,74] Therefore, through the review of these large independent trials, it can be concluded that the use of CA 125 as a sole biomarker for ovarian cancer screening is not effective.

2.9.1 CA 125 and Lung Cancer

The relationship between ovarian cancer and elevated levels of CA 125 have been widely researched, but it has not been comprehensively evaluated in association with lung cancer. Previous studies have explored the relationship between CA 125 and lung cancer but to a varying extent. Some explored the relationship between CA 125 and lung cancer in regards to survival prediction[18-20] while others evaluated CA 125 against other cancer biomarkers such as carcinoembryonic antigen (CEA) and carbohydrate antigen 19-9 (CA 19-9) [21-24] or investigated the mechanistic role of CA 125 in lung cancer. For the purpose of the current study, this section will focus mainly on the general association and mechanistic role of CA 125 with lung cancer.

2.9.1.1 MUC16 and Lung Cancer

In 2001, molecular cloning of CA 125 lead to the discovery of MUC16.[75] The MUC16 gene product belongs to a group of mucins that protect and lubricate epithelial surfaces that line the internal organs within the body.[26] Although their function is to protect epithelial surfaces, this mucin has been found to be involved in the development of cancer.[76] MUC16 often splits and sheds into the bloodstream making it easy to be found and measured through a blood test.[77] MUC16 was initially believed to be specifically an ovarian cancer biomarker, but over time and through various research, it

has become evident that this marker could also be detected in patients with cancers such as gastric, colorectal, lung and pancreatic.[12,78] This agreed with Kim and colleagues' study[90], which reported MUC16 to be highly mutated among cancers including lung cancer.

Ma and colleagues[91] explored the prognostic values of CA 125 (MUC16) and other biomarkers in 164 patients (101 males, 63 females) with stage I NSCLC who underwent surgery. The authors found a 5.7% positive result for CA 125(MUC16) in 131 adenocarcinoma patients and a 3.1% positive result for CA 125 in 43 non-adenocarcinoma patients.[79] Although CA 125 (MUC16) levels were elevated, the authors concluded that more research needs to be conducted in order to confirm the use of CA 125 as a biomarker in lung cancer.[79]

In contrast to the previous study, Kanwal and colleagues[24]examined and obtained MUC16 mRNA levels in NSCLC tissues and their adjacent non-malignant tissues in 84 patients (51 males and 33 females) residing in air-polluted regions in China. When compared with matched adjacent noncancerous tissues, the MUC16 mRNA levels were significantly increased in 48.8% (41/84) of the NSCLC tissues.[26] However, it was noted that MUC16 mRNA expression did not correlate with gender ($p=0.74$), age ($p=0.27$) or histologic type ($p=0.53$).[26] Although this study had a small sample size, it provided promising results and demonstrated that MUC16 up-regulation induced by gene mutation may be involved in the development and progression of lung cancer.[26]

Although MUC16 has been shown to be involved in the growth and metastasis of several cancers, its role in lung carcinoma remains unclear and needs to be further studied in larger, more robust studies to test the reproducibility of the results in the studies mentioned previously in this section.

CHAPTER III - METHODS

3.1 Introduction

This chapter includes the study design and methods of the present study. Details on the source data, recruitment and data collection methods are described within. Moreover, this chapter concludes with the analytic and evaluative strategies used to address the study objectives described in Chapter 1.

3.2 Source Data – PLCO Trial

The present research was conducted using secondary randomized controlled trial (RCT) data from the PLCO trial sponsored by the U.S. National Cancer Institute (NCI).[28] The PLCO cancer screening trial was a multi-centre trial that assessed whether cancer screening examinations could reduce mortality due to prostate, lung, colorectal and ovarian cancers.[28]

3.2.1 Study Sample Data

Each of the ten screening centers were responsible for identifying and establishing its procedures for recruiting and identifying participants for the trial based on the guidelines developed by the NCI.[28,80] Between November 1993 and April 2001, screening centers contacted potential participants through direct mail, rosters of names and addresses from profit and not-for-profit (including government) organizations, community outreach, and use of mass media.[27,28] The target goal for enrollment was roughly 75,000 males and 75,000 females between the ages of 55-74 years.[27,28] Once the screening centers identified potential participants, information about the participants was collected to determine their eligibility for the trial.[27,28] Informed consent was

obtained from all participants at all sites. In total, 154,938 participants (78,234 females and 76,704 males) were enrolled in the study.[27,28]

3.2.2 Inclusion & Exclusion Criteria

An individual was eligible to participate in the PLCO cancer screening trial if they did not have any of the exclusion criteria. This exclusion criteria included:[81]

- Having an age of less than 55 or greater than 74 years at time of randomization
- A history of prostate, lung, colorectal or ovarian cancer, or current treatment for any cancer except basal or squamous cell skin cancer
- Individuals who were participating in another cancer screening or cancer primary prevention trial
- Females with previous surgical removal of ovaries before October 1996 (criteria was changed after this date as it was identified that this could affect design power for detecting mortality reduction in ovarian cancer)
- Individuals with prior surgical removal of the entire colon, or one lung
- Individuals who have participated in previous cancer screening trials
- Females who had taken the medication Tamoxifen or Evista/Raloxifene within the previous 6 months before randomization
- Individuals who had a colonoscopy, sigmoidoscopy, or barium enema within 3 years before randomization
- Individuals who did not or were not willing to sign the consent form.

3.2.3 Randomization and Screening Process

Random assignment was implemented using compiled software and encrypted files on microcomputers.[81] Block randomization was used to approximately distribute participants equally into the intervention arm, which included screening, or the control arm, which included standard medical care.[28] The block randomization scheme includes various lengths of random block permutations, which were stratified by screening center, age, and gender.[28] A total of 39,105 females and 38,340 males were randomized into the intervention arm. All repeat screening interventions were completed within six years. The intervention arm was then followed up for at least an additional seven years.[27,28] 38,111 females and 38,345 males were randomized into the control arm.[27,28] Participants in the control arm were followed up for 13 years after enrollment but were not given any screening examinations.[27,28] Females who were randomized into the screening arm received chest x-rays, flexible sigmoidoscopy, CA 125 blood tests, and transvaginal ultrasound.[81] For individuals with negative screen results, a follow up call was implemented to keep track of regular screening attendance.[28] A similar procedure was performed for those individuals with suspicious or positive results by screening, but for whom it did not reveal a prostate, lung, colorectal or ovarian cancer.[82] Baseline information on sociodemographic characteristics, risk factors for cancers being studied, and screening history were collected from all participants. It was also required for all participants to complete a dietary questionnaire and annual information on their health status.[80]

3.2.3.1 Lung Cancer Screening

For lung cancer screening, a posteroanterior chest X-ray was taken by a qualified technologist and later interpreted by a radiologist.[82] The X-ray exam was classified as positive (suspicious) for lung cancer if the evaluation exposed any of the following pulmonary abnormalities: nodule, mass, major

atelectasis/lobar collapse, hilar or mediastinal lymph node enlargement infiltrate/consolidation/alveolar opacity, or pleural mass.[82] The X-ray exam was classified as negative for lung cancer if the evaluation showed midline structure, the heart to be of normal size (not displaced or enlarged), and the pulmonary parenchyma revealed no abnormality suspicious for cancer.[82] In December 1998, a few changes were implemented for lung cancer screening. The remaining third annual chest X-ray exams were offered only to individuals who smoked and follow up was extended to 3 years for all participants to be followed up for at least 13 years from the time of randomization.[82]

3.2.3.2 CA 125 Assay

The original CA 125 assay used in the trial was Centocor CA-125 radioimmunoassay (RIA) assay.[72] It was later replaced in October 1995 by an improved assay, which was the Centocor CA-125II RIA assay.[72] All samples tested with the original CA 125 assay were retested using the CA-125II RIA assay.[72] Among the 5371 participants that were in the initial screening of the PLCO trial, the original CA-125 RIA assay had a positivity rate of 0.6% while the CA-125II assay had a positivity rate of 2.4%.[72] In both assays, an abnormal (positive) case was defined as a result of ≥35 U/mL.[28] Up to 45mL of blood was drawn for the CA 125 assay.[28] The CA 125 assay was performed at the initial visit at the entry into the PLCO trial and then annually for five years. The collected blood was centrifuged, then the serum was separated from the clot and frozen within 2-4 hours of blood collection.[28] Samples were run in duplicate. The assay precision was represented by its coefficient of variation (CV).[72] The CVs and its 95% confidence intervals were 4.07% (3.92-4.22) at the lower concentration of 52.7U/mL and 3.78% (3.64-3.92) at the higher concentration of 106.5U/mL. Therefore, these results were in good agreement to those that were reported by the manufacturers in the product inserts.[81] A result of

≥35U/mL was classified as abnormal (positive).[72] Any sample with an abnormal result was re-analyzed to verify the value.[72] Samples that showed discrepant results between duplicate results (CV over 10%) were re-analyzed.[72] Samples were stored at -70°C or colder and were shipped weekly overnight on dry ice to the central laboratory at UCLA.[28] Blood that was not used for the CA 125 test was stored at -70°C in a central repository where it could be used for research in the future, (not limited to just the use by UCLA).[28] The results of the CA 125 test were transmitted electronically by the UCLA laboratory to NCI microcomputers.[28]

3.2.4 Diagnostic and Therapeutic Follow-Up

Participants who received abnormal results on the screening tests were notified within three weeks following the test. These participants were then referred to a qualified medical professional of their choice to receive a definitive clinical investigation and diagnosis and, if needed, treatment. The PLCO cancer screening trial had no direct control over the treatments or interventions given once diagnosis was made.[28]

3.3 - Modeling and Analysis

The following section describes the approach taken to address the study objectives of the current study. Topics among this section will include statistical methodology, data preparation, data cleaning, variable selection, model building, and model evaluation.

3.3.1 Statistical Approach

Data preparation, model building, model evaluation and analysis were performed using Stata 15 statistical software.[83]

3.3.2 Data Preparation

Data cleaning was conducted to prepare data for analysis. Data cleaning involves the process of identifying and managing outliers (values which are unrealistic and not proportional) to prevent biased results. Outliers can be detected visually using scatter plots and box plots using the Stata command *scatter* and command *graph box*. Candidate predictors were assessed for possible outliers, missing data, and implausible values.

3.3.2.1 Candidate Variables

Candidate explanatory variables were selected based on previous literature identifying lung cancer risk factors as well as previous studies from the PLCO cancer screening trial. Candidate variables and potential confounders can be found in Table 4. For all objectives, the primary exposure of interest was CA 125, and the primary outcome of interest was lung cancer. CA 125 was composed of two variables, CA 125 results and CA 125 levels. CA 125 results is a dichotomous variable divided into levels of abnormal (positive) ≥35U/mL and normal (negative) <35 U/mL. CA 125 levels is a continuous variable measuring levels of CA 125. Both variables were tested in separate models. Lung cancer was dichotomized into levels of yes or no. CA 125 levels are presented in the results and were used to draw conclusions for the present study as the continuous variable provides more robust information about the biomarker compared to a dichotomous yes/no variable. Dichotomous associations are also presented because past clinical decision-making has been based on such dichotomous categorizations. The following sociodemographic characteristics were included: age at study entry, race/ethnicity, education and current body mass index (BMI). Potential confounders in this study are categorized under smoking exposures, medical history and co-morbidities. Variables under each category are further described in Table 4.

3.3.3 Descriptive Statistics, Univariate Analysis and Exploratory Analysis

For categorical variables such as race/ethnicity, education, and cigarette smoking status, frequency, percentage and proportion were assessed. For continuous variables such as age, smoking duration and BMI, distributions, measures of central tendency and variability were assessed. For normally distributed quantitative variables, means and standard deviations were calculated. Student's independent sample t-tests were conducted to compare the differences in means between two groups in these variables. For quantitative variables with skewed distribution, medians and interquartile ranges were calculated. If an independent sample t-test, cannot be used with variables that have skewed distribution, a Wilcoxon rank-sum test was used. This non-parametric test identifies if the two populations, where the two samples are selected from, have the same distribution.[84] Once these study covariates were summarized, additional analyses were done stratifying by lung cancer. To evaluate differences in distribution between two categorical variables, chi-square tests of independence were conducted. Univariate associations between each covariate and the outcome of lung cancer were tested to see if any relationships exist. Predictors with a p-value of <0.25 were included into the main logistic regression model for further evaluation.

Exploratory analysis was conducted to determine if any covariates had a relationship with CA 125 levels. Univariate associations between each covariate and CA 125 levels were tested to see if any relationships exist. Predictors with a p-value of <0.10 were included into the main linear regression model. Backward stepwise regression was used to build the final linear regression model, using a p-value of <0.05.

Table 4. Candidate variables and potential confounders for evaluating associations with lung cancer

Variable Categories	Variable Names	Additional Descriptions
Sociodemographic (n=3)	Age	In years
	Race/Ethnicity	White, Non-Hispanic
		Black, Non-Hispanic
		Hispanic
		Asian
		American Indian or Alaskan Native
		Native Hawaiian or Pacific Islander
	Education	Less than high school
		High school graduate
		Post high school training other than college
		Some college
		College graduate
		Post graduate degree
	Current Body Mass Index (BMI)	kg/m^2
Smoking exposure(n=5)	Cigarette smoking status	Never
		Former
		Current
	Smoking intensity	# of cigarettes smoked per day
	Smoking duration	In years
	Quit time	# of years since person quit smoking
	Pack-years	# of packs of cigarettes smoked per day by # of years smoked
Medical history(n=11)	Family history of cancer	Family history of lung cancer?
		No
		Yes
	Personal history of any cancer	Personal history of any cancer at baseline?
		No
		Yes
	Chronic Bronchitis	Ever diagnosed with chronic bronchitis?
		No
		Yes
	Chronic obstructive pulmonary disease (COPD)	Ever diagnosed with COPD?
		No
		Yes
	Emphysema	Ever diagnosed with emphysema?

	No
	Yes
Female Hormone Status	Never
	Former
	Current
Female Hormones	Ever used female hormones?
	No
	Yes
Age at menopause	<40
	41+
Benign or Fibrocystic Breast Disease	No
	Yes
Gallbladder stones/inflammation	Ever had gallbladder stones or inflammation?
	No
	Yes
Colorectal polyps	Ever had colorectal polyps?
	No
	Yes

3.3.4 Model Building Strategy

3.3.4.1 Handling Quantitative Variables

Multivariable logistic regression models were used to address the research objectives due to the dependent variable being dichotomous (lung cancer, yes/no). A manual backwards stepwise regression was used to build final models. If the variables were found to have a p-value of <0.05, they were considered statistically significant and were used in the final model. However, COPD which was found to be significant in a priori literature, was also included in the model regardless of statistical significance. The association between CA 125 and lung cancer was tested in the final multivariable model.

The linearity of continuous predictor variables in the model was assessed by using multivariate fractional polynomials (MFP) using the Stata command *mfp*. This method is considered more flexible in identifying non-linear effects compared to other methods.[85] The following default set of fractional polynomial powers (FP) was used: -2, -1, -0.5, 0 (log), 0.5, 1, 2 and 3. Multi-level ordinal variables used in a model were assessed for the possibility of pseudo-continuity by evaluating the trends in effect sizes when treated as an indicator variable and, if possible, when treated as continuous to reduce the degrees of freedom in models.

3.3.4.2 Assessing Potential Collinearity

The model was tested for correlations to address potential collinearity. Next, variance inflation factors (VIF) and tolerance values (1/VIF) was used to evaluate the degree of multicollinearity amongst all continuous candidate predictor variables. VIFs work by indicating

the factor by which the variance of a certain predictor is inflated due to the lack of independence among all other predictors.

3.3.4.3 Confounders & Interactions

Once the final model was built for each objective, the interaction of interest (smoking status) and confounding of the CA 125 association with lung cancer was tested in the model. A confounding variable was defined as a variable that results in a 15% change in the coefficient of the exposure variable once removed.[86] Variables that were found to be confounders were controlled analytically and kept in final models.

3.4 - Model Evaluation

3.4.1 Fit Diagnostics

The Hosmer-Lemeshow Goodness-of-Fit (GOF) test, McFadden's Pseudo-R^2, Akaike's Information Criteria (AIC) and Bayes Information Criteria (BIC) were used to evaluate the overall model fit. GOF test takes a logistic regression model's fitted probabilities and creates ten groups.[87]Once this process is complete, observed and expected frequencies are analyzed across subgroups.[87]A non-significant GOF test indicates that there is an acceptable agreement between the observed and expected frequencies, concluding that the model correctly fits with the hypothesis being tested.[87] McFadden's Pseudo-R^2, which has also been known as the "likelihood ratio index," compares the proposed model without any predictors to a model with all predictors to determine the fit of the model.[88] AIC provides an estimate of in-sample errors of the proposed model.[89] Schwarz's Bayesian Information Criterion (BIC) produces a similar

outcome to the AIC. The difference between the two model selection tools is that the BIC prefers simpler models when estimating the model performance compared to the AIC.[90]

3.4.2 Influential Observations

Influential observations were investigated in the analysis to determine the effect they have on the model, as well as to possibly identify data entry errors, which could potentially influence the effect estimates. Two methods that were used to investigate observation influence are Pearson standardized residuals and deviance residuals. Pearson standardized residuals produce the differences between observed and expected frequencies, while deviance residuals measure the disagreement between the observed and fitted log-likelihood functions.[91] After the fit is assessed, potential outlying and influential covariates were identified and evaluated using standardized Pearson residuals.

CHAPTER IV - RESULTS

This chapter describes the study participants and summarizes the results found studying the associations between CA 125 and lung cancer along with predictor variables. The chapter explores how these associations between CA 125 and lung cancer are affected when stratified by smoking status (never/ever) and lung cancer stage (early/advanced). Moreover, exploratory analysis will be presented to show the relationship between important factors and CA 125. Model assumptions were evaluated, and goodness of model fits were determined.

4.1 Data Preparation

There was less than 10% missingness in the predictor variables, therefore, multiple imputations for missing data were not performed. There were no extreme values or outliers detected in explanatory variables. CA 125 levels were natural log transformed according to the results in Stata's *ladder of power* command, as CA 125 levels was not normally distributed and natural log transformation provided the best normalization. After transforming, CA 125 levels were normally distributed, therefore, bootstrapping was not needed to determine confidence intervals.

4.2 Sample Characteristics

There were 23,938 participants with no lung cancer and 578 participants with lung cancer diagnosed during the follow-up period. The start of follow up for the earliest participants began around October 1993 and was continued until April 2001. The characteristics of study participants with and without lung cancer are presented in Table 5. Univariate logistic regressions were carried out for all explanatory variables. Univariate odds ratios, 95%

confidence interval estimates and p-values for unadjusted effects are shown in Table 5. The results of univariate analyses informed which predictors were to be considered in multivariable analysis.

Table 5. Characteristics of overall participants by lung cancer and univariate logistic associations with lung cancer

Explanatory Variables	No lung cancer (n=29,938)	Lung cancer (n=576)	Univariate odds ratio (95% CI, *P-value*)
Age, year, mean (SD)	62.38 (5.36)	64.07 (5.21)	1.06 (1.04-1.07, P<0.001)
Race/ethnicity, number (%)			
Non-Hispanic White	26,574 (98.06)	526 (1.94)	Reference group
Non-Hispanic Black	1,622 (98.18)	30 (1.82)	0.68 (0.30-1.52, P=0.345)
Hispanic	488 (98.68)	6 (1.32)	0.52 (0.29-0.95, P=0.034)
Asian	1,063 (98.98)	11 (1.02)	1.12 (0.36-3.54, P=0.843)
Pacific Islander	135 (97.83)	3 (2.17)	n/a
American Indian/Alaskan Native	86 (100.00)	0 (0.00)	n/a
Education, number (%)			0.87 (0.83-0.92, P<0.001)
Less than high school	1,746 (96.20)	69 (3.80)	Reference group
High school graduate	8,036 (98.04)	161 (1.96)	0.51 (0.38-0.68, P<0.001)
Post High school training	3,865 (98.12)	74 (1.88)	0.48 (0.35-0.68, P<0.001)
Some college	6,898 (98.14)	131 (1.86)	0.48 (0.36-0.65, P<0.001)
College graduate	4,786 (98.56)	70 (1.44)	0.37 (0.26-0.52, P<0.001)
Postgraduate degree	4,566 (98.51)	69 (1.49)	0.38 (0.27-0.54, P<0.001)
Current Body Mass Index (BMI), mean (SD)	27.08 (5.47)	26.02 (4.66)	0.96 (0.94-0.98, P<0.001)
Family history of Lung cancer, self-reported (%)			
No	25,727 (98.36)	428 (1.64)	Reference group
Yes	3,401 (96.73)	115 (3.27)	2.13 (1.75-2.58, P<0.001)
Personal history of cancer, self-reported (%)			
No	28,011 (98.21)	511 (1.79)	Reference group
Yes	1,924 (96.73)	65 (3.27)	1.85 (1.42-2.40, P<0.001)
COPD, self-reported, number (%)			
No	27,832 (98.33)	472 (1.62)	Reference group
Yes	1,980 (95.33)	97 (4.67)	2.89 (2.31-3.61, P<0.001)
Chronic Bronchitis, self-reported, number (%)			
No	28,162 (98.27)	497 (1.73)	Reference group
Yes	1,651 (95.82)	72 (4.18)	2.47 (1.92-3.18, P<0.001)

Table 5, continued. Characteristics of overall participants by lung cancer and univariate logistic associations with lung cancer

Explanatory Variables	No lung cancer (n=29,938)	Lung cancer (n=576)	Univariate odds ratio (95% CI, *P-value*)
Emphysema, self-reported, number (%)			
No	29,321 (98.26)	518 (1.74)	Reference group
Yes	518 (90.88)	52 (9.12)	5.68 (4.22-7.66, P<0.001)
Smoking status, number (%)			3.95 (3.52-4.43, P<0.001)
Never	16,927 (99.55)	76 (0.45)	Reference group
Former	10,339 (97.34)	283 (2.66)	6.09 (4.73-7.86, P<0.001)
Current	2,660 (92.46)	217 (7.54)	28.26 (13.95-23.67, P<0.001)
Cigarettes smoked per day, mean (SD)	20.35 (12.13)	24.83 (12.03)	1.02 (1.01-1.03, P<0.001)
Smoking duration, year, mean (SD)	26.47 (13.76)	38.19 (10.44)	1.08 (1.07-1.09, P<0.001)
Smoking Quit-time, year, mean (SD)	19.89 (12.07)	12.57 (9.90)	0.94 (0.93-0.95, P<0.001)
Pack-years, mean (SD)	28.76 (24.34)	47.59 (26.51)	1.02 (1.01-1.02, P<0.001)
Female hormone status*, number (%)			0.85 (0.78-0.94, P<0.001)
Never	10,516 (97.80)	237 (2.20)	Reference group
Former	5,138 (98.02)	104 (1.98)	0.90 (0.71-1.13,P=0.366)
Current	14,284 (98.38)	235 (1.62)	0.73 (0.61-0.88, P<0.001)
Female hormones†, number (%)			
No	10,516 (97.80)	237 (2.20)	Reference group
Yes	19,422 (98.28)	339 (1.72)	0.77 (0.65-0.92, P=0.003)
Age at menopause (<40 vs 40+)			
<40	25,673 (98.18)	475 (1.82)	Reference group
40+	4,015 (97.69)	95 (2.31)	1.28 (1.02-1.59, P=0.030)
Benign or fibrocystic breast disease, number (%)			
No	21,101 (98.00)	431 (2.00)	Reference group
Yes	8,367 (98.41)	135 (1.59)	0.79 (0.65-0.96, P=0.018)

Table 5, continued. Characteristics of overall participants by lung cancer and univariate logistic associations with lung cancer

Explanatory Variables	No lung cancer (n=29,938)	Lung cancer (n=576)	Univariate odds ratio (95% CI; *P-value*)
Gallbladder Inflammation, number (%)			
No	25,219 (98.07)	497 (1.93)	Reference group
Yes	4,581 (98.47)	71 (1.53)	0.79 (0.61-1.01, P=0.060)
Colorectal polyps, number (%)			
No	28,123 (98.21)	512 (1.79)	Reference group
Yes	1,671 (96.70)	57 (3.30)	1.87 (1.42-2.47, P<0.001)

*Female Hormone Status: "Have you ever used female hormones (tablets, pills or creams) for menopause?"

† Female Hormone: "Are you currently using female hormones?"

Abbreviations: COPD: Chronic Obstructive Pulmonary Disease

4.3 Predictor Associations with Lung Cancer

4.3.1 Sociodemographic

Participants' ages at study entry ranged from 52 to 78 years (mean=62.4; SD=5.4). Age was related to a higher odds of lung cancer. The majority of participants (88.8%) were Non-Hispanic White. Being Hispanic was found to have a significant protective effect against lung cancer (OR: 0.52; 95% CI: 0.29-0.95, P=0.034). Overall, education had a protective effect against lung cancer (0.87; 95% CI: 0.83-0.92, P<0.001).

4.3.2 Medical History

The current BMI in participants with no lung cancer was 1.06 kg/m^2 higher than for participants with lung cancer, indicating an inverse association with lung cancer (OR: 0.96; 95% CI: 0.94-0.98, P<0.001). Family history of lung cancer, personal history of cancer, COPD, chronic bronchitis, emphysema and colorectal polyps were all related to a higher odds of lung cancer (P<0.001, See Table 5).

4.3.3 Exposures

Current female hormone use and ever use of female hormones demonstrated a protective effect against lung cancer. On average, participants with no lung cancer smoked 20.4 cigarettes per day and for approximately 26.5 years. Participants with lung cancer smoked 24.8 cigarettes per day on average and for approximately 38.2 years. Therefore, participants with lung cancer, on average, smoked 4.48 cigarettes/day more (OR: 1.02; 95% CI: 1.01-1.03, P<0.001) and 11.7 years longer than participants with no lung cancer (OR: 1.08; 95% CI 1.07-1.09, P<0.001). In individuals who previously smoked, the average number of years since

smoking cessation in participants with no lung cancer was 19.9 years and 12.6 years in participants with lung cancer. Therefore, participants with no lung cancer had a smoking cessation period that was 7.3 years longer than in participants with lung cancer (OR: 0.94; 95% CI: 0.93-0.95, P<0.001).

Stata's *ladder of power* command showed natural log transformation was the most appropriate for CA 125 levels. Table 5 results of univariate analyses served as a guide for variable selection in multivariable analysis. A summary of univariate logistic regression for lung cancer and dichotomous CA 125 results (positive vs. negative) by screening rounds are provided in Table 6. Table 7 summarizes univariate logistic regression for lung cancer and log transformed continuous CA 125 levels by screening rounds.

Table 6. Univariate logistic regression for lung cancer (yes vs no) and dichotomous CA 125 results* by screening rounds (T0-T5)

Variables	No lung cancer (n=29,938) (%)	Lung cancer (n=576) (%)	Odds ratio (95% CI, P-value)
CA 125 Result T0			
Negative	27,786 (98.14)	527 (1.86)	Reference group
Positive	385 (96.25)	15 (3.75)	2.05 (1.21-3.46, P=0.007)
CA 125 Result T1			
Negative	26,444 (98.27)	466 (1.73)	Reference group
Positive	412 (95.81)	18 (4.19)	2.47 (1.53-4.01, P<0.001)
CA 125 Result T2			
Negative	25,460 (98.35)	428 (1.65)	Reference group
Positive	459 (97.04)	14 (2.96)	1.81 (1.05-3.11, P=0.031)
CA 125 Result T3			
Negative	24,381 (98.47)	380 (1.53)	Reference group
Positive	408 (95.77)	18 (4.23)	2.83 (1.74-4.58, P<0.001)
CA 125 Result T4			
Negative	19,380 (98.71)	254 (1.29)	Reference group
Positive	313 (95.72)	14 (4.28)	3.41 (1.96-5.91, P<0.001)
CA 125 Result T5			
Negative	21,348 (98.61)	302 (1.39)	Reference group
Positive	350 (96.15)	14 (3.85)	2.83 (1.63-4.88, P<0.001)

*CA 125 results are divided into levels of abnormal (positive) ≥35 U/mL and normal (negative) <35 U/mL

Table 7. Univariate logistic regression for lung cancer (yes vs no) and LN CA 125 levels by screening rounds (T0-T5)

Variables	N total	No lung cancer LN CA 125 Mean (SD)	Lung cancer LN CA 125 Mean (SD)	Combined Mean (SD)	Odds ratio (95% CI, P-value)
CA 125 Level T0	28,712	12.27 (21.46)	12.89 (10.82)	12.28 (21.31)	1.16 (0.99-1.37, P=0.066)
CA 125 Level T1	27,340	12.17 (31.19)	14.46 (18.89)	12.21 (31.02)	1.34 (1.13-1.59, P<0.001)
CA 125 Level T2	26,361	12.31 (25.05)	13.38 (11.99)	12.33 (24.89)	1.26 (1.05-1.53, P=0.010)
CA 125 Level T3	25,187	12.29 (32.94)	15.19 (19.53)	12.33 (32.77)	1.48 (1.23-1.78, P<0.001)
CA 125 Level T4	19,961	12.14 (11.69)	14.56 (13.28)	12.17 (11.72)	1.58 (1.25-1.97, P<0.001)
CA 125 Level T5	22,014	12.13 (10.59)	14.37 (11.49)	12.16 (10.61)	1.61 (1.31-1.99, P<0.001)

Abbreviation: LN: Natural log-transformed

4.4 Key Study Findings

4.4.1 Summary of Tables

The associations between explanatory variables and lung cancer were evaluated using multivariable logistic regressions. To give an example of the performance of the models, results from screening round year 3 for dichotomous CA 125 results (yes/no) and log transformed (LN) continuous CA 125 levels are presented in detail in this chapter (See Tables 8 and 9). Supplementary tables for each model for dichotomous CA 125 results (yes/no) and log transformed (LN) continuous CA 125 levels by study year are in the Appendix (S1-S12).

Table 8. Multivariable logistic regression for lung cancer and CA 125 dichotomous results* in screening round T3

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
CA 125 Results, positive vs negative	2.20 (1.30-3.72; 0.003)
Age (years)	1.05 (1.03-1.07; <0.001)
Education, self-reported level of,	0.93 (0.87-1.00; 0.047)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-0.99; 0.011)
Family history of lung cancer, self-reported, yes vs no	1.10 (1.04-1.16; 0.001)
Personal history of cancer, self-reported, yes vs no	1.60 (1.15-2.21; 0.005)
COPD, self-reported, yes vs no	1.05 (0.76-1.44; 0.781)
Cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.06; <0.001)
n total; AIC, BIC	24,471; 3341.213, 3422.265
P of overall model, GOF P-value	<0.001, 0.0029
Pseudo R^2	0.1544
Area under ROC curve	0.8233

*CA 125 results are divided into levels of abnormal (positive) ≥35 U/mL and normal (negative) <35 U/mL
Abbreviation: AIC: Akaike information criterion; BIC: Bayesian information criterion; COPD: Chronic Obstructive Pulmonary Disease; GOF: Goodness-of-Fit test, LN: Natural log-transformed

Table 9. Multivariable logistic regression for lung cancer and LN CA 125 levels in screening round T3

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
LN CA 125 Levels	1.37 (1.13-1.67; <0.001)
Age (years)	1.05 (1.03-1.07; <0.001)
Education, self-reported level of,	0.93 (0.87-0.99; 0.042)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-1.00; 0.012)
Lung cancer, self-reported family history of, yes vs no	1.10 (1.04-1.16; <0.001)
Any type of cancer, self-reported personal, yes vs no	1.58 (1.14-2.19; 0.006)
COPD, self-reported history of, yes vs no	1.03 (0.75-1.43; 0.836)
How many cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total; AIC, BIC	24,471; 3338.684, 3419.736
P of overall model, GOF P-value	<0.001, 0.2249
Pseudo R^2	0.1551
Area under ROC curve	0.8234

Abbreviation: AIC: Akaike information criterion; BIC: Bayesian information criterion; COPD: Chronic Obstructive Pulmonary Disease; GOF: Goodness-of-Fit test, LN: Natural log-transformed

4.4.2 Summary of Predictor Variable Findings for Lung Cancer~CA 125

Race/ethnicity, chronic bronchitis, emphysema, female hormone status (current/former), ever use of female hormones, age at menopause, benign or fibrocystic breast disease, gallbladder disease, colorectal polyps, smoking status, smoking quit time in individuals who previously smoked were removed from the lung cancer~CA125 model as they did not approach statistical significance. Each multivariable logistic regression model for lung cancer~CA 125 was adjusted for education, age, current body mass index (BMI), family history of lung cancer, personal history of cancer, chronic obstructive pulmonary disease (COPD), average number of cigarettes smoked per day and number of years smoked. Despite being statistically insignificant in some models, COPD was added into the multivariable logistic regression models to make the models comparable (See Appendix). Education, and current BMI had a protective effect against lung cancer across all 6 screening round models for dichotomous CA 125 results (positive vs. negative) and continuous LN CA 125 levels (T0-T5). Age, family history of lung cancer, personal history of cancer, average number of cigarettes smoked per days, number of years smoked were all associated with a high odds of lung cancer in dichotomous CA 125 results (positive vs. negative) and continuous LN CA 125 levels for all screening rounds (T0-T5). Each screening round model by dichotomous CA 125 results (positive vs. negative) or continuous LN CA 125 levels was overall statistically significant ($P<0.001$). Lung cancer~ CA 125 dichotomous results were statistically significant for screening rounds T1, T3, and T4 (Figure 2). Lung cancer~LN continuous CA 125 levels results were statistically significant for screening rounds T1, T3, T4 and T5 (Figure 3).

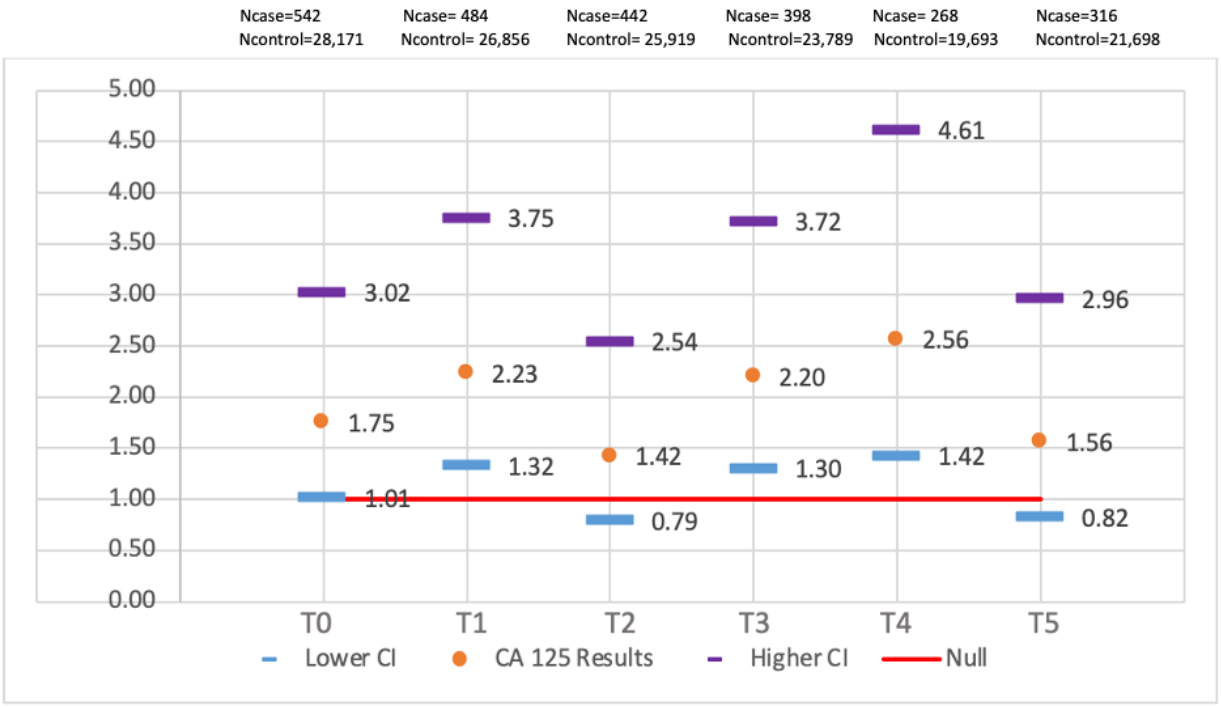

Figure 2. Odds ratio for lung cancer by dichotomous CA 125 results (positive/negative) by screening rounds T0-T5

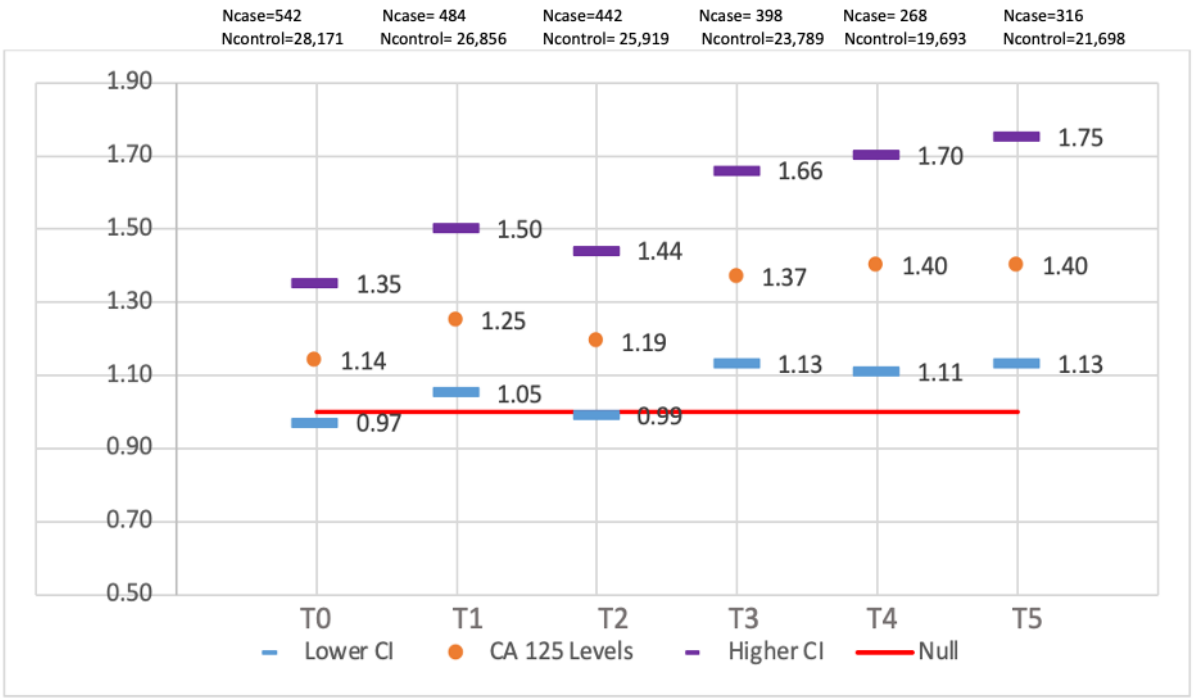

Figure 3. Odds ratio for lung cancer by LN CA 125 levels by screening rounds T0-T5

4.4.3 Smoking Status – Individuals who smoked vs. individuals who never smoked among total study sample

Table 10 summarizes the odds ratio results of multivariable logistic regression by predictors CA 125 dichotomous results (positive/negative) and continuous log-transformed (LN) CA 125 levels by smoking status (Individuals who never smoked/Individuals who smoked/Combined) for screening round T3. When smoking status was combined (Individuals who never smoked/Individuals who smoked), there were statistically significant associations between lung cancer and dichotomous CA 125 results (positive vs. negative) among screening rounds T0, T1, T3 and T4 (See Table 10). Similarly, there were statistically significant associations between lung cancer and continuous LN CA 125 levels for combined smoking status in screening rounds T1-T5 (See Table 10). For lung cancer and dichotomous CA 125 results (positive vs. negative) among individuals who smoked, statistical significance associations were found among screening rounds T1, T3, and T4. (See Table 10). Among individuals who smoked, there was a statistically significant higher risk of lung cancer associated with continuous LN CA 125 levels in screening rounds T1, T3, T4, and T5 (See Table 10). Among individuals who never smoked, there was no statistically significant association between lung cancer and dichotomous CA 125 results. There was statistically significant association between lung cancer and continuous LN CA 125 levels only in screening round T3 (See Figure 4 and 5). Despite non-significance, the direction of effect, magnitude of effect and consistency among the results suggest that there could be an association.

Table 10. Multivariable* logistic regression odds ratios for lung cancer and predictors CA 125 dichotomous† results and log transformed continuous CA 125 levels, by smoking status

	Individuals who never smoked		Individuals who smoked		Combined	
Screening Rounds	**CA 125 Results OR (95% CI; *P-value*)**	**LN CA 125 Levels OR (95% CI; *P-value*)**	**CA 125 Results OR (95% CI; *P-value*)**	**LN CA 125 Levels OR (95% CI; *P-value*)**	**CA 125 Results OR (95% CI; *P-value*)**	**LN CA 125 Levels OR (95% CI; *P-value*)**
T0	2.24 (0.54-9.25; 0.264)	1.20 (0.76-1.89; 0.442)	1.72 (0.95-3.11; 0.075)	1.15 (0.96-1.37; 0.140)	1.75 (1.01-3.02; 0.047)	1.14 (0.97-1.35; 0.113)
T1	2.06 (0.50-8.52; 0.318)	1.32 (0.82-2.14; 0.249)	2.32 (1.32-4.07; 0.003)	1.25 (1.04-1.51; 0.020)	2.23 (1.32-3.75; 0.003)	1.25 (1.05-1.50; 0.012)
T2	0.96 (0.13-6.97; 0.965)	1.38 (0.85-2.25; 0.192)	1.50 (0.81-2.77; 0.193)	1.17 (0.96-1.44; 0.128)	1.42 (0.09-2.54; 0.240)	1.19 (0.99-1.44; 0.068)
T3	2.06 (0.50-8.60; 0.320)	**1.61 (1.03-2.51; 0.036)**	2.25 (1.28-3.98; 0.005)	1.33 (1.08-1.65; 0.008)	2.20 (1.30-3.72; 0.003)	1.37 (1.13-1.66; 0.001)
T4	N/A	1.46 (0.79-2.71; 0.228)	3.00 (1.63-5.51; <0.001)	1.39 (1.08-1.80; 0.011)	2.56 (1.42-4.61; 0.002)	1.40 (1.11-1.77; 0.005)
T5	N/A	1.29 (0.74-2.25; 0.375)	1.76 (0.92-3.38; 0.089)	1.43 (1.13-1.82; 0.003)	1.56 (0.82-2.96; 0.171)	1.40 (1.13-1.75; 0.003)

Abbreviations: CI: Confidence Intervals, LN: Natural log-transformed; OR: Odds Ratio.

*Models adjusted for: age, education, current body mass index (BMI), family history of lung cancer, personal history of cancer, chronic obstructive pulmonary disease (COPD), number of cigarettes smoked per day, number of years smoked

† CA 125 results are divided into levels of abnormal (positive) ≥35 U/mL and normal (negative) <35 U/mL

N/A: CA 125 results did not resolve in the model due to 0 being in the denominator

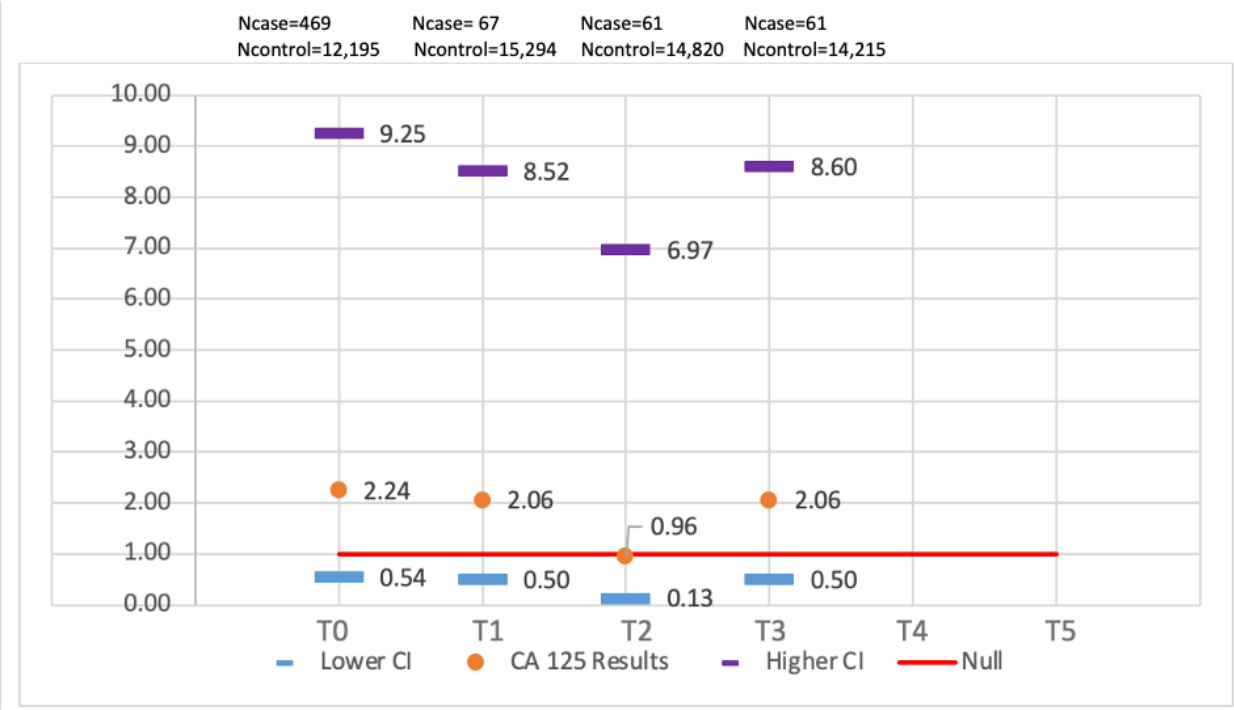

Figure 4. Odds ratio for lung cancer by dichotomous CA 125 results (positive/negative) by screening rounds T0-T5 among individuals who never smoked

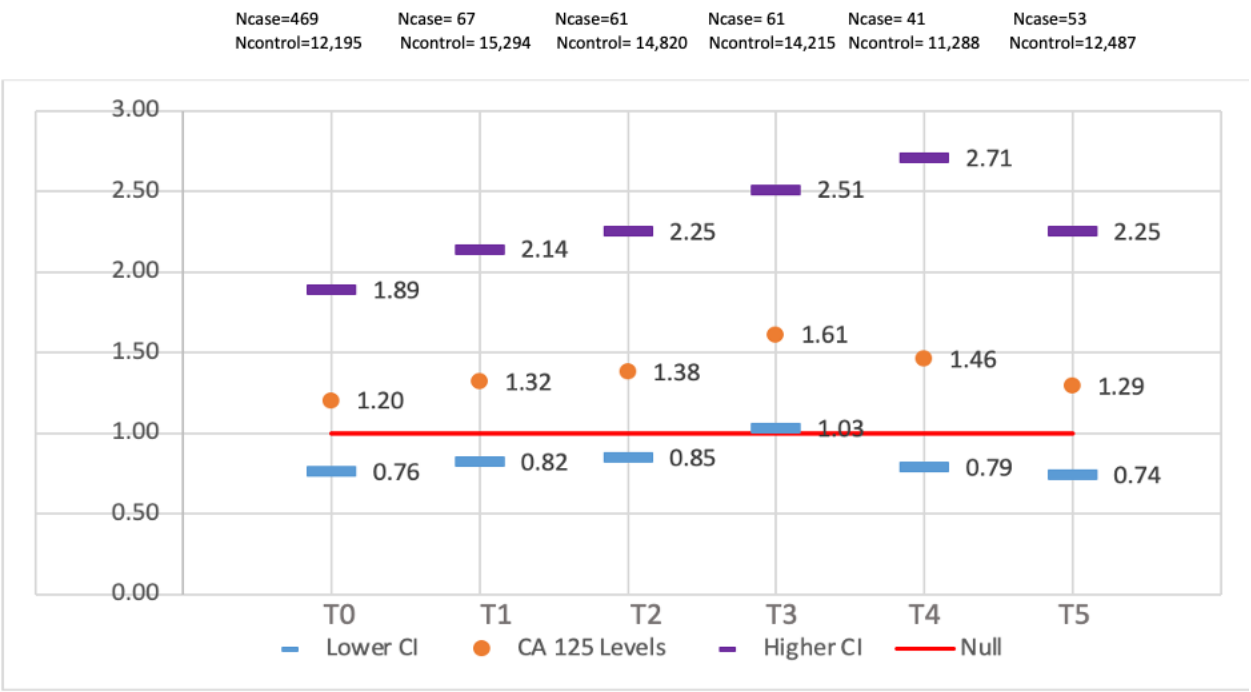

Figure 5. Odds ratio for lung cancer by LN CA 125 levels by screening rounds T0-T5 among individuals who never smoked

The interactions between smoking status and dichotomous CA 125 results (positive vs. negative) and smoking status and continuous log transformed (LN) CA 125 levels for lung cancer were not statistically significant (Table 11). Therefore, it can be concluded that smoking status does not affect the association between lung cancer and CA 125.

Table 11. CA 125 (dichotomous and log transformed continuous) interaction with smoking status (individuals who smoked vs individuals who never smoked) in multivariable* logistic regression for lung cancer

Screening Rounds	Smoking status by CA 125 results interaction term OR (95% CI; *P-value*)	Smoking status by LN CA 125 levels interaction term OR (95% CI; *P-value*)
T0	0.86 (0.19-3.98; 0.845)	0.98 (0.60-1.59; 0.930)
T1	1.24 (0.27-5.66; 0.786)	0.96 (0.58-1.60; 0.890)
T2	1.65 (0.21-13.13; 0.636)	0.87 (0.51-1.47; 0.601)
T3	1.22 (0.26-5.61; 0.800)	0.86 (0.52-1.41; 0.548)
T4	N/A	1.02 (0.53-1.98; 0.944)
T5	N/A	1.17 (0.64-2.14; 0.605)

Abbreviations: CI: Confidence Intervals, LN: Natural log-transformed, OR: Odds Ratio.
*Models adjusted for: age, education, current body mass index (BMI), family history of lung cancer, personal history of cancer, chronic obstructive pulmonary disease (COPD), number of cigarettes smoked per day, number of years smoked
N/A: CA 125 results did not resolve in the model due to 0 being in the denominator

4.4.4 Early Stage vs. Advanced Stage Lung Cancer among Total Study Sample

For dichotomous CA 125 results (positive vs. negative) and early-stage lung cancer, statistical significance was found in screening rounds T0, T1, T3 and T4 (See Figure 6). For continuous log transformed (LN) CA 125 levels and early-stage lung cancer, statistical significance was found in all screening rounds (See Figure 7). For advanced stage lung cancer and dichotomous CA 125 results (positive vs. negative), statistical significance was found among screening rounds T3 and T4 (See Table 12). Among continuous LN CA 125 levels and advanced stage lung cancer, statistical significance was found among T3, T4 and T5 (See Table 13). Therefore, an important finding was that the biomarker CA 125 may be useful for early detection of lung cancer.

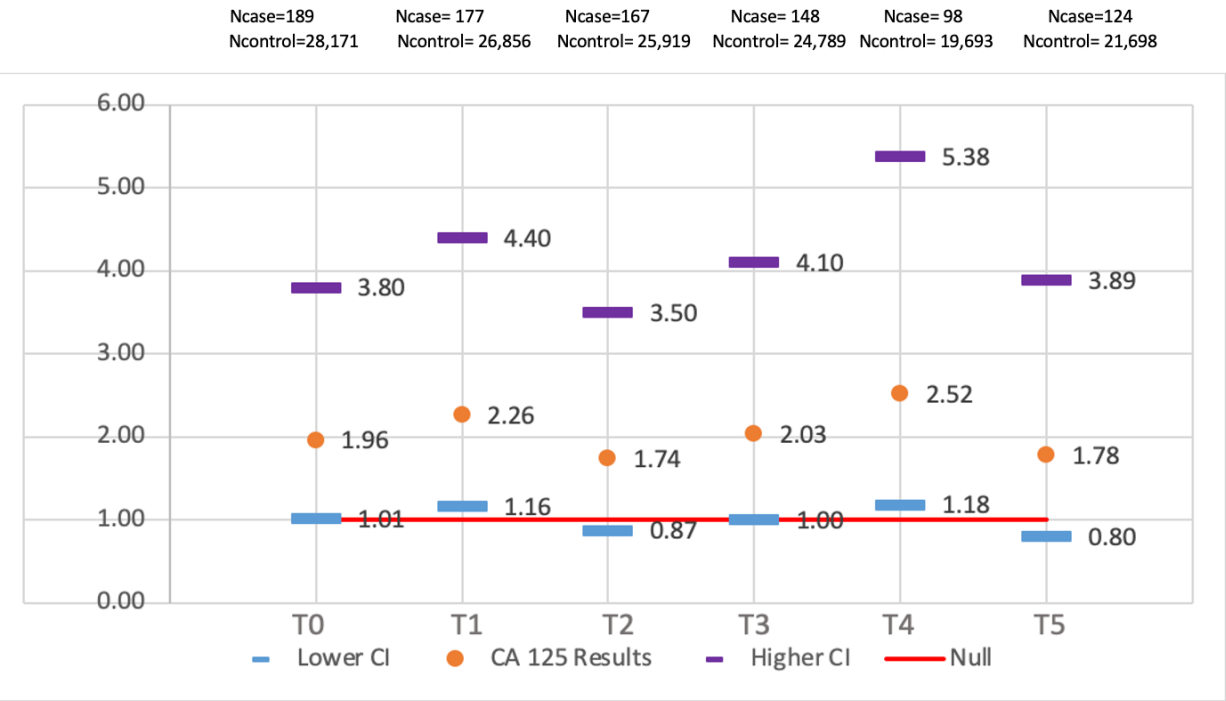

Figure 6. Odds ratio for lung cancer by dichotomous CA 125 results (positive/negative) by screening rounds T0-T5 by early-stage lung cancer

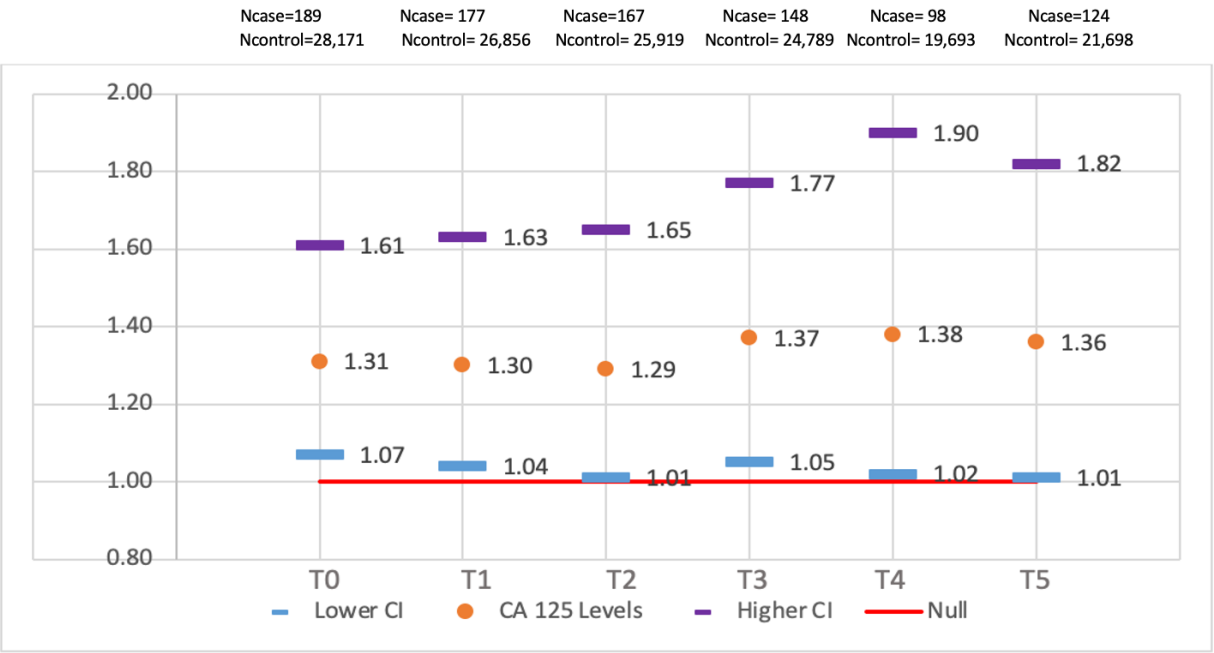

Figure 7. Odds ratio for lung cancer by LN CA 125 levels by screening rounds T0-T5 by early-stage lung cancer

Table 12. Multivariable* logistic regression odds ratios for lung cancer and predictors CA 125 dichotomous† results by lung cancer stage, early vs advanced

Screening Rounds	Early stage‡ lung cancer			Advanced stage‡ lung cancer		
	OR (95% CI; *P-value*)	n total	LC total	OR (95% CI; *P-value*)	n total	LC total
T0	1.96 (1.01-3.80; 0.047)	28,360	189	1.47 (0.59-3.64; 0.408)	28,446	275
T1	2.26 (1.16-4.40; 0.016)	27,033	177	2.13 (0.98-4.64; 0.057)	27,092	236
T2	1.74 (0.87-3.50; 0.118)	26,086	167	1.00 (0.36-2.73; 0.995)	26,134	215
T3	2.03 (1.00-4.10; 0.049)	24,937	148	2.40 (1.14-5.00; 0.021)	24,983	194
T4	2.52 (1.18-5.38; 0.017)	19,791	98	2.58 (1.09-6.10; 0.031)	19,821	128
T5	1.78 (0.80-3.89; 0.159)	21,822	124	1.26 (0.45-3.41; 0.654)	21,845	147

Abbreviations: CI: Confidence Intervals, LC: Lung Cancer, OR: Odds Ratio
*Models adjusted for age, education, current body mass index (BMI), family history of lung cancer, personal history of cancer, chronic obstructive pulmonary disease (COPD), number of cigarettes smoked per day, number of years smoked
† CA 125 results are divided into levels of abnormal (positive) ≥35 U/mL and normal (negative) <35 U/mL
‡ Early-stage lung cancer includes stages 1, 1A, 1B, 2A, and 2B; Advanced stage lung cancer includes stages 3A, 3B and 4.

Table 13. Multivariable* logistic regression odds ratios for lung cancer and predictors log transformed continuous CA 125 levels by lung cancer stage, early vs advanced

Screening Rounds	Early stage† lung cancer			Advanced stage† lung cancer		
	OR (95% CI; *P-value*)	n total	LC total	OR (95% CI; *P-value*)	n total	LC total
T0	1.31 (1.07-1.61; 0.010)	28,360	189	0.93 (0.71-1.21; 0.578)	28,446	275
T1	1.30 (1.04-1.63; 0.020)	27,033	177	1.19 (0.91-1.55; 0.202)	27,092	236
T2	1.29 (1.01-1.65; 0.042)	26,086	167	1.08 (0.81-1.43; 0.595)	26,134	215
T3	1.37 (1.05-1.77; 0.019)	24,937	148	1.40 (1.06-1.86; 0.019)	24,983	194
T4	1.38 (1.02-1.90; 0.040)	19,791	98	1.44 (0.02-2.04; 0.040)	19,821	128
T5	1.36 (1.01-1.82; 0.042)	21,822	124	1.46 (1.06-2.01; 0.019)	21,845	147

Abbreviations: CI: Confidence Intervals, LC: Lung cancer, OR: Odds Ratio

*Models adjusted for age, education, current body mass index (BMI), family history of lung cancer, personal history of cancer, chronic obstructive pulmonary disease (COPD), number of cigarettes smoked per day, number of years smoked

†Early-stage lung cancer includes stages 1, 1A, 1B, 2A, and 2B; Advanced stage lung cancer includes stages 3A, 3B and 4.

4.4.5 Prediction Models

Year T3 was used to examine the sensitivity, specificity and ROC curves of CA 125 levels (See Table 14). In the univariate logistic regression results for lung cancer~LN continuous CA 125 levels, when the probability for positivity was set to P≥0.0135, the sensitivity was 78.39% (95% CI: 0.74-0.82), specificity was 23.87% (95% CI: 0.23-0.24), the positive predictive value was 1.63%, and the negative predictive value was 98.57%. The probability for positivity threshold was set to P≥0.0135 in order for the study's results to be comparable with another biomarker which will be mentioned later in the Discussion chapter. The area under the curve (AUC) of the lung cancer prediction models with and without CA 125 levels added was 0.8234 and 0.8225 respectively.

Table 14. Comparison of ROC curve of prediction models with and without CA 125

	N total	ROC Curve	Std. Error	95% CI
Model with CA 125 levels	24,471	0.8234	0.0115	0.80-0.85
Model without CA 125 levels	24,471	0.8225	0.0116	0.80-0.85
P-value	0.64			

4.4.3 Exploratory Analysis

Table 15 summarizes the exploratory analysis of continuous log transformed (LN) CA 125 levels and potential covariates using multivariable linear regression for screening round T3. In this study's exploratory analysis, the overall multivariable linear regression model between continuous LN CA 125 levels and covariates was statistically significant (P<0.001) (Table 15). Age, current BMI, family history of lung cancer, personal history of cancer, COPD, number of years smoked, and smoking status were all associated with continuous LN CA 125 (P<0.001) (Table 15).

Table 15. Multivariable linear regression for LN CA 125 levels and potential covariates in screening round T3

Explanatory Variables	**Beta Coefficient (95% Confidence Interval; *P-value*)**
Age (years)	0.004 (0.003-0.006; <0.001)
Current Body Mass Index (BMI) (kg/m^2)	-0.006 (-0.007- -0.005;<0.001)
Number of cigarettes smoked per day	0.0005 (0.00007- -0.001; 0.025)
COPD, self-reported history of, yes vs no	0.097 (0.072-0.123; <0.001)
Any type of cancer, self-reported personal, yes vs no	0.041(0.016-0.067; 0.001)
n total	24,824
P of overall model	<0.001
Adjusted R^2	0.0093

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease; GOF: Goodness-of-Fit test, LN: Natural log-transformed

4.5 Assumption Checking

4.5.1 Independence of Errors

Observations were independent of one another in all models, participants were not entered more than once.

4.5.2 Assessment of Non-Linear Associations

For all logistic regressions, possible non-linear associations were considered by conducting multivariate fractional polynomial (MFP) analysis. Generally, there were no strong non-linear associations found among the variables.

4.5.3 Assessment of Collinearity

Collinearity was evaluated with variance inflation factors (VIFs) and tolerance values (Table 16). There was a mean VIF of 1.30. Tolerances were substantially greater than 0.1. The smallest tolerance was 0.466. Since the VIFs for all explanatory variables were near 1.0 and were not above 10 (VIF cut-off), collinearity was not a major concern for the models.

4.5.4 Lack of Influential Outliers

As shown in Figure 8 and 9, influential outliers were inspected by plotting Pearson standardized residual for the models. Influential outliers were not a cause of concern in the model.

Table 16. Collinearity evaluation for model

Explanatory Variables	Variance Inflation Factors (VIF)	Tolerance
Age (years)	1.03	0.968
Education, self-reported level of	1.04	0.963
Current Body Mass Index (BMI) (kg/m^2)	1.03	0.973
Lung cancer, self-reported family history of, yes vs no	1.01	0.994
Any type of cancer, self-reported personal, yes vs no	1.00	0.997
COPD, self-reported history of, yes vs no	1.03	0.968
How many cigarettes smoked per day	2.15	0.466
Number of years smoking	2.15	0.466
Mean VIF	1.30	

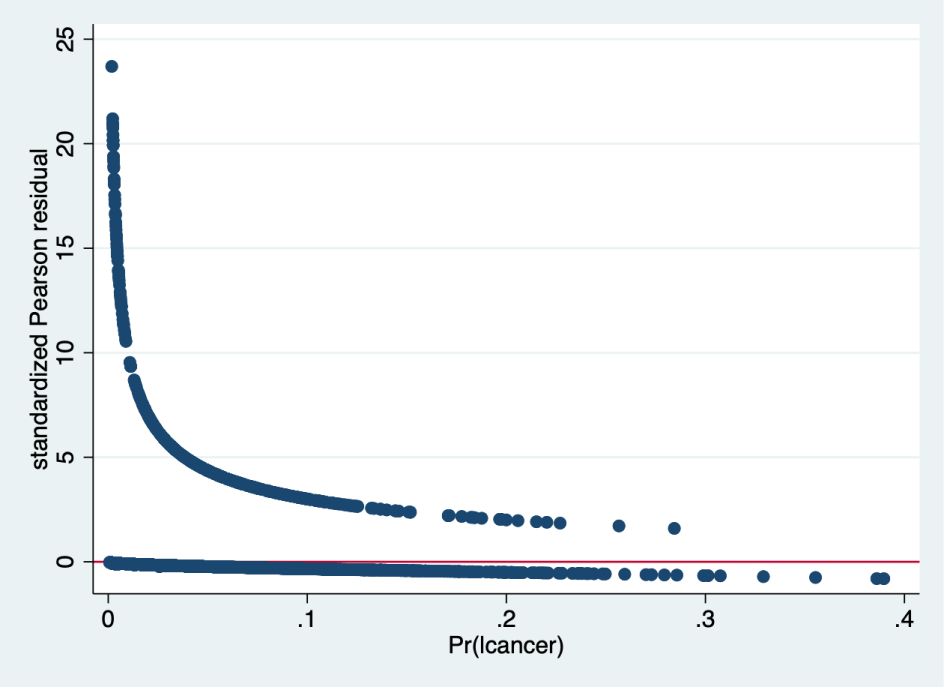

Figure 8. Inspection of influential observations with Pearson standardized residual in T3 dichotomous CA125 results model

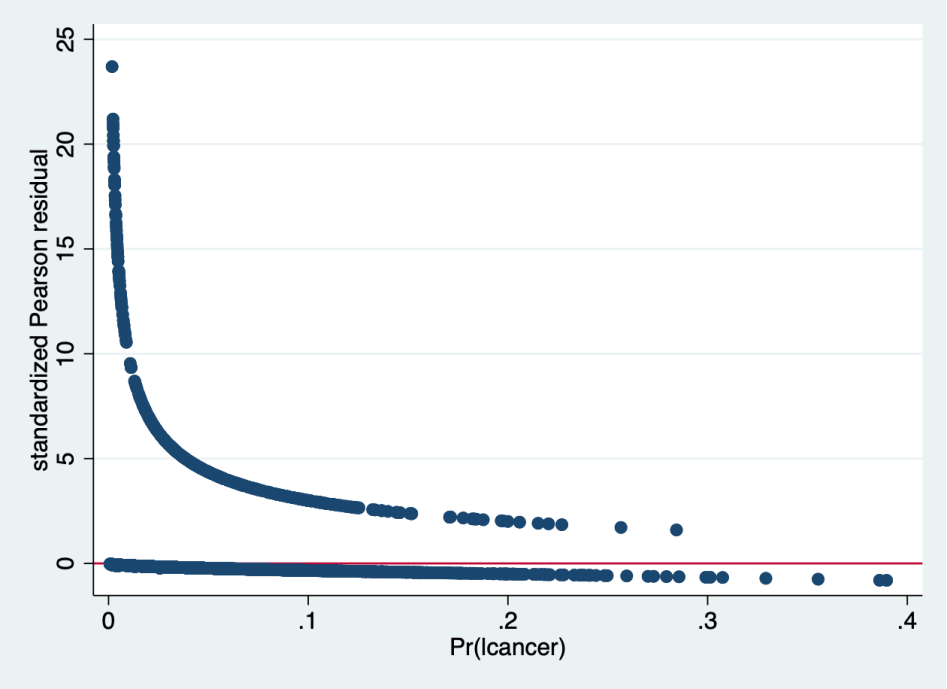

Figure 9. Inspection of influential observations with Pearson standardized residual in T3 log transformed CA125 Levels model

4.6 Fit Diagnostics

The Hosmer-Lemshow Goodness-of Fit (GOF) test, Akaike information criterion (AIC), Bayesian information criterion (BIC) and R^2 or pseudo-R^2 of intermediate analysis were evaluated to find optimal models. AIC/BIC results for each model can be found in the supplementary tables (See Appendix). GOF test results were statistically significant in dichotomous CA 125 results T0-T5 and continuous CA 125 levels T0 and T1. However, for T2-T5 for continuous CA 125 levels, GOF tests were not statistically significant. Since this study has a large sample size, small trivial differences can be significant which can make a model appear to lack goodness of fit. It is also possible that due to the inclusion of the explanatory variable, COPD, the GOF test became statistically significant, in turn causing some of the models to lack goodness of fit as well.

CHAPTER 5 - DISCUSSION

This chapter will explore how this study builds upon previous literature on CA 125 and lung cancer as well as the impact the study findings have on the potential uses of this biomarker in the early detection of lung cancer, and in turn public health.

5.1 Main Findings

This study examined the relationship between CA 125 and lung cancer and the impact of smoking status and lung cancer stage on this association. The main findings of this study include: i) CA 125 was significantly and independently associated with lung cancer; ii) CA 125 was found to be associated with early-stage lung cancer; iii) there might be a relationship between CA 125 and lung cancer in individuals who never smoked; and iv) there was an association between CA 125 and lung cancer in individuals who smoked.

Age, family history of lung cancer, personal history of cancer, and smoking history all were found to have an association with lung cancer, which aligns with research studies that include these variables into lung cancer risk prediction models.[92] Despite COPD not being statistically significant, it was included in the model a priori, as it is a known risk factor for lung cancer and was statistically significant in univariate analyses. In this study, when CA 125 was removed from the multivariable logistic regression model predicting the outcome of lung cancer, COPD became statistically significant. Once CA 125 was added back into the model, COPD became statistically non-significant. The study results conclude that CA 125 has a stronger relationship to lung cancer than COPD.

Although the study objectives of previous literature are different compared to this study, the general findings are consistent. Molina and colleagues demonstrated that high CA 125 levels were found in adenocarcinomas and large cell lung cancer and could potentially be considered for lung

cancer risk prediction.[93] Wu and researchers also concluded that CA 125 had an association with lung cancer. Wang and colleagues demonstrated that CA 125 was more useful in the diagnosis of NSCLC compared to the biomarker, carcinoembryonic antigen (CEA).[94] Moreover, Ying and colleagues reported that when comparing CA 125 and CEA, only CA 125 was found to be an independent predictive marker for prognosis in patients with NSCLC.[95] [96] This study adds further evidence for an association between CA 125 and lung cancer.

5.1.1 Smoking Status, CA 125 and Lung Cancer

The relationship between smoking and lung cancer has been extensively investigated in previous research.[33] This study evaluated whether the association between CA 125 and lung cancer differed between individuals who never smoked and individuals who smoked (currently or former). The study found that an elevated CA 125 level was associated with a higher risk of lung cancer in individuals who smoked. Currently, there is no literature for us to compare these findings to.

The association between CA 125 and lung cancer in individuals who never smoked was not statistically significant in this study. Despite this association not being statistically significant, the direction of effect, magnitude of effect and consistency among the results suggest that there could be an association. Due to the current study not having enough statistical power to test this association, future research should test the association between CA 125 and lung cancer in individuals who never smoked using a larger study sample of individuals who never smoked and are diagnosed with lung cancer. Currently, there is no literature for us to compare to this study's findings.

5.1.2 Early-stage Lung Cancer and CA 125

An important study finding was that CA 125 may be useful for early detection of lung cancer. Although the study shows promising results, CA 125 did not have a large effect on this study's lung cancer prediction models. The AUC for the model including CA 125 was only 0.009 higher than when CA 125 was excluded from the model. Which indicated that CA 125 is not a strong enough predictor to be used solely in lung cancer screening. In the univariate lung cancer prediction model with CA 125, when the probability for positivity was set to P≥0.0135, the sensitivity was 78.39% (95% CI: 0.74-0.82) and the specificity was 23.87% (95% CI: 0.23-0.24). These findings were compared to another promising lung cancer biomarker, pro-surfactant protein B.[97] When it was solely examined in a univariate model, pro-surfactant protein B, which when probability for positivity was set to P≥0.032, had a sensitivity of 80.4%, and specificity of 40.1%.[97] Similarly, to the findings of CA 125 and its association with early stage lung cancer, pro surfactant protein B was found to be associated with early stage lung cancer. [97] Since CA 125 is not a strong enough biomarker to be used alone to predict lung cancer risk, future research should explore the combination of CA 125 and pro-surfactant protein B or other high performing biomarkers in regards to early stage lung cancer prediction.

5.1.3 Sex Differences among CA 125

This study's sample only included females, as the original use of CA 125 was for evaluating its use for ovarian cancer screening. When we tested for the association between lung cancer~CA 125, there were no sex-specific variables (such as ever used female hormones, history of breast disease or age at menopause) that were statistically significant. Therefore, this study's results suggest that female sex-specific factors do not have an influence on the relationship between lung cancer~CA 125. For studies that looked at lung cancer~CA 125 that included males and females, there were no results

suggesting any differences in the association of lung cancer~CA 125 between male and female participants.[79,93,97,98]

5.1.4 Exploratory Analysis

This study reported exploratory results between CA 125 and potential covariates. The results suggested that age, smoking intensity, COPD and any personal history of cancer increase CA 125 levels. In contrast, this study's results found that BMI was found to decrease levels of CA 125. Currently, there are only studies exploring the association between CA 125 and COPD. Bulut and colleagues suggested that elevated levels of biomarkers such as CA 125 in participants with COPD may be related to the severity of COPD.[99] In Li et al.'s study on the correlation of CA 125 with pleural effusions and COPD-related complications, the authors suggested that CA 125 levels were correlated with pulmonary heart disease, acute exacerbations and pulmonary hypertension.[100] Moreover, in a study by Fortun and colleagues on the use of CA 125 to distinguish pulmonary tuberculosis from other pulmonary infections, the study determined that CA 125 levels increased in patients with pulmonary tuberculosis and declined to normal values during treatment.[101] Therefore, this study's findings add to previous literature exploring the relationship between CA 125 and COPD. In part, it appears that CA 125 lies in the causal pathway between COPD and lung cancer as the removal of CA 125 from our study model led to a stronger association for COPD with lung cancer.

5.2 Impact on Public Health

Since we concluded that CA 125 is not a strong enough predictor to be used alone in lung cancer screening, the next step would be exploring the potential of combining CA 125 with a panel of other promising biomarkers that have demonstrated an association to lung cancer. In previous

literature, CA 125 has most often been examined in combination with other promising biomarkers such as carcinoembryonic antigen (CEA), cytokeratin 19 fragment (CYFRA21-1) and neuron-specific enolase (NSE).[95,96,102] In Wu et al.'s meta-analysis on the combined detection of CEA and CA 125 for the diagnosis of lung cancer, researchers found that under ideal study circumstances, the combination of CEA and CA 125 had a higher diagnostic efficiency for the detection of lung cancer than CEA detection alone. In future research, it would be beneficial to reproduce this study on a larger scale to further examine the impact of using various combinations of biomarkers (such as a study with CEA and CA 125 as mentioned earlier) dependent on age, sex and/or underlying lung conditions when screening for lung cancer. If a panel of biomarkers was found to be favorable in lung cancer risk prediction, external validations in various populations with a longitudinal study design would need to be conducted. If the usage of CA 125 in a biomarker panel for lung cancer risk prediction was externally validated, it might be useful to integrate this panel into a comprehensive lung cancer risk prediction model and test it in large samples. The inclusion of the panel of biomarkers in the model must demonstrate better risk prediction, than the original model for it to be considered for implementation for lung cancer screening.

5.3 Limitations

There were several limitations to this study. First, males were not measured for CA 125 levels, therefore the study findings are not applicable to the male population, see 5.1.3 for further discussion. Second, the questionnaire responses were self-reported by the participants, which may have caused misreporting bias. This type of bias could have happened at random and could have caused an underestimation of the study results.

The PLCO cancer screening trial sample used in this study was not representative of the general population in the US. The participants of the PLCO cancer screening trial were found to be a part of a higher socioeconomic status than the general population in the US and Canada.[103] This sample had better health outcomes and lower general mortality than the general population of similar demographics.[103] However, this study primarily focused on the biological aspects of CA 125 and its association with lung cancer which is expected to be present in all socioeconomic status levels. Therefore, sampling may not have been an important issue in the present study.

This study's statistical analysis was not optimal as more sophisticated methods are available, see "5.4 Future Directions" for further discussion. We did not assess the association between CA 125 and lung cancer by lung cancer histological type as the dataset had a large number of adenocarcinomas, but a small sample of squamous cell carcinomas. Small sample sizes of different tumours would have produced large confidence intervals causing us to be unable to draw a conclusion.

5.4 Strengths

This study has several strengths. The sample size and number of outcome events were adequate to find statistically significant results regarding the relationship between CA 125 and lung cancer. It also provided effect estimates with precise confidence intervals. Moreover, the prospective longitudinal study design allowed for the clarity of temporal sequence, avoided selection bias at enrollment through participant randomization and collected detailed information on multiple potential confounders.

5.5 Future Direction for CA 125 Research

Future research should explore time to diagnosis analysis to see whether CA 125 allows for detection of lung cancer in a time-window when the cancer is at an early stage and treatment probability of cure is high. Multi-level analysis or Cox proportional-hazards modelling is recommended for future research. These types of analyses provide a look into the biological nature of change of CA 125 and could provide a better trajectory as to how often a CA 125 test needs to be done when monitoring for lung cancer risk. Diagnosis of lung cancer as early as possible is vital as it improves the chance of early intervention and treatment. If identified early, NSCLC has the possibility of surgical resection and 5-year survival rates of 70-90%.[104-106] Unfortunately, lung cancer symptoms usually occur in patients when the lung cancer has advanced, and treatments are less effective against lung cancers that have spread. Approximately 75% of unscreened patients have advanced stage lung cancer at the time of symptomatic diagnosis.[107] Therefore, it is imperative for the continuous development and improvement of accurate biomarkers in lung cancer risk prediction for the detection of early stage lung cancer, along with screening individuals who never smoked who would not originally qualify for screening according to current eligibility criteria.

5.5 Conclusion

In conclusion, this study demonstrates that CA 125 is significantly and independently associated with lung cancer and that CA 125 is associated with early-stage lung cancer. CA 125 is not a strong enough independent predictor of lung cancer; however, it may be useful in a panel of complimentary biomarkers. Future research is needed to explore whether a panel of complimentary biomarkers including CA 125 may be a valuable addition to existing lung cancer risk prediction models and be useful in guiding selection of individuals into lung cancer screening programs.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021.
2. Centers for Disease Control and Prevention (CDC); Leading causes of death in the United States [Internet]; 2021. [updated 2021; cited 2021 Feb 7]. Available from: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm.
3. de Groot PM, Wu CC, Carter BW, Munden RF. The epidemiology of lung cancer. Transl Lung Cancer Res. 2018; 7(3):220-233.
4. Brenner DR, Weir HK, Demers AA, et al. Projected estimates of cancer in Canada in 2020. CMAJ. 2020; 192(9): E199-E205.
5. Canadian Cancer Society. Canadian Cancer Statistics: A 2020 special report on cancer incidence by stage [Internet]; 2020 [updated 2021; cited 2021 Feb 7]. Available from https://www.cancer.ca/~/media/cancer.ca/CW/cancer%20information/cancer%20101/Canadian%20cancer%20statistics/Canadian-cancer-statistics-2020_special-report_EN.pdf?la=en
6. Statistics Canada. Canadian Tobacco, Alcohol and Drugs Survey (CTADS): summary of results for 2017 [Internet]. Ottawa, ON: Health Canada; 2017 [updated 2017; cited 2021 Feb 7]. Available from https://www.canada.ca/en/health-canada/services/canadian-tobacco-alcohol-drugs-survey/2017-summary.html
7. Chyou PH, Nomura AM, Stemmermann GN. A prospective study of the attributable risk of cancer due to cigarette smoking. Am J Public Health. 1992; 82(1):37-40.

8. Boer R, Moolgavkar SH, Levy DT. Chapter 15: Impact of tobacco control on lung cancer mortality in the United States over the period 1975-2000--summary and limitations. Risk Anal. 2012; 32 Suppl 1: S190-201.

9. Couraud S, Zalcman G, Milleron B, Morin F, Souquet PJ. Lung cancer in never smokers--a review. Eur J Cancer. 2012; 48(9):1299-1311.

10. Mounawar M, Mukeria A, Le Calvez F, et al. Patterns of EGFR, HER2, TP53, and KRAS mutations of p14arf expression in non-small cell lung cancers in relation to smoking history. Cancer Res. 2007; 67(12):5667-5672.

11. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. Nat Rev Cancer. 2007; 7(10):778-790.

12. Streppel MM, Vincent A, Mukherjee R, et al. Mucin 16 (cancer antigen 125) expression in human tissues and cell lines and correlation with clinical outcome in adenocarcinomas of the pancreas, esophagus, stomach, and colon. Hum Pathol. 2012; 43(10):1755-1763.

13. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010; 5(6):463-466.

14. Lu KH, Skates S, Hernandez MA, et al. A 2-stage ovarian cancer screening strategy using the Risk of Ovarian Cancer Algorithm (ROCA) identifies early-stage incident cancers and demonstrates high positive predictive value. Cancer. 2013; 119(19):3454-3461.

15. Jacobs IJ, Menon U, Ryan A, et al. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. Lancet. 2016; 387(10022):945-956.

16. Skates SJ. Ovarian cancer screening: development of the risk of ovarian cancer algorithm (ROCA) and ROCA screening trials. Int J Gynecol Cancer. 2012; 22 Suppl 1: S24-26.

17. Buys SS, Partridge E, Black A, et al. Effect of screening on ovarian cancer mortality: the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Randomized Controlled Trial. JAMA. 2011; 305(22):2295-2303.

18. Diez M, Cardan FJ, Ortega MD, Torres A, Picardo A, Balibrea JL. Evaluation of serum CA 125 as a tumor marker in non-small cell lung cancer. Cancer. 1991; 67(1):150-154.

19. Diez M, Torres A, Maestro ML, et al. Prediction of survival and recurrence by serum and cytosolic levels of CEA, CA125 and SCC antigens in resectable non-small-cell lung cancer. Br J Cancer. 1996; 73(10):1248-1254.

20. Kimura Y, Fujii T, Hamamoto K, Miyagawa N, Kataoka M, Iio A. Serum CA125 level is a good prognostic indicator in lung cancer. Br J Cancer. 1990; 62(4):676-678.

21. Molina JR, Yang, P., Cassivi, S.D., Schild, S.E., & Adjei, A.A. Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment and Survivorship. Mayo Clin Proc. 2008; 85(5):584–594.

22. Salgia R, Harpole D, Herndon JE, 2nd, Pisick E, Elias A, Skarin AT. Role of serum tumor markers CA 125 and CEA in non-small cell lung cancer. Anticancer Res. 2001; 21(2B):1241-1246.

23. Lakshmanan I, Salfity S, Seshacharyulu P, et al. MUC16 Regulates TSPYL5 for Lung Cancer Cell Growth and Chemoresistance by Suppressing p53. Clin Cancer Res. 2017; 23(14):3906-3917

24. Gube M, Taeger D, Weber DG, et al. Performance of biomarkers SMRP, CA125, and CYFRA 21-1 as potential tumor markers for malignant mesothelioma and lung cancer in a cohort of workers formerly exposed to asbestos. Arch Toxicol. 2011; 85(3):185-192.

25. Hassan R, Schweizer C, Lu KF, et al. Inhibition of mesothelin-CA-125 interaction in patients with mesothelioma by the anti-mesothelin monoclonal antibody MORAb-009: Implications for cancer therapy. Lung Cancer. 2010; 68(3):455-459.

26. Kanwal M, Ding XJ, Song X, Zhou GB, Cao Y. MUC16 overexpression induced by gene mutations promotes lung cancer cell growth and invasion. Oncotarget. 2018; 9(15):12226-12239.

27. Gohagan JK, Prorok PC, Hayes RB, Kramer BS, Prostate LC, Ovarian Cancer Screening Trial Project T. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. Control Clin Trials. 2000; 21(6 Suppl):251S-272S.

28. Prorok PC, Andriole, G.L., Bresalier, R.S., Buys, S.S., Chia, D., Crawford, E.D., Fogel, R., Gelmann, E.P., Gilbert, F., Hasson, M.A., Hayes, R.B., Johnson, C.C., Mandel, J.S., Oberman, A., O'Brien, B., Oken, M.M., Rafla, S., Rutt, W., Weissfeld, J.L., Yokochi, L., & Gohagan, J.K... Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Control Clin Trials. 2000; 21.

29. World Health Organization. Cancer: Key Facts; 2021 [updated 2021; cited 2021 Feb 18] Available from https://www.who.int/news-room/fact-sheets/detail/cancer

30. International Agency for Research on Cancer (IARC). Tobacco Smoking; 2018 [updated 2018; cited December 16, 2019]. Available from https://monographs.iarc.fr/wpcontent/uploads/2018/06/mono100E-6.pdf

31. Lubuzo B, Ginindza T, Hlongwana K.The barriers to initiating lung cancer care in low and middle-income countries. Pan Afr Med J. 2020; 35:38.

32. American Cancer Society. What is cancer? [Internet]; 2019 [updated 2019; cited 2021 Feb 18]. Available from http://www.cancer.org/cancer/lung-cancer/about/what-is.html

33. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. Clin Chest Med. 2011; 32(4):605-644.

34. Carney DN, De Leij L. Lung cancer biology. Semin Oncol. 1988; 15(3):199-214.

35. National Cancer Institute. Understanding Cancer: What is cancer? [Internet]; 2015 [updated 2015; cited 2021 Feb 18]. Available from https://www.cancer.gov/about-cancer/understanding/what-is-cancer

36. El-Telbany A, Ma PC. Cancer genes in lung cancer: racial disparities: are there any? Genes Cancer. 2012; 3(7-8):467-480.

37. Collins LG, Haines, C., Perkel, R., & Enck, R. Lung Cancer: Diagnosis and Management. American Family Physician. 2007; 75(1).

38. van Meerbeeck JP, Fennell DA, De Ruysscher DK. Small-cell lung cancer. Lancet. 2011; 378(9804):1741-1755.

39. Travis WD, Brambilla, E., Muller-Hermelink, H.K., Haris, C.C., World Health Organization. Organization classification of tumours: pathology and genetics: tumours of the lung, pleura, thymus and heart. IARC Press. 2004; 10.

40. College of American Pathologists. Lung Cancer: Squamous Cell Carcinoma [Internet]; 2019 [updated 2019; cited 2021 Jan 10]. Available from http://www.cap.org/apps/docs/reference/myBiopsy/LungSquamousCellCarcinoma.pdf

41. College of American Pathologists. Lung Cancer: Lung Adenocarcinoma [Internet]; 2019 [updated 2019; cited 2021 Jan 10]. Available from http://www.cap.org/apps/docs/reference/myBiopsy/LungAdenocarcinoma.pdf

42. American Joint Commission on Cancer. (2010). Cancer Staging Manual. (S. B. Edge, D. R. Byrd, C. C. Compton, A. G. Fritz, F. L. Greene, & A. I. Trotti, Eds.) (7th Ed.). Chicago, IL: Springer.

43. Mirsadraee S, Oswal D, Alizadeh Y, Caulo A, van Beek E, Jr. The 7th lung cancer TNM classification and staging system: Review of the changes and implications. World J Radiol. 2012; 4(4):128-134.

44. Wynder EL, Muscat JE. The changing epidemiology of smoking and lung cancer histology. Environ Health Perspect. 1995; 103 Suppl 8:143-148.

45. Alberg AJ, Brock MV, Ford JG, Samet JM, Spivack SD. Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2013; 143(5 Suppl): e1S-e29S.

46. Brownson RC, Chang, J.C., and Davis, J.R. Gender and Histologic Type Variations in Smoking-Related Risk of Lung Cancer. Epidemiology. 1992; 3(1).

47. Creamer MR, Wang TW, Babb S, et al. Tobacco Product Use and Cessation Indicators Among Adults - United States, 2018. MMWR Morb Mortal Wkly Rep. 2019; 68(45):1013-1019.

48. Bohadana A, Nilsson F, Rasmussen T, Martinet Y. Gender differences in quit rates following smoking cessation with combination nicotine therapy: influence of baseline smoking behavior. Nicotine Tob Res. 2003; 5(1):111-116.

49. Sorensen G, Pechacek TF. Attitudes toward smoking cessation among men and women. J Behav Med. 1987; 10(2):129-137.

50. Keohavong P, DeMichele MA, Melacrinos AC, Landreneau RJ, Weyant RJ, Siegfried JM. Detection of K-ras mutations in lung carcinomas: relationship to prognosis. Clin Cancer Res. 1996; 2(2):411-418.

51. McCarthy WJ, Meza R, Jeon J, Moolgavkar SH. Chapter 6: Lung cancer in never smokers: epidemiology and risk prediction models. Risk Anal. 2012; 32 Suppl 1: S69-84.

52. Patel JD. Lung cancer in women. J Clin Oncol. 2005; 23(14):3212-3218.

53. Subbaraman N. Public health: A burning issue. Nature. 2014; 513(7517): S16-17.

54. Doll R, & Bradford Hill, A. Smoking and Carcinoma of the Lung. British Medical Journal. 1950.

55. Alberg AJ, Brock MV, Samet JM. Epidemiology of lung cancer: looking to the future. J Clin Oncol. 2005; 23(14):3175-3185.

56. Siegel R, Ward E, Brawley O, Jemal A. Cancer statistics, 2011: the impact of eliminating socioeconomic and racial disparities on premature cancer deaths. CA Cancer J Clin. 2011; 61(4):212-236

57. Ryberg D, Hewer A, Phillips DH, Haugen A. Different susceptibility to smoking-induced DNA damage among male and female lung cancer patients. Cancer Res. 1994; 54(22):5801-5803.

58. Slatore CG, Chien JW, Au DH, Satia JA, White E. Lung cancer and hormone replacement therapy: association in the vitamins and lifestyle study. J Clin Oncol. 2010; 28(9):1540-1546.

59. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018; 68(1):7-30.

60. Torre LA, Siegel RL, Ward EM, Jemal A. Global Cancer Incidence and Mortality Rates and Trends--An Update. Cancer Epidemiol Biomarkers Prev. 2016; 25(1):16-27.

61. Dalton SO, Frederiksen BL, Jacobsen E, et al. Socioeconomic position, stage of lung cancer and time between referral and diagnosis in Denmark, 2001-2008. Br J Cancer. 2011; 105(7):1042-1048.

62. Gao Y, Goldstein AM, Consonni D, et al. Family history of cancer and nonmalignant lung diseases as risk factors for lung cancer. Int J Cancer. 2009; 125(1):146-152.

63. Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. Br J Cancer. 2005; 93(7):825-833.

64. Lissowska J, Foretova L, Dabek J, et al. Family history and lung cancer risk: international multicentre case-control study in Eastern and Central Europe and meta-analyses. Cancer Causes Control. 2010; 21(7):1091-1104.

65. Brenner DR, Hung RJ, Tsao MS, et al. Lung cancer risk in never-smokers: a population-based case-control study of epidemiologic risk factors. BMC Cancer. 2010; 10:285.

66. Bast Jr RC, Badgwell, D., Lu, Z., Marquez, R., Rosen, D., Liu, J., Baggerly, K.A., Atkinson, E.N., Skates, S., Zhang, Z., Lokshin, A., Menon, U., Jacobs, I., & Lu, K. New tumor markers: CA125 and beyond. Int J Gynecol Cancer. 2005; 15:274-281.

67. Miralles C, Orea M, Espana P, et al. Cancer Antigen 125 Associated With Multiple Benign and Malignant Pathologies. Annals of Surgical Oncology. 2003; 10(2):150-154.

68. Moss EL, Hollingworth J, Reynolds TM. The role of CA125 in clinical practice. J Clin Pathol. 2005; 58(3):308-312.

69. Terada KY, Elia J, Kim R, Carney M, Ahn HJ. Abnormal CA-125 levels in menopausal women without ovarian cancer. Gynecol Oncol. 2014; 135(1):34-37.

70. Einhorn N, Sjovall K, Knapp RC, et al. Prospective evaluation of serum CA 125 levels for early detection of ovarian cancer. Obstet Gynecol. 1992; 80(1):14-18.

71. Jacobs I, Davies AP, Bridges J, et al. Prevalence screening for ovarian cancer in postmenopausal women by CA 125 measurement and ultrasonography. BMJ. 1993; 306(6884):1030-1034.

72. Buys SS, Partridge E, Greene MH, et al. Ovarian cancer screening in the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial: findings from the initial screen of a randomized trial. Am J Obstet Gynecol. 2005; 193(5):1630-1639.

73. U.S. Preventive Services Task Force (USPSTF). (2018). Screening for ovarian cancer: recommendation statement. JAMA, 319(6).

74. Canadian Task Force on Preventive Health Care. (2013). Screening for Ovarian: U.S. Preventive Services Task Force Reaffirmation Recommendation Statement [Internet]; 2013 [updated 2013; cited 2021 Feb 17]. Available from https://canadiantaskforce.ca/wp-content/uploads/2016/05/2013-ovarian-cancer-en.pdf

75. Yin BW, Lloyd KO. Molecular cloning of the CA125 ovarian cancer antigen: identification as a new mucin, MUC16. J Biol Chem. 2001; 276(29):27371-27375.

76. Rachagani S, Torres MP, Moniaux N, Batra SK. Current status of mucins in the diagnosis and therapy of cancer. Biofactors. 2009; 35(6):509-527.

77. Gubbels JA, Felder M, Horibata S, et al. MUC16 provides immune protection by inhibiting synapse formation between NK and ovarian tumor cells. Mol Cancer. 2010; 9:11.

78. Boivin M, Lane D, Piche A, Rancourt C. CA125 (MUC16) tumor antigen selectively modulates the sensitivity of ovarian cancer cells to genotoxic drug-induced apoptosis. Gynecol Oncol. 2009; 115(3):407-413.

79. Ma S, Shen L, Qian N, Chen K. The prognostic values of CA125, CA19.9, NSE, AND SCC for stage I NSCLC are limited. Cancer Biomark. 2011; 10(3-4):155-162.

80. Gohagan JK, Prorok, P.C., Hayes, R.B., Kramer, B.S. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: History, Organization, and Status. Contemporary Clinical Trials. 2000; 21(6):251S-272S.

81. Simpson NK, Johnson CC, Ogden SL, et al. Recruitment strategies in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial: the first six years. Control Clin Trials. 2000; 21(6 Suppl):356S-378S.

82. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Control Clin Trials. 2000; 21(6 Suppl):273S-309S.

83. StataCorp. 2018. Stata Statistical Software: Release 15. College Station, TX: StataCorp LP

84. Field, A. P. (2009). Discovering statistics using SPSS (and sex, drugs and rock “n” roll). Los Angeles: SAGE Publications, 2009.

85. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. Int J Epidemiol. 1999; 28(5):964-974.

86. Greenland S. Interactions in epidemiology: relevance, identification, and estimation. Epidemiology. 2009; 20(1):14-17.

87. Archer KJL, S.; Goodness-of-fit test for logistic regression model fitted using survey sample data. The Stata Journal. 2006; 6(1).

88. McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics (Edited by P.Zarembka), 105-42. Academic Press, New York.

89. Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory (pp. 267-281). Budapest: Akademiai Kiado.

90. Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics 6(2), pp. 461-464.

91. Hosmer DWL, S... Applied Logistic Regression (Second Edi.). John Wiley & Sons, Inc. 2000.

92. Tammemagi MC, Ten Haaf K, Toumazis I, et al. Development and Validation of a Multivariable Lung Cancer Risk Prediction Model That Includes Low-Dose Computed Tomography Screening Results: A Secondary Analysis of Data From the National Lung Screening Trial. JAMA Netw Open. 2019; 2(3): e190204.

93. Molina R, Filella X, Auge JM, et al. Tumor markers (CEA, CA 125, CYFRA 21-1, SCC and NSE) in patients with non-small cell lung cancer as an aid in histological diagnosis and prognosis. Comparison with the main clinical and pathological prognostic factors. Tumour Biol. 2003; 24(4):209-218.

94. Wang H, Zhu Z, Xiao Y, Ma N, Li H, Wen Z. [The value of serum tumor marker CA125 and CEA in the diagnosis of non-small cell lung cancer.]. Zhongguo Fei Ai Za Zhi. 2008; 11(1):97-100.

95. Ying L, Wu J, Zhang D, et al. Preoperative serum CA125 is an independent predictor for prognosis in operable patients with non-small cell lung cancer. Neoplasma. 2015; 62(4):602-609.

96. Wu LX, Li XF, Chen HF, et al. Combined detection of CEA and CA125 for the diagnosis for lung cancer: A meta-analysis. Cell Mol Biol (Noisy-le-grand). 2018; 64(15):67-70.

97. Sin DD, Tammemagi CM, Lam S, et al. Pro-surfactant protein B as a biomarker for lung cancer prediction. J Clin Oncol. 2013; 31(36):4536-4543.

98. Dai H, Liu J, Liang L, et al. Increased lung cancer risk in patients with interstitial lung disease and elevated CEA and CA125 serum tumour markers. Respirology. 2014; 19(5):707-713.

99. Bulut I, Arbak P, Coskun A, et al. Comparison of serum CA 19.9, CA 125 and CEA levels with severity of chronic obstructive pulmonary disease. Med Princ Pract. 2009; 18(4):289-293.

100. Li S, Ma H, Gan L, et al. Cancer antigen-125 levels correlate with pleural effusions and COPD-related complications in people living at high altitude. Medicine (Baltimore). 2018; 97(46): e12993.

101. Fortun J, Martin-Davila P, Mendez R, et al. Ca-125: a useful marker to distinguish pulmonary tuberculosis from other pulmonary infections. Open Respir Med J. 2009; 3:123-127.

102. Yang Q, Zhang P, Wu R, Lu K, Zhou H. Identifying the Best Marker Combination in CEA, CA125, CY211, NSE, and SCC for Lung Cancer Screening by Combining ROC Curve and Logistic Regression Analyses: Is It Feasible? Dis Markers. 2018; 2018:2082840.

103. Pinsky PF, Miller A, Kramer BS, et al. Evidence of a healthy volunteer effect in the prostate, lung, colorectal, and ovarian cancer screening trial. Am J Epidemiol. 2007; 165(8):874-881.

104. Shah R, Sabanathan S, Richardson J, Mearns AJ, Goulden C. Results of surgical treatment of stage I and II lung cancer. J Cardiovasc Surg (Torino). 1996; 37(2):169-172.

105. Nesbitt JC, Putnam JB, Jr., Walsh GL, Roth JA, Mountain CF. Survival in early-stage non-small cell lung cancer. Ann Thorac Surg. 1995; 60(2):466-472.

106. Goldstraw P, Chansky K, Crowley J, et al. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol. 2016; 11(1):39-51.

107. Walters S, Maringe C, Coleman MP, et al. Lung cancer survival and stage at diagnosis in Australia, Canada, Denmark, Norway, Sweden and the UK: a population-based study, 2004-2007. Thorax. 2013; 68(6):551-564.

Supplementary Tables

Appendix Table of Content

Table S1.	Multivariable logistic regression for lung cancer and CA 125 results by screening round T0	89
Table S2.	Multivariable logistic regression for lung cancer and CA 125 results by screening round T1	89
Table S3.	Multivariable logistic regression for lung cancer and CA 125 results by screening round T2	90
Table S4.	Multivariable logistic regression for lung cancer and CA 125 results by screening round T3	90
Table S5.	Multivariable logistic regression for lung cancer and CA 125 results by screening round T4	91
Table S6.	Multivariable logistic regression for lung cancer and CA 125 results by screening round T5	91
Table S7.	Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T0	92
Table S8.	Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T1	92
Table S9.	Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T2	93
Table S10.	Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T3	93
Table S11.	Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T4	94
Table S12.	Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T5	94

Table 1S. Multivariable logistic regression for lung cancer and CA 125 results by screening round T0

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
CA 125 Results, positive vs negative	1.75 (1.01-3.02; 0.047)
Age (years)	1.05 (1.03-1.06; <0.001)
Education, self-reported level of,	0.94 (0.89-1.00; 0.051)
Current Body Mass Index (BMI) (kg/m^2)	0.96 (0.95-0.98; <0.001)
Lung cancer, self-reported family history of, yes vs no	1.09 (1.04- 1.15; <0.001)
Any type of cancer, self-reported personal, yes vs no	1.46 (1.09- 1.95; 0.011)
COPD, self-reported history of, yes vs no	1.22 (0.95-1.58; 0.120)
How many cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	27,893
P of overall model	<0.001
Pseudo R^2	0.1654

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 2S. Multivariable logistic regression for lung cancer and CA 125 results by screening round T1

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
CA 125 Results, positive vs negative	2.23 (1.32-3.75; 0.003)
Age (years)	1.05 (1.03-1.06; <0.001)
Education, self-reported level of,	0.94 (0.89-1.01; 0.081)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-0.99; 0.001)
Lung cancer, self-reported family history of, yes vs no	1.09 (1.03-1.14; 0.001)
Any type of cancer, self-reported personal, yes vs no	1.53 (1.13-2.07; 0.006)
COPD, self-reported history of, yes vs no	1.17 (0.89-1.55; 0.264)
How many cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	26,554
P of overall model	<0.001
Pseudo R^2	0.1614

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 3S. Multivariable Logistic Regression for lung cancer and CA 125 results by Screening Round T2

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
CA 125 Results, positive vs negative	1.42 (0.79-2.54; 0.240)
Age (years)	1.05 (1.03-1.07; <0.001)
Education, self-reported level of,	0.94 (0.88-1.00; 0.057)
Current Body Mass Index (BMI) (kg/m^2)	0.98 (0.96-1.00; 0.033)
Lung cancer, self-reported family history of, yes vs no	1.10 (1.04-1.15; <0.001)
Any type of cancer, self-reported personal, yes vs no	1.54 (1.12-2.12; 0.008)
COPD, self-reported history of, yes vs no	1.25 (0.94-1.66; 0.127)
How many cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	25,608
P of overall model	<0.001
Pseudo R^2	0.1628

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 4S. Multivariable logistic regression for lung cancer and CA 125 results by screening round T3

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
CA 125 Results, positive vs negative	2.20 (1.30-3.72; 0.003)
Age (years)	1.05 (1.03-1.07; <0.001)
Education, self-reported level of,	0.93 (0.87-1.00; 0.047)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-0.99; 0.011)
Lung cancer, self-reported family history of, yes vs no	1.10 (1.04-1.16; 0.001)
Any type of cancer, self-reported personal, yes vs no	1.60 (1.15-2.21; 0.005)
COPD, self-reported history of, yes vs no	1.05 (0.76-1.44; 0.781)
How many cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.06; <0.001)
n total	24,471
P of overall model	<0.001
Pseudo R^2	0.1544

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 5S. Multivariable logistic regression for lung cancer and CA 125 results by screening round T4

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
CA 125 Results, positive vs negative	2.56 (1.42-4.61; 0.002)
Age (years)	1.04 (1.02-1.07; <0.001)
Education, self-reported level of,	0.91 (0.84-1.00; 0.040)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-1.00; 0.023)
Lung cancer, self-reported family history of, yes vs no	1.07 (1.00-1.15; 0.065)
Any type of cancer, self-reported personal, yes vs no	1.65 (1.12-2.45; 0.012)
COPD, self-reported history of, yes vs no	1.24 (0.85-1.80; 0.263)
How many cigarettes smoked per day	1.02 (1.01-1.03; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	19,445
P of overall model	<0.001
Pseudo R^2	0.1556

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 6S. Multivariable logistic regression for lung cancer and CA 125 results by screening round T5

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
CA 125 Results, positive vs negative	1.56 (0.82-2.96; 0.171)
Age (years)	1.05 (1.02-1.07; <0.001)
Education, self-reported level of,	0.93 (0.86-1.00; 0.059)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-1.00; 0.026)
Lung cancer, self-reported family history of, yes vs no	1.09 (1.02-1.16; 0.007)
Any type of cancer, self-reported personal, yes vs no	1.06 (1.11-2.32; 0.012)
COPD, self-reported history of, yes vs no	1.18 (0.83-1.69; 0.355)
How many cigarettes smoked per day	1.02 (1.01-1.03; <0.001)
Number of years smoking	1.06 (1.05-1.06; <0.001)
n total	21,406
P of overall model	<0.001
Pseudo R^2	0.4444

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 7S. Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T0

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
LN CA 125 Levels	1.14 (0.97-1.35; 0.113)
Age (years)	1.05 (1.03-1.06; <0.001)
Education, self-reported level of,	0.94 (0.89-1.00; 0.051)
Current Body Mass Index (BMI) (kg/m^2)	0.96 (0.95-0.98; <0.001)
Lung cancer, self-reported family history of, yes vs no	1.09 (1.04-1.15; <0.001)
Any type of cancer, self-reported personal, yes vs no	1.45 (1.09-1.94; 0.012)
COPD, self-reported history of, yes vs no	1.22 (0.94-1.57; 0.132)
How many cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	27,892
P of overall model	<0.001
Pseudo R^2	0.1652

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 8S. Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T1

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
LN CA 125 Levels	1.25 (1.05-1.49; 0.012)
Age (years)	1.04 (1.03-1.06; <0.001)
Education, self-reported level of,	0.94 (0.89-1.01; 0.075)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-0.99; 0.001)
Lung cancer, self-reported family history of, yes vs no	1.09 (1.03-1.14; 0.001)
Any type of cancer, self-reported personal, yes vs no	1.54 (1.14-2.08; 0.005)
COPD, self-reported history of, yes vs no	1.16 (0.88-1.53; 0.298)
How many cigarettes smoked per day	1.01 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	26,554
P of overall model	<0.001
Pseudo R^2	0.1610

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 9S. Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T2

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
LN CA 125 Levels	1.19 (0.99-1.44; 0.068)
Age (years)	1.04 (1.03-1.06; <0.001)
Education, self-reported level of,	0.94 (0.88-1.00; 0.055)
Current Body Mass Index (BMI) (kg/m^2)	0.98 (0.96-1.00; 0.037)
Lung cancer, self-reported family history of, yes vs no	1.09 (1.04-1.15; <0.001)
Any type of cancer, self-reported personal, yes vs no	1.54 (1.12-2.12; 0.008)
COPD, self-reported history of, yes vs no	1.23 (0.93-1.64; 0.150)
How many cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	25,608
P of overall model	<0.001
Pseudo R^2	0.1633

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 10S. Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T3

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
LN CA 125 Levels	1.37 (1.13-1.67; <0.001)
Age (years)	1.05 (1.03-1.07; <0.001)
Education, self-reported level of,	0.93 (0.87-0.99; 0.042)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-1.00; 0.012)
Lung cancer, self-reported family history of, yes vs no	1.10 (1.04-1.16; <0.001)
Any type of cancer, self-reported personal, yes vs no	1.58 (1.14-2.19; 0.006)
COPD, self-reported history of, yes vs no	1.03 (0.75-1.43; 0.836)
How many cigarettes smoked per day	1.02 (1.01-1.02; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	24,471
P of overall model	<0.001
Pseudo R^2	0.1551

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 11S. Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T4

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
LN CA 125 Levels	1.40 (1.11-1.77; 0.005)
Age (years)	1.04 (1.02-1.07; 0.001)
Education, self-reported level of,	0.92 (0.84-1.00; 0.046)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-1.00; 0.029)
Lung cancer, self-reported family history of, yes vs no	1.07 (1.00-1.15; 0.051)
Any type of cancer, self-reported personal, yes vs no	1.65 (1.11-2.44; 0.013)
COPD, self-reported history of, yes vs no	1.23 (0.84-1.78; 0.287)
How many cigarettes smoked per day	1.02 (1.01-1.03; <0.001)
Number of years smoking	1.06 (1.05-1.07; <0.001)
n total	19,445
P of overall model	<0.001
Pseudo R^2	0.1554

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease

Table 12S. Multivariable logistic regression for lung cancer and log transformed CA 125 levels by screening round T5

Explanatory Variables	**Odds Ratio (95% Confidence Interval; *P-value*)**
LN CA 125 Levels	1.40 (1.13-1.75; 0.003)
Age (years)	1.04 (1.02-1.07; <0.001)
Education, self-reported level of,	0.93 (0.86-1.00; 0.06)
Current Body Mass Index (BMI) (kg/m^2)	0.97 (0.95-1.00; 0.04)
Lung cancer, self-reported family history of, yes vs no	1.09 (1.03-1.16; 0.006)
Any type of cancer, self-reported personal, yes vs no	1.60 (1.11-2.31; 0.012)
COPD, self-reported history of, yes vs no	1.16 (0.82-1.66; 0.402)
How many cigarettes smoked per day	1.02 (1.01-1.03; <0.001)
Number of years smoking	1.06 (1.05-1.06; <0.001)
n total	21,406
P of overall model	<0.001
Pseudo R^2	0.1467

Abbreviation: COPD: Chronic Obstructive Pulmonary Disease