Evaluation of the PLCO_{M2012} Risk Prediction Model and National Lung Screening Trial Criteria for Selecting Individuals for Lung Cancer Screening

Meera Mahmud, BSc (Honors)

Submitted in partial fulfilment 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

© 2019
ABSTRACT

Background: Lung Cancer (LC) is the leading cause of cancer death in North America. Cancer screening trials, such as the PLCO (Prostate, Lung, Colorectal and Ovarian) and NLST (National Lung screening trial) evaluate LC mortality. There has been a growing interest in risk prediction modelling for selecting high-risk individuals for LC screening. Increased risk for developing LC is however associated with greater risk of dying from non-LC (competing) causes; the presence of coexisting illnesses may negate the net benefit of LC screening.

Purpose: The focus of this study was to compare the two selection criteria methods, the NLST criteria and the PLCO\textsubscript{M2012} model (at 1.5\% and 2\% 6-year risk), for selecting smokers age of 55 to 75 for screening who may develop LC, and to evaluate which risk factors are strongly associated with competing causes of death (CCoD).

Methods: Contingency table and logistic regression analysis, using STATA software, were used to analyze the results of applying both criteria on the PLCO ever-smokers population (N= 74 207). A CCoD logistic regression model was developed using 5-year follow-up data to assess the association between each PLCO\textsubscript{M2012} model predictor and non-LC death. Predictor variables were ranked by their ability to predict 6-year LC incidence and 5-year non-LC death. Predictive performance, discrimination (area under the receiving-operating-characteristic curve [ROC-AUC]) and calibration were assessed.

Results: Significantly higher LC proportions were found using the PLCO\textsubscript{M2012} model than the NLST criteria in various individual characteristics. Increasing the model threshold resulted in less false positives, higher positive predictive value and probability of 6-year LC incidence. The PLCO\textsubscript{M2012} model was also shown to be significantly associated with 5-year non-LC death (p<0.001). The CCoD model demonstrated fair discrimination (AUC=0.7114) and good
calibration. There was an overall agreement in the rank order of variables predicting LC incidence and non-LC death, with the top four variables being age, smoking intensity and duration, and body mass index (BMI), however the effects were opposite in nature for the latter.

**Conclusion:** Findings show that the PLCO\textsubscript{M2012} risk prediction model accurately identifies higher proportions of LC cases compared to the NLST criteria, however it also selects individuals for LC screening who are at risk of dying from competing causes who will not benefit from screening. Decisions on LC screening among individuals at high-risk of developing LC should be based on an appropriate assessment of each person’s health status and life expectancy, in order to maximize the benefit-harm ratio from LC screening.

**Keywords:** Lung cancer, epidemiology, risk prediction modelling, lung cancer screening, competing causes of death
ACKNOWLEDGEMENTS

These past three years could not have been possible without the support from my supervisor, Dr. Martin Tammemagi. I am grateful to have had such an inspiring role model to guide me throughout this challenging, but rewarding, process. The knowledge you have shared with me throughout my graduate studies has shaped me into the student I am today. You have instilled in me a strong passion for Epidemiology and I am excited for the journey ahead, where I can utilize the knowledge and skills you have taught me.

I would like to thank my committee members, Dr. Dawn Prentice and Dr. Sinead McElhone, for taking time out of their busy schedules to look over my thesis and provide me with their feedback. Both of you have made me look at my research from different perspectives which has expanded my knowledge and helped me grow as a researcher. I would also like to extend my gratitude to the wonderful professors in the Faculty of Health Sciences, such as Dr. Pietrangeli, who has always kept her office door and her ears open for me whenever I needed to talk.

I would also like to mention the wonderful friends and family who have showed me support and have edited my work throughout these years. All the wonderful people I met at Brock gave me a reason to come on campus every day. I am thankful for the friends that have taken the time to read over and provide feedback on my thesis work, I am forever in your debt. Most importantly, I would like to thank my family for staying patient with me throughout this process and providing me unconditional support and love. All of this could not have been possible without my mother. She has been my rock throughout this entire process, and I know that me finishing my program finally gives her peace.
# TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

Chapter 1: Introduction ........................................................................................................................................... 1

1.1 Research Gaps.................................................................................................................................................. 6

1.2 Purpose.......................................................................................................................................................... 6

1.3 Study Aims.................................................................................................................................................... 6

1.4 Study Objective 1........................................................................................................................................... 7

1.5 Study Objective 2........................................................................................................................................... 7

1.6 Study Objective 3........................................................................................................................................... 8

Chapter 2: Background ........................................................................................................................................ 9

2.1 Biology of Lung Cancer................................................................................................................................... 9

   2.1.1 Description and Pathology .................................................................................................................... 9

   2.1.2 Types of Lung Cancer .......................................................................................................................... 10

2.2 Lung Cancer Epidemiology .......................................................................................................................... 10

   2.2.1 Trends in Lung Cancer ........................................................................................................................ 10

   2.2.2 Mortality and Survival ........................................................................................................................ 12

   2.2.3 Diagnosis, Treatment and Prognosis................................................................................................. 12

2.3 Risk Factors .................................................................................................................................................. 13
4.2 Study Population Characteristics .............................................................. 60

4.2.1 Characteristics of PLCO smokers associated with a Lung Cancer Diagnosis .... 60

4.2.2 Impact of Participant Characteristics on Lung Cancer Risk in Univariate Analysis ......................................................................................................................... 61

4.3 Study Objective 1: NLST versus PLCO_{M2012} Model Risk \( \geq 1.5\% \) \& \( \geq 2.0\% /6yr \) Criteria for Selecting LC Screenees .............................................................................................................................. 66

4.4 Study Objective 2: PLCO_{M2012}+/NLST criteria– vs. NLST criteria+/PLCO_{M2012} – for selecting LC Screenees by PLCO_{M2012} Model Predictors ........................................................................................................... 69

4.4.1 Differences in 6-year Lung Cancer Incidence Risk in Discordant groups by Risk Predictor Status ........................................................................................................................................................... 71

4.4.2 In the High-Risk Strata of each Model Risk Predictor: What are the Odds of 1) Being selected for Screening and 2) Being Diagnosed with Lung Cancer by the PLCO_{M2012} Model vs. NLST Criteria? ................................................................................................................................. 72

4.5 Study Objective 3: Univariate & Multivariable Models for 5-year non-LC deaths .... 78

4.5.1 Evaluation of PLCO_{M2012} Model Predictors in Univariate and Multivariable Analysis .............................................................................................................................................. 83

4.5.1.1 Age .............................................................................................................. 83

4.5.1.2 Race/Ethnicity .............................................................................................. 83

4.5.1.3 Education ..................................................................................................... 84

4.5.1.4 Body Mass Index .......................................................................................... 85

4.5.1.5 Personal History of Cancer .......................................................................... 87

4.5.1.6 Family History of Lung Cancer ..................................................................... 87

4.5.1.7 History of Chronic Obstructive Pulmonary Disease ....................................... 88

4.5.1.8 Smoking Status ............................................................................................ 88
4.5.1.9 Smoking Intensity .................................................................................. 89

4.5.1.10 Smoking Duration ................................................................................ 90

4.5.1.11 Smoking Quit-time ............................................................................... 91

4.5.2 Competing Causes of Death Model Performance ........................................ 93

4.5.3 Competing Causes of Death Model Assumption & Diagnostics ..................... 96

Chapter 5: Discussion .......................................................................................... 97

5.1 PLCO_{M2012} Model Superior in Selecting High-Risk Individuals for LC Screening .... 97

5.1.1 The PLCO_{M2012} Model at $\geq 1.5\%/6$yr Risk Threshold .......................... 97

5.1.2 The PLCO_{M2012} Model at $\geq 2.0\%/6$yr Risk Threshold .......................... 98

5.2 Components of the PLCO_{M2012} model select individuals with highest 6-year LC risk.. 99

5.3 Predicting Competing Causes of Death .......................................................... 101

5.3.1 Age ........................................................................................................... 101

5.3.2 Smoking Intensity and Duration ............................................................... 103

5.3.3 Body Mass Index. ....................................................................................... 104

5.3.4 CCoD negatively impacts Lung Cancer Screening Outcomes. ....................... 105

5.4 Study strengths ............................................................................................. 106

5.5 Study limitations ........................................................................................... 107

5.6 Implications .................................................................................................. 109

5.7 Conclusion .................................................................................................... 116

References ......................................................................................................... 118
Table 1. Projected new cases for the most common cancers, by age group and sex in Canada, 2017.

Table 2. Characteristics of the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207) and univariate logistic regression odds ratios predicting lung cancer.

Table 3a. Proportion of participants eligible, 6-year lung cancer risk and lung cancer cases detected by NLST criteria and PLCO\textsubscript{M2012} model risk $\geq 1.5\%/6$yr criterion status in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207).

Table 3b. Proportion of participants eligible, 6-year lung cancer risk and lung cancer cases detected by NLST criteria and PLCO\textsubscript{M2012} model risk $\geq 2.0\%/6$yr criterion status in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207).

Table 4. Accuracy of lung cancer classification by each criterion in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=72 407).

Table 5a. Distribution of Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207) in discordant groups by PLCO\textsubscript{M2012} risk factors at $\geq 1.5\%/6$yr PLCO\textsubscript{M2012} model risk.

Table 5b. Distribution of Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207) in discordant groups by PLCO\textsubscript{M2012} risk factors at $\geq 2.0\%/6$yr PLCO\textsubscript{M2012} model risk.

Table 6. Univariate and multivariable logistic regression models for 6-year lung cancer incidence and 5-year risk of dying from competing causes in Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207) using PLCO\textsubscript{M2012} model predictors.

Table 7. PLCO\textsubscript{M2012} model predictors ranked, from greatest to least (1-10) predictive ability, by decrease in AUC values in multivariable lung cancer incidence and competing causes (non-lung cancer) death models.

Table 8. Predictive performance of multivariable competing causes of death model in Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207).
LIST OF FIGURES

Figure 1. Non-linear relationship between education (6 levels) and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207), adjusted for other predictors in the PLCO\textsubscript{M2012} risk prediction model.

Figure 2. Non-linear relationship between body mass index and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207), adjusted for other predictors in the PLCO\textsubscript{M2012} risk prediction model.

Figure 3. Non-linear relationship between smoking intensity and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207), adjusted for other predictors in the PLCO\textsubscript{M2012} risk prediction model.

Figure 4. The unadjusted non-linear relationship between smoking duration and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207).

Figure 5. The univariate non-linear relationship between smoking quit-time and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207).

Figure 6. Calibration plot: Observed vs. predicted 5-year non-lung cancer death probabilities.

Figure 7. Distribution of competing causes deaths and no competing causes death in years by 5-year non-LC death model risk (scatter diagram).

Figure 8. LOWESS plot of 5-year non-lung cancer deaths vs. PLCO\textsubscript{M2012} lung cancer risk in Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial smokers (N=74 207).
LIST OF ABBREVIATIONS

ACRIN – American College of Radiology Imaging Network
AUC – Area under the curve
BMI – Body mass index
CARET – Beta-Carotene and Retinol Efficacy Trial
CCO – Cancer Care Ontario
CCoD – Competing Causes of Death
CDC – Disease Control and Prevention
CHD – Coronary heart disease
CI – Confidence intervals
CLPTMIL – Cleft lip and palate transmembrane 1 like gene
COPD – Chronic obstructive pulmonary disorder
CT – Computed tomography
CVD – Cardiovascular Disease
CXR – Chest radiography
EFGR – Epidermal growth factor receptor
GED – General education development
HL – Hosmer-Lemeshow
IMS – Information Management System
IQR – Interquartile range
LC – Lung cancer
LDCT – Low-dose helical computed tomography
LLP – Liverpool Lung Project
LOWESS – Locally weighted scatterplot smoothing
LSS – Lung Screening Study
Lung-RADS – Lung Imaging Reporting and Data System
MFP – Multivariable fractional polynomial
NCI – National Cancer Institute
NDI – National Death Index
NIH – National Institute of Health
NLST – National Lung Screening Trial
NNS – Number needed to screen
NRI – Net reclassification improvement
NSCLC – Non-small cell lung cancer
OR – Odds ratio PAH – Polycyclic aromatic hydrocarbon
Pan-Can – Pan-Canadian Early Detection of Lung Cancer Study
PLCO – Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial
PMRT – Post-mastectomy radiotherapy
PPV – Positive predictive value
QALY – Quality-adjusted life years.
RCT – Randomized controlled trial
ROC – Receiver operator characteristic curve
SCLC – Small cell lung cancer
SES – Socioeconomic status
SD – Standard deviation
SHS – Second-hand smoke
SMS – Study Management System
SC – Squamous cell carcinoma

TERT – Human telomerase reverse transcriptase gene

USPSTF – Unites States Prevention Service Task Force

VIF – Variation inflation factor
1. **Introduction**

Lung cancer (LC) is the leading cause of cancer related death in North America.\(^{(1,2)}\) However, due to the decrease in smoking behavior, which is the main cause of LC, incidence rates of LC have declined in North America since the late 20\(^{th}\) century. Several studies have shown the inverse relationship between LC and smoking cessation.\(^{(3)}\) In the United States and Canada, the prevalence of male smokers started to decrease in the mid-1960s, and to date, males are still considered to have higher smoking rates, on average, than females.\(^{(4,5)}\) Continuous reduction in tobacco use has had a positive impact on LC mortality rates in North America, however smoking rates remain high and millions of current and former-smokers remain at great risk for developing LC.\(^{(5,6)}\)

It is important to consider why LC deaths are so high despite the decrease in smoking prevalence. Research suggests that while smoking cessation reduces the risk of LC, former-smokers remain at an elevated risk in relation to never-smokers.\(^{(7,8)}\) Another reason behind the high mortality rate associated with LC is the poor prognosis of the disease, due to the fact that LC is most often diagnosed at an advanced stage when effective treatment is no longer helpful.\(^{(9)}\) Due to this poor prognosis, fewer people with LC are alive beyond two years after diagnosis.\(^{(5)}\) This suggests that early diagnosis of LC can lead to early management and treatment and in turn can decrease the risk of dying from LC.

The search for an ideal screening test to detect early stage LC goes back to the 1960s. Several screening trials across Europe and the United States demonstrated that early detection of LC using screening tests showed improvement in survival, due to more indolent cancers being identified, however did not lead to a mortality reduction.\(^{(10)}\) The use of sputum cytology and chest radiography (CXR) for screening failed to identify a significant decrease in LC mortality.
One of the biggest cancer screening trials known as the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening trial also demonstrated no significant reduction in LC mortality from annual CXR. \(^{15,16}\)

The National Lung Screening Trial (NLST) is the largest randomized LC screening study to date which demonstrated that screening high-risk individuals with low-dose helical computed tomography (LDCT) significantly reduces LC mortality by 20\%. \(^{17}\) LDCT provides images with excellent resolution to detect small nodules in areas of the lung and has been shown to be more effective in detecting nodules and early-stage LCs than CXR. \(^{17–21}\) Due to the significant results from the NLST trial, using LDCT to screen high-risk individuals has been recommended by many organizations in the United States. \(^{22–26}\) Two important organizations which issued national coverage for LC screening were the United States Prevention Task Force (USPSTF) and Centers for Medicare and Medicaid. Details on LC screening trials will be discussed in upcoming chapters.

The NLST results sparked an interest in the benefits of LDCT screening on LC mortality, but more importantly prompted the important question of who should be screened and what qualifies individuals to be considered at high-risk for LC. The NLST trial identified high-risk individuals using three criteria; the first being individuals between the ages of 55-74 years, the second being those with a history of cigarette smoking of at least 30 pack years (packs smoked per day multiplied by the years spent smoking), and lastly, former-smokers who have quit within the last 15 years. \(^{17}\) Older age is an important risk factor for LC and is a prominent characteristic in those diagnosed with LC; studies show that LC mortality rates are higher in older cohorts made up of heavy smokers as compared to younger cohorts composed of heavy smokers. \(^{27}\) It has also been found that shortened survival after LC diagnosis is associated with
older age.\textsuperscript{(28)} Since smoking is the most significant risk factor for LC in not only current, but also former-smokers, smoking history (including smoking cessation) is also important factor to study when evaluating LC risk.\textsuperscript{(3,8)}

There are risk factors other than age and smoking history that are strongly associated with LC which are not included in the NLST criteria. For instance, an individual’s medical history can put them at great risk for LC; this includes family history of LC, personal history of cancer or underlying lung conditions such as pulmonary diseases.\textsuperscript{(29–32)} Studies have also shown an association between low body mass index (BMI) or socioeconomic status (SES) and increased LC risk, and this relationship was observed in many countries including the U.S, Canada, the Netherlands and China, even after adjusting for smoking.\textsuperscript{(33–37)} Disparities in LC incidence between different ethnicities also exist, where African American men are at greatest risk.\textsuperscript{(38–40)} Therefore, these risk factors along with smoking characteristics such as smoking duration, intensity and number of years since smoking cessation, are important indicators for LC. Risk factors for LC will be discussed in more detail in the following chapter. Recognizing important predictor variables for LC increases the chance of selecting high-risk individuals for screening, and as a result might lead to more early-stage LCs diagnosed and treated, therefore averting more deaths.

Developing prediction algorithms to identify high-risk individuals has gained importance in the medical field. Risk prediction models estimate the probability of a certain health incidence during a given period.\textsuperscript{(41–43)} Early risk prediction models have shown promising results in selecting high-risk individuals for LC screening. However, many of these models have limitations such as a narrow number of valuable predictors, inefficient measurements of certain predictors and overall poor or modest calibration and discrimination when validated in different
populations, suggesting that these models may not be successful in improving screening programs.\(^{(41,44–48)}\)

The PLCO Cancer Screening Trial 2012 Model (PLCOM\(_{2012}\)) is a logistic regression model based on the 6-year LC incidence among PLCO control smokers. This model predicts the probability of LC incidence within 6 years for ever-smokers (current or former-smokers) to determine their eligibility for LC screening. Studies have shown that using the PLCOM\(_{2012}\) risk prediction model (at ≥1.5% six years risk threshold) to select individuals for screening is significantly more sensitive in detecting those who will be diagnosed with LC and could therefore save more lives than using the NLST criteria.\(^{(49,50)}\) A risk threshold value of 1.5%/6yr was first chosen by observing the number of LC deaths in the NLST intervention (LDCT) group and examining at which risk the NLST LDCT arm mortality rates were consistently below the CXR control arm’s rates. This value was at the PLCOM\(_{2012}\) model 65\(^{th}\) percentile risk which is equivalent to a risk value of 0.015 in PLCO smokers.\(^{(50)}\) The PLCOM\(_{2012}\) model contains variables which do not require collection of biospecimens for laboratory testing, making it feasible.\(^{(51)}\) The PLCOM\(_{2012}\) models’ variables, along with its advantages over the NLST criteria, will be discussed in the next chapter.

Other risk thresholds of the PLCOM\(_{2012}\) model have been evaluated; the Pan-Canadian Early Detection of Lung Cancer (Pan-Can) Study used a modified version of the PLCOM\(_{2012}\) model with a risk threshold value of ≥2%/6yr in different centers across Canada, and this model identified a greater proportion of early stage LCs than the NLST criteria.\(^{(52)}\) Recently, Cancer Care Ontario (CCO) began its Lung Cancer Screening Pilot for People at High Risk, in which individuals in four screening centers across Ontario were recruited for LC screening based on the PLCOM\(_{2012}\) model using a risk threshold of ≥2%/6yr. The primary purpose of the pilot program
is to determine how to best implement organized provincial LC screening, and initial interim results indicate that the program is successful at identifying a relatively large number of LCs in which a high proportion are early stage. Significant results from this pilot could translate into LC screening becoming available throughout Ontario as part of public healthcare using the PLCO_{M2012} model as the eligibility criterion.\(^{53}\).

As previously mentioned, to reduce the number of deaths from LC though a screening program, screening should be targeted towards high-risk individuals. Unfortunately, individuals at risk for developing LC are also at risk of dying from competing causes of death (CCoD). Significant competing risks exist for heavy smokers such as cardiovascular disease, respiratory diseases, hypertension and other types of cancers.\(^{54-56}\) Individuals with co-existing illnesses are at risk of dying from these causes and may be unable to withstand diagnostic or treatment procedures. Not to mention, risks of screening among this population are magnified, such as overdiagnosis, increased emotional stress and side effects from LDCT screening or downstream evaluations (i.e. biopsies). Therefore, to increase the efficacy of a selection criterion, such as the PLCO_{M2012} risk prediction model risk, individuals who are at high-risk of dying from competing causes should not be selected for screening as these individuals are not likely to benefit from a screening program.\(^{55-57}\) The effectiveness of a screening program are measured by several factors which include LC specific mortality, life-years gained, and quality of life-years gained (QALY), all of which are negatively impacted when screening individuals at high-risk for CCoD; this will be further discussed in the upcoming chapters. Overall, individual comorbid situations should be considered when predicting LC risk and recommending screening for high-risk individuals.
1.1 Research Gaps

Gaps in literature exist regarding the performance of the PLCOM2012 risk prediction model at a risk threshold of ≥2%/6yr. This threshold is relevant and novel as it has been recently evaluated in Canada. Utilizing a stricter risk threshold for model-based screening selection, as compared to a lesser risk threshold, translates into less false positive results and selecting those for screening who are at higher risk of developing LC. To the best of our knowledge, the PLCOM2012 model at ≥2%/6yr risk has not been evaluated against the NLST criteria among ever-smokers. The predictive performance of the model at this threshold therefore needs to be evaluated, and this was addressed in our study. Gaps in knowledge also remain as to whether individuals with co-existing illnesses, who are at high-risk for developing LC as well as dying from non-LC causes, should be offered LC screening. The impact of CCoD on LC screening outcomes is a developing area of research and the association between risk prediction modelling and CCoD is yet to be investigated.

1.2 Purpose:

The purpose of this study was to identify components of the PLCOM2012 risk prediction model that were superior to the NLST criteria in selecting high-risk individuals for LC screening, and might also have been deleterious if they were strongly associated with non-LC death.

1.3 Study Aims

The specific aims of this study were:

1) To evaluate the overall performance of the PLCOM2012 model at 1.5% and 2% 6-year-risk thresholds, specifically comparing it to the NLST criteria.

2) To validate previously identified PLCOM2012 model predictors by assessing which components of the model were increasing the selection of high-risk individuals for LC screening.
3) To investigate which components of the PLCO_{M2012} model were increasing the selection of individuals at high-risk for non-LC death, who would not benefit from LC screening.

1.4 Study Objective 1

The first objective of this study was to examine whether the PLCO_{M2012} model, at 1.5% and 2% six-year LC risk threshold, identified high-risk individuals for LC incidence better than the NLST criteria. The NLST criteria and the PLCO_{M2012} model (with both risk thresholds) were applied to a subset of ever-smokers from the PLCO trial (N=74 207). By applying both criteria, PLCO smokers were eligible for screening by either the model or the NLST criteria. The number of participants eligible for screening, total number of LC cases diagnosed, and 6-year LC incidence risk was estimated for those eligible by either criterion and the accuracy of each selection method was compared by calculating sensitivities, specificities and positive predictive values (PPV).

1.5 Study Objective 2

The second objective was to investigate if components (predictors) of the PLCO_{M2012} model selected high-risk individuals who were truly at higher risk for developing LC, and whether the PLCO_{M2012} model detected a greater proportion LC cases than the NLST criteria among individuals in the high-risk strata for each predictor.

When both selection criteria were applied to PLCO smokers, as done in the first study objective, some participants were eligible for screening by one method and not the other, and these discordant groups are informative. The second objective directly compared these discordant groups – individuals eligible for screening by the NLST criteria but not the PLCO_{M2012} model and individuals eligible for screening by the PLCO_{M2012} model but not the NLST criteria.
For each discordant group, PLCO\textsubscript{M2012} model predictors were evaluated in their ability to identify high-risk individuals; this was done by computing the number of participants eligible for screening, number of LC cases detected, and 6-year LC incidence risk for the high-risk strata for each predictor variable. This will demonstrate which PLCO\textsubscript{M2012} predictors are important in improving the identification of individuals who develop LC by the PLCO\textsubscript{M2012} criterion versus the NLST criteria.

1.6 Study Objective 3

The last objective of this study was to assess if any PLCO\textsubscript{M2012} model predictors, predicted the risk of dying from non-LC (competing) causes more so, less than or to the same extent as the risk of developing LC, specifically observing which components of the PLCO\textsubscript{M2012} model were increasing the selection of individuals that might be dying within 5-year times of competing causes. The reasoning behind choosing a time-frame of 5 years will be discussed in the Methodology chapter. The association and predictive performance between each individual model predictor and 5-year non-LC death was assessed. The PLCO\textsubscript{M2012} model overall was also evaluated in its ability to predict 5-year non-LC death. This study objective is important because if strong LC incidence predictors are weakly associated with CCoD, then they can be weighted more heavily in prediction models, whereas weak LC incidence predictors which are strongly associated with CCoD can be removed or discounted in prediction models.
2. **Background**

This chapter provides the fundamental background knowledge of LC, LC screening and risk prediction modelling. Examining the previous literature in this area will help the reader appreciate the rationale of this research study. To start, the biological nature and the different types of LC will be discussed, followed by descriptive statistics on LC in the general population. Next, an overview of important known and suspected risk factors for LC will be presented to shed light on potential predictors. A review of LC screening and the PLCO and NLST screening trials will also be reviewed to help the reader understand the research goals and the study population. A description of risk prediction modeling and the PLCO\textsubscript{M2012} model will be provided to explain the efficiency of using LC prediction models for screening, as well to provide a basis for the Methodology section. Lastly, an analysis of the different competing risks of deaths associated with LC will be presented. This will help the reader to better understand the importance of the last study aim and the influence of competing risks of death on the selection of high-risk individuals for LC.

2.1 **Biology of Lung Cancer**

2.1.1 **Description and Pathology**

LC is a malignant tumor that originates in the cells of the lung. Various genetic, epigenetic and protein expression mutations lead to the activation of oncogenes (genes that form tumor cells) and inactivation of tumor suppressor genes (promote programmed cell death and regulate the cell cycle). A common tumor growth promoting protein associated with LC is Epidermal Growth Factor Receptor (EFGR). EFGR is involved in the regulation of oncogenic functions such as cell proliferation, survival, differentiation, invasion and metastasis. Activating mutations in EFGR leads to abnormal growth of lung epithelial cells which is associated with
LC. Tumor suppressor genes, such as tumor protein 53 (TP53), play a crucial role in regulating normal cell growth and gene expression, therefore loss of function of these genes is significantly related to carcinogenesis (cancer formation).\(^{58}\)

### 2.1.2 Types of Lung Cancer

LC can be divided into two major types that differ in their biological behavior, non-small cell LC (NSCLC) and small cell LC (SCLC). NSCLC is the most common type of LC, accounting for 80%-85% of LCs. NSCLC can be further subdivided into several types. Adenocarcinoma is the most common type of NSCLC which originates in mucus producing cells located in the epithelial lining of the lung. This type of cancer is most typically found in never-smokers, women and younger populations. Squamous cell carcinoma is a type of NSCLC originating in the bronchi and is more common in current and former-smokers. Other and least common forms of NSCLC include large cell carcinoma, adenosquamous carcinoma and sarcomatoid carcinoma.\(^{1,59}\)

SCLC usually starts in the central region of the lungs and is the most aggressive form of LC. The main types of SCLC are small cell carcinoma and combined small cell carcinoma, and these are most prevalent in smokers. SCLC are small cells but they grow and spread very quickly to different parts of the body and because of this, the survival of individuals with SCLC is usually short.\(^{59}\)

### 2.2 Lung Cancer Epidemiology

#### 2.2.1 Trends in Lung Cancer

Observing the trend of LC incidence worldwide, and specifically in North America, will help understand the overall burden this disease causes today. During the late 19\(^{th}\) century there was an increase in the mass production of tobacco products, mainly cigarettes, which initiated
the tobacco epidemic. Not long after, the LC epidemic began in the 1930s and LC incidence rates were on the rise. In 2002, highest LC incidence rates in men were in Central and Eastern Europe and in North America, and for women, were highest in North America and Northern Europe. Few decades after the start of the tobacco epidemic, the intensity of cigarette smoking also increased while the age of smoking initiation decreased; overtime this led LC to be the second leading and leading cause of cancer death in females and males, respectively, globally.\(^{(60)}\)

In North America, the prevalence of cigarette smoking peaked earlier in men (~1950s) than in women (~1960s), so the incidence of LC in women started later.\(^{(60)}\) In America, more men are diagnosed with LC each year, however, LC incidence rates in males has decreased over the past 40 years, while it has risen for women, due to the increase in prevalence of cigarette smoking among women. LC rates peaked for men in the early 1980s (102.1 per 100 000) and began to decline at a faster rate than for women. LC incidence rates in women peaked in the late 1990s (52.9 per 100 000) and recently started to decrease; the gender gap in LC incidence is expected to close.\(^{(61)}\)

In the beginning of the 20\(^{th}\) century, LC incidence and smoking rates were greater in those with higher SES living in high income countries compared to those with lower SES, however during the 1950s and early 1960s, smoking prevalence shifted to lower SES groups living in low-middle income countries. This was due to increased knowledge of the harmful effects of smoking on health which caused higher SES individuals to quit smoking.\(^{(61)}\) Today, LC rates are continuing to decline; in 2017 there were 222 500 new LC cases in the U.S \(^{(62)}\) and 28 600 new LC cases in Canada.\(^{(5)}\)
To understand the nature of the problem, statistics on LC mortality and survival are necessary. Even though LC rates have been declining in North America, survival rates are very low, and more people die from LC than any other type of cancer. In the U.S., LC mortality rates decreased 43% from ~85-90 deaths per 100,000 in 1990 to ~50 deaths per 100,000 in 2014 among males and 17% from ~43-45 deaths per 100,000 in 2002 to ~35 deaths per 100,000 in 2014 among females due to reduced tobacco use and improvements in treatment.\(^{(62)}\) In Canada, the mortality rate of LC started to level off in the late 1980s and has been declining since (2.1% per year since 1992). However, the mortality rate for females was rising until recently, the mortality rate decreased 0.8% per year between 2006 and 2012. Overall, males have been estimated to continue to have a higher LC mortality rate (59 per 100,000) than females (45 per 100,000) in 2017.\(^{(5)}\)

The 5-year survival rate for LC in the U.S during 2001 to 2007 was 16.3%, compared to 12.3% in the late 1970s.\(^{(63)}\) Based on 2006-2008 Canadian statistics, the 5-year survival rates for lung, colorectal and prostate cancer are 17%, 64% and 95% respectively.\(^{(64)}\) There have not been large advances in survival for LC over the past several decades and this is because survival rates vary depending on the stage of diagnosis; diagnosis made at the local stage is associated with a 52% 5-year survival rate, 24% survival rate at the regional stage and 4% at the distant stage.\(^{(63)}\)

### 2.2.3 Diagnosis, Treatment and Prognosis

Symptoms often do not appear in the early stages for LC. When they do appear however, symptoms may include chest pain, that worsens by coughing or deep breathing, blood stained mucus, frequent chest infections, wheezing and atelectasis. Individuals with more subtle symptoms such as consistent coughing, shortness of breath, fatigue, or loss of appetite generally
mistake these symptoms for other health conditions and often do not see their physicians, therefore decreasing chances of early diagnosis and treatment.\(^{(65)}\)

If it is suspected that an individual may have LC, due to suspicious or unusual results from physical exams, blood tests or screenings, then diagnosis is the next step. Diagnostic procedures may include sputum cytology, bronchoscopy and biopsy. Sputum cytology involves examining the mucus coughed up by the lungs (sputum) because large squamous cell tumors may shed cancer cells into the sputum. Bronchoscopy is used to look inside large airways of the lung to look for tumor blockage and allows biopsies of tumors and lymph nodes for diagnosis.\(^{(66)}\)

Treatment of LC depends on the stage of the cancer, patients’ overall health and their personal preference. Early stage disease may be surgically resected with curative intent. NSCLC can be treated with different forms of surgery depending on where the cancer is located. Radiation therapy, chemotherapy or chemoradiation (combination of radiation therapy and chemotherapy) are possible treatment methods for advanced disease.\(^{(67)}\) 75% of patients with LC present with symptoms due to progressive local, or regional or metastatic disease and this is when treatment is no longer effective. This results in a poor prognosis of individuals diagnosed with LC.\(^{(5)}\) Hence, LC screening is important because it can lead to early detection of localized stage cancer which improves survival and long-term quality of life. LC screening programs have been executed with this objective and their techniques and results will be discussed shortly.

### 2.3 Risk Factors

#### 2.3.1 Tobacco Smoking

Smoking accounts for 80-90% of LC cases and compared to never-smokers, current smokers are at a 20-fold increased risk for LC.\(^{(68)}\) A large pooled analysis evaluating the impact of smoking on the risk of developing major cell types of LC found higher risk estimates in
squamous cell carcinoma (SqCC) and SCLC.\(^{(69)}\) All 4 major types of LC (adenocarcinoma, squamous cell carcinoma, large-cell carcinoma and small-cell carcinoma) are caused by smoking and the risk for LC increases with the time spent smoking (smoking duration) and the number of cigarettes smoked per day (smoking intensity).\(^{(63)}\) A 20-year prospective study on British Doctors, examining the association between tobacco smoking and cause-specific mortality, found that a stronger effect of smoking duration was demonstrated than of smoking intensity. It was estimated that a dose response relationship exists between smoking intensity and LC risk, however smoking duration was shown to have an exponential effect on the risk of LC, demonstrating that those who start smoking at a younger age have a higher chance of developing LC.\(^{(70)}\) However, it is important to mention that using risk prediction modeling, smoking intensity demonstrates a non-linear relationship with LC risk and has a stronger association with LC risk than smoking duration.\(^{(49)}\)

The risk of death during the first and second 20 years of the 40 years’ follow-up in Doll and Peto’s study were 264 per 100,000 and 314 per 100,000, respectively. This demonstrated that LC mortality increased among smokers in the second 20 years,\(^{(71)}\) while another analysis of the same cohort found that the relative risk of death caused from LC in heavy current-smokers (≥25 cigarettes per day) was 1.83 times the relative risk for never-smokers in the cohort born in the 19th century.\(^{(72)}\) These results demonstrated that longer duration of smoking and greater smoking intensity increases the risk of developing LC.

Smokers can benefit from smoking cessation at any age, regardless of sex and type of tobacco smoked. The Lung Health Study, a prospective randomized trial, demonstrated the efficacy of smoking cessation interventions on decreasing LC deaths.\(^{(73)}\) The U.S. Surgeons General’s 1990 report on the health benefits of smoking cessation show that as the period of
abstinence from smoking cigarettes increases, the risk of acquiring LC decreases. However, even for durations of abstinence of 40 years, the risk of LC among former-smokers remains higher than for never-smokers.\(^8\)

2.3.2 Age

Older age is associated with greater risk of developing LC. In 2009, the American Cancer Society recorded that LC causes the most cancer deaths in those over 60 years of age.\(^{74}\) One study revealed that the individual LC mortality rate in an older, heavy smoking cohort born in 1910–1914 is 250 per 100 000 at the age of 70, but for the cohort born in 1940–1944, this risk is 150 per 100 000, representing a decrease of 40\%.\(^{27}\) A similar trend is observed in Canada; in those over the age of 50, lung, breast, prostate and colorectal cancer accounts for over 47\% of all new cancer cases, in which LC is most common for people aged 60-80 years.\(^{5,74}\) In both men and women, a drastic increase in LC death from smoking is seen after ages 50-60 years (Table 1).\(^{3,5}\)

<table>
<thead>
<tr>
<th>Age</th>
<th>Lung and bronchus</th>
<th>Colorectal</th>
<th>Prostate</th>
<th>Breast</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total* Males</td>
<td>Females</td>
<td>Total* Males</td>
<td>Females</td>
</tr>
<tr>
<td>0-19</td>
<td>28,600</td>
<td>14,400</td>
<td>14,200</td>
<td>26,800</td>
</tr>
<tr>
<td>20-29</td>
<td>25</td>
<td>10</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>30-39</td>
<td>100</td>
<td>35</td>
<td>60</td>
<td>350</td>
</tr>
<tr>
<td>40-49</td>
<td>610</td>
<td>270</td>
<td>340</td>
<td>1,150</td>
</tr>
<tr>
<td>50-59</td>
<td>3,600</td>
<td>1,650</td>
<td>2,050</td>
<td>3,800</td>
</tr>
<tr>
<td>60-69</td>
<td>8,500</td>
<td>4,300</td>
<td>4,200</td>
<td>8,100</td>
</tr>
<tr>
<td>70-79</td>
<td>9,600</td>
<td>5,000</td>
<td>4,600</td>
<td>9,700</td>
</tr>
<tr>
<td>80+</td>
<td>6,300</td>
<td>3,400</td>
<td>3,100</td>
<td>6,700</td>
</tr>
</tbody>
</table>

Table 1. Projected new cases for the most common cancers, by age group and sex, Canada, 2017


Some research suggests that increased inflammation and decreased immune functions due to the aging process can make mutations more liable and increase proliferation and can therefore
support the alteration of lung cells or tissues to malignancy. However, the relationship between age and overall cancer risk due to genetic mutations is not yet fully understood.\(^{(75)}\)

2.3.3 Family History of Lung Cancer

It has been shown that having a history of LC in one or more relatives increases one’s risk for LC by 2 to 4-fold in ever-smokers as well as non-smokers. LC risk increases with the number of first degree relatives, even after controlling for smoking and other risk factors.\(^{(30,76,77)}\) Familial aggregation (or occurrence) of LC may be caused by genetic, lifestyle or environmental factors, or a combination of all three. Several case control studies have supported family aggregation of LC where LC inheritance was based on genetic susceptibility.\(^{(78–82)}\) In a previous study, first-degree relatives of a LC proband, which consisted of never-smokers, had a greater probability of developing cancer compared with first-degree relatives of never-smokers with no LC, suggesting the presence of a familial effect independent of tobacco smoking. This study demonstrated that never-smokers with LC were 40\% more likely, than never-smoking controls, to report a family history of LC in their first-degree relatives, adjusting for sex, race/ethnicity, age, smoking status, occupation, industry, chronic bronchitis, emphysema, and tuberculosis.\(^{(81)}\) Studies conducted on groups of twins are beneficial for distinguishing between genetic and non-genetic (environmental or infectious) risk factors of familial aggregations of LC. Observing higher associations of LC between monozygotic (genetically identical) compared to dizygotic twins (share 50\% of their segregated genes), demonstrates the influence of genetic factors on the familial pattern of LC. One large study assessed the risk of LC in 45,000 monozygotic and dizygotic twins and found that twins of individuals with LC had an increased risk for LC in both the monozygotic group (7.7 -fold increased risk) and the dizygotic group (6.7-fold increased risk). This suggested a combined effect of genetic and environmental factors on familial
occurrence of LC in twins.\(^{(83)}\) Various twin studies have shown that this association is due to similar environmental and lifestyle (mostly smoking) factors and that approximately 8% of LCs are inherited or occur because of a genetic predisposition.\(^{(79,84,85)}\)

Genome wide association studies have found chromosomal regions 5p15, 6p21 and 15q25 to be associated with LC development. The 5p15.33 locus contains the human telomerase reverse transcriptase (TERT) gene and the cleft lip and palate transmembrane 1 like (CLPTMIL) gene.\(^{(86)}\) The TERT gene is essential for telomerase enzyme production and maintenance of telomeres. Mutations of these gene can lead to overexpression of the telomerase enzyme which can lead to formation of cancerous cells; 90% of human tumor tissues show telomerase activity, meaning that TERT plays a vital role in carcinogenesis. The CLPTMIL gene is expressed in the lung tissue and is postulated to stimulate apoptosis of lung cells under toxic conditions (for e.g., tobacco carcinogen related stress). Therefore, deregulated or disrupted pathways of apoptosis leads to proliferations and increased susceptibility to LC.\(^{(86)}\) Abnormal expression of certain genes also located on 6p21 region are associated with tumor growth and are strongly related to LC.\(^{(87)}\) Chromosomal region 15q25.1 has been shown to be correlated with nicotine dependency and LC development. Nicotinic acetylcholine receptor subunit genes found in this region have been theorized to take a part in the signaling pathways that promote carcinogenesis.\(^{(88)}\)

2.3.4 Personal History of Cancer

Personal history of cancer appears to be associated with LC in never-smokers and increases risk in smokers. In one study, previous reproductive primary cancer and a history of radiotherapy was significantly linked with LC in women, suggesting that endocrine factors possibly play a role in lung carcinogenesis in women.\(^{(32)}\) In another study, the risk for LC increased (relative risk = 2.0, 95% CI 1.0-4.3) in women who had previous irradiation experience
versus those who did not 10 years after their original breast cancer diagnosis.\(^{(89)}\) An association between post-mastectomy radiotherapy (PMRT) for breast cancer and LC risk was seen in those with a smoking history; compared with women without PMRT exposures, women who smoked cigarettes and had PMRT demonstrated a 37-fold increase of LC risk in the ipsilateral lung.\(^{(90)}\) Several reasons contribute to this association such as the interaction between radiotherapy and cigarette smoking, genetic predisposition or previous cancer treatments which make individuals more predisposed to the disease.\(^{(91,92)}\)

### 2.3.5 Pulmonary Diseases

A strong association exists between airflow obstruction (which can be caused by chronic obstructive pulmonary disorder (COPD)\(^{(29,31)}\) and LC, as well as fibrotic disorders and LC.\(^{(93)}\) In a large population-based case control study, LC risk was higher for those with a history of chronic bronchitis (odds ratio (OR) = 2.0, 95% CI 1.5–2.5), emphysema (OR = 1.9, 95% CI 1.4–2.8), or COPD (OR = 2.5, 95% CI 2.0–3.1), after adjusting for smoking. It was also found that in light smokers, the link between chronic bronchitis and LC was the strongest.\(^{(31)}\) Tobacco exposure stimulates inflammation in the lung cells potentially causing LC and COPD through an inflammation pathway known as nuclear factor-kB pathway.\(^{(29,63)}\) There are several mechanisms that explain the relationship between COPD and LC, independent of smoking. For example, a deficiency of alpha 1-antitrypsin, which is a protein that protects the body from neutrophil elastase (enzyme which can attack normal lung tissue), increases the risk of LC by almost 2-fold.\(^{(94)}\) Regardless of the mechanism, the presence of COPD is a clinically useful risk indicator for LC incidence.
2.3.6 Body Mass Index

An inverse relationship between BMI and LC exists, indicating that a higher BMI reduces the risk of LC.\(^{(37,95)}\) It has been suggested that leanness plays an etiological role in cancer via carcinogen related oxidative DNA damage; smoking-related DNA compounds in blood lymphocytes and oxidative DNA damage (determined by levels of urinary 8-hydroxy-deoxyguanosine) appear to be inversely associated with BMI.\(^{(96)}\) The inverse relationship between BMI and LC risk is more prevalent in smokers than never-smokers,\(^{(37,96,97)}\) however various studies have shown this association to exist even after adjusting for smoking status and smoking intensity.\(^{(98,99)}\) One study demonstrated that the relationship between BMI and LC risk was not explained by confounding variables, such as smoking, because similar magnitude of association was seen in women who were never-smokers. Women in the upper quintile of BMI (≥30.70 kg/m\(^2\)) were found to be at a decreased risk for all major LC subtypes, and this association was statistically significant for squamous cell carcinoma (relative risk = 0.22, 95% CI 0.08-0.64).\(^{(98)}\) The association between BMI and LC risk is difficult to study because of various confounding variables such as lifestyle factors, including smoking and alcohol drinking, which affect body weight over time.

2.3.7 Socioeconomic Status

Socioeconomic status (SES) describes the social standing of an individual or group, and is an indicator for certain environmental, occupational or behavioral exposures.\(^{(100)}\) SES is often measured as a combination of education, income and occupation. Research shows that children from low-SES communities develop academic skills at a slower rate than children from higher SES groups. This may be due to the poor cognitive development, language, memory and emotional processing seen in children from low SES communities. Schools in lower SES
communities are usually under resourced, consequently leading to increased drop-out rates and decreased academic achievement in these communities.\(^{(101)}\)

The relationship between less educated populations and LC may reflect the issue of poor school systems, which can lead to poor income and health in adulthood. Lack of knowledge on risk factors of LC can increase the probability of acquiring the disease, as less educated individuals may be less receptive to prevention messages and are less likely to change their health behavior and use the health care system.\(^{(102)}\) In North America, the risk of LC is inversely related to income and education, and lower SES is significantly linked to poorer LC survival. This correlation may reflect difficulties in accessing and attaining proper treatment on time in lower SES communities and the fact that individuals in these communities are more likely to be diagnosed with LC at a later stage due to lack of knowledge of the disease.\(^{(34,103)}\)

Studies have been conducted to assess the role of smoking in explaining the increase LC incidence rates in lower SES communities. Smoking has been shown to be associated with SES, where smoking habits and risk taking behaviors are more prevalent in lower SES communities.\(^{(104,105)}\) Some studies have shown smoking to account for about 40% of the educational differences in LC incidence,\(^{(106,107)}\) however the inverse relationship between higher education/SES status and LC risk has been consistently seen even after adjusting for smoking status, duration or intensity.\(^{(35,108,109)}\)

2.3.8 Race/Ethnicity

LC incident rates are similar among African American and White women, however are 47% higher among African American men compared to White men. These differences reflect historical differences in smoking prevalence,\(^{(110)}\) susceptibility to tobacco carcinogens, and occupational exposures.\(^{(38)}\) Racial and ethnic disparities seen in LC incidence, mortality and
survival are often also associated with SES where African American men have greater LC mortality and low survival rate due to delayed diagnosis and lack of access to LC treatment.\(^{(38,39,111–113)}\)

In the Disease Control and Prevention (CDC) report of racial disparities in LC incidence from 1998 to 2006, it was seen that the annual incidence of LC was highest among African Americans (76.1 per 100 000), followed by Whites (69.7 per 100 000), then American Indians/Alaska Natives (48.4 per 100 000), Asian/Pacific Islanders (38.4 per 100 000) and lastly, Hispanics (37.3 per 100 000).\(^{(40)}\) Observing LC incident rates from 2004 to 2008, it was found that African Americans had the highest age-adjusted LC incidence rates (72.7 per 100 000) compared to Whites (63.3 per 100 000), American Indians/Alaskan Natives (44.5 per 100 000), Asians/Pacific Islanders (39.0 per 100 000), and Hispanics (32.5 per 100 000).\(^{(110)}\) These differences initially were attributed to the variations in smoking behavior among diverse ethnic and racial groups. The decrease in LC incidence observed among African Americans greatly reflects the decrease in initiation of tobacco smoking among individuals in this population since the 1970s.\(^{(38)}\)

Tobacco exposure alone does not contribute to the higher incidence of LC among African Americans. Occupational exposure to asbestos/insulation and coal is greater in African Americans compared to their White counterparts.\(^{(114)}\) Literature on the association between genetic variation and susceptibility to LC in African Americans are inconclusive.\(^{(115)}\) African Americans continue to have a higher incidence of LC than Whites, adjusting for smoking, therefore further research is needed to attempt to delineate factors attributing to this difference.
2.3.9 Secondhand Smoking

Exposure to second-hand smoke (SHS) is defined as being exposed to someone else’s cigarette smoke at home, work, or in public and outdoor places on a regular basis. The 2006 U.S Surgeon General’s report identifies SHS as the major cause of LC and coronary heart disease in healthy nonsmokers, accounting for approximately 3 000 LC deaths per year in the U.S. Nonsmokers who live with a smoker have a 20%-30% increased risk for LC and components of SHS such as nicotine, carbon monoxide, and tobacco specific carcinogens can be found in exposed nonsmokers.\(^{116}\) Some prediction models use SHS as an important predictor for LC. Spitz et al. showed that SHS was predictive in never-smokers and had a statistically significant association with LC (OR = 1.8, 95% CI 1.2 - 2.6).\(^{47}\)

2.3.10 Diet

Evidence has shown fruit and vegetable consumption to be inversely associated with LC risk, however findings for vegetable consumption has been less consistent than fruit. Fruits and vegetables contain certain antioxidants, such as carotenoids, which help prevent oxidative stress and DNA damage induced by free radicals. An inverse relationship has been seen between carotenoids and LC.\(^{117}\) The 2007 World Cancer Research Fund reports that out of 19 effect estimates from 17 cohort studies, 14 showed reduced risk with higher level of vegetable consumption, which was statistically significant in 3. This report also states that a pooled analysis from 8 cohort studies that assessed the association between fruit intake and LC risk, showed a statistically significant reduction in LC risk in the highest intake fruit group when compared to the lowest (relative risk = 0.77, 95% CI 0.67–0.87). These data suggested a convincing correlation between fruit intake and decreased LC risk.\(^{118}\) However, the relationship between diet and LC risk remains uncertain because of the effect of smoking. Cigarette smoking
is closely linked to an unhealthy lifestyle and relationships between dietary factors and LC are much weaker than the association between active tobacco smoking and LC.\(^{(63)}\)

### 2.3.11 Occupational Exposures

LC is linked to many workplace exposures. Metals such as arsenic, chromium and nickel are known to increase LC risk. In one study, higher number of LC cases were observed in individuals exposed to occupational exposures such as paint dust, metals, polycyclic aromatic hydrocarbons (PAH), welding fumes and asbestos.\(^{(35)}\)

Asbestos is a well-known occupational carcinogen and refers to several forms of naturally appearing silicate materials. Although the exact mechanism of how asbestos exposure causes LC is unknown, the relationship between asbestos exposure and occupational LC is strong.\(^{(119,120)}\) In the LC risk prediction model by Bach et al., asbestos exposure was identified by occupation in asbestos-associated trades such as insulation, sheet metal work, plumbing, plasterboard application, ship-fitting, ship electrical work, boiler making, or ship scaling and was associated with an independent increase in LC risk (hazard ratio = 1.24, 95% CI 1.04 – 1.48; \(p = 0.02\)).\(^{(44)}\) A pooled analysis looking at 11 case control studies found a dose response trend in LC risk when looking at the effects of diesel exhaust exposure, another occupational agent containing carcinogens.\(^{(121)}\) Overall there is a weak correlation between diesel exhaust exposure and risk of developing LC.

### 2.3.12 Radon Exposure

Radon is an inert gas produced from radium in the natural decay series of uranium. Radon has several isotopic forms and radon-222 is one that is found in significant concentration in the environment. The decay products of radon produce alpha particles which can cause DNA damage of the respiratory epithelium cells and has been described as the second most important
cause of LC after cigarette smoking.\textsuperscript{(122)} Studies of underground miners of uranium and other minerals have documented radon exposure as a cause of LC and exposure from radon in indoor air is also associated with an increased risk for LC.\textsuperscript{(63,122)}

2.3.13 Ionizing Radiation

In general, a good proportion of the U.S population acquires radiation through X-radiation and computed tomography (CT). The use of CT in the U.S has increased more than 3-fold since 1993. In one study, it was estimated that 29,000 cancers were related to CT scans performed in the U.S in 2007 and out of these, LC was the most common expected radiation related cancer.\textsuperscript{(123)} Risk models show that the risk of radiation on cancer would only be an issue if people got screened by LDCT at an early age, for example before the age of 50.\textsuperscript{(124)} More research is needed to understand the association between the radiation from CT and x-rays and LC risk.

2.3.14 Arsenic

A systematic review of the literature exploring the association between arsenic in drinking water and the risk of LC showed that elevated concentration of arsenic in drinking water is associated with increased risk of LC and is overall an important risk factor in many parts of the world, including North America.\textsuperscript{(125)}

2.3.15 Outdoor and Indoor Air pollution

Several studies have shown the correlation between LC risk and concentrations of carcinogens from fossil fuel combustion. Carcinogens produced by combustion of fossil fuels include polycyclic aromatic hydrocarbons and metals such as arsenic, nickel, and chromium.\textsuperscript{(63)} In developed countries, indoor pollution that usually affects nonsmokers are SHS and radon,\textsuperscript{(116)} and in developing countries the risk comes from using soft coal and biomass fuels for cooking
and space heating.\textsuperscript{(126)} The effect on LC risk from burning biomass seems to be less than the effect of burning fossil fuels, however both are important risk factors in determining LC risk.

2.4 Lung Cancer Screening

To decrease the burden of LC in the population, attention has been given to preventative strategies, such as promoting smoking cessation. However, there has been a growing interest in screening for LC to increase survival and reduce mortality in the population. As mentioned before, severe symptoms for LC usually do not appear, or individuals do not report symptoms, until the disease is at an advance stage. Screening is a method to detect disease at its earliest stages when potentially curative treatment is possible, therefore reducing the mortality associated with LC in a high-risk screened population. Mortality reduction is one of the fundamental measures of a screening program’s efficacy.\textsuperscript{(10)} When looking at screening techniques, it is important that the benefits of screening outweigh the risks. Potential risks include exposure to radiation, false positive or false negative results, diagnostic complications and unnecessary psychological stress.

Resources providing details of a LC screening regimen are available,\textsuperscript{(127,128)} however a general overview of a screening program will be discussed in this thesis. A LC screening program involves an interdisciplinary team of primary care clinicians, pulmonologists, and other LC specialists, as well as support from nurses and hospital and clinical administration. Information about screening can be provided to clinicians and patients through various means, such as health system newsletters, web-based resources, decision aids, health fairs, letters to patients, pamphlets, a call line or journal articles. If a patient is deemed eligible, then the program schedules a shared decision-making visit for the patient to further assess eligibility and discuss the benefits and harms of screening; these visits can be conducted with a nurse
practitioner who is trained as a counselor or with the primary physician. In addition, smoking cessation resources are made available for patients who smoke. Once eligible patients are informed with all the details of the program and give their consent, they undergo screening using LDCT and results are structurally reported through a management system (i.e. Lung-RADS). Lung Imaging Reporting and Data System (Lung-RADS) helps to manage indeterminant nodules more efficiently. It is used to structure reporting and management of LC screening results and findings by defining categories of positive results and providing management recommendations for each category. Although Lung-RADS is currently being used in North America, there are better nodule management protocols, based on volumetric nodule assessment, being used elsewhere. Patients with negative findings are asked to return for their next scheduled screening, and follow-up of abnormal findings is assessed by Lung-RADS and a collaboration with relevant specialists, and patients are then appropriately referred to diagnostic evaluation and then further to treatment evaluations as needed.

Several LC screening trials have been conducted since the 1950s using sputum cytology and chest radiography. Sputum cytology is the examination of sputum which contains cells derived from the buccal cavity, the pharynx, the larynx, trachea and the lungs. Chest radiography, or x-rays, involves exposing a part of the body to a small dose of ionized radiation to create pictures of the inside of the body. Neither of these screening techniques have been shown to reduce LC mortality. Studies using these methods have had methodological limitations and discouraging results. Studies also demonstrate no benefit from adding sputum cytology to annual CXR and do not address the effectiveness of CXR alone.
2.4.1 Prostate, Lung, Colorectal, Ovarian Cancer Screening Trial

Since previous screening trials using CXR were relatively small with low power and had potential biases,\(^{[11,13,14]}\) there was concern that the true positive effect of CXR screening went undetected in these studies. Due to this and the lack of benefit on LC mortality seen in the previous trials, the National Cancer Institute (NCI) initiated the PLCO trial in 1993 to determine whether screening with CXR reduced death from prostate, lung, colorectal and ovarian cancers. Focusing on the lung component, this trial differed from past LC screening trials because of its large sample size (N = 154,901). It also included women and individuals who never smoked. The PLCO trial ran from November 1993 until May 2012 and was the first large and well conducted randomized controlled trial (RCT) to evaluate the effectiveness of annual screening with CXR. It also represented the largest screening trial for LC with a population from different backgrounds, sexes, and smoking history.\(^{[16,133]}\) Screening was carried out in 10 centers across the U.S. recruiting individuals from the general population. Men and women aged 55-74 years who were healthy were targeted and recruited through direct mail. Informed consent by both the participants and NCI were signed and submitted. Participants also filled out a baseline questionnaire on demographics, smoking history, family history, medical history and past screening.

Out of the 154,901 individuals registered, 77,445 and 77,456 were randomized to the intervention group and control group, respectively, and were followed up for 13 years; details on the randomization and recruitment process in the PLCO trial will be provided in the methods section. The intervention group included annual single-view posterior-anterior chest radiograph screening for 3 years, and the control group received their usual care as recommended by their physician. There was a roughly equal proportion of men and women in the intervention group,
The majority of these participants were 55-64 years of age, from a White, non-Hispanic background and never-smokers. Current and former-smokers underwent baseline CXR screening followed by three annual screens and never-smokers had an initial CXR screening followed by two annual screens. Positive CXRs were specified when the radiologist identified a mass, nodule or any other abnormality as “suspicious” for cancer. Upon a positive screen, participants were sent to their primary health physician for evaluation and management of the disease and medical records containing information on diagnostic follow-up were obtained. Cancers were categorized as NSCLCs (adenocarcinomas, squamous carcinomas, large-cell carcinomas), SCLCs, or carcinoid tumors. (16)

Overall compliance to CXR screening was 83.5% and was generally consistent throughout the screening rounds. (134) At each screening round, never-smokers had the lowest positivity rate and the rates increased with each smoking risk category (current>former>never-smokers, p<0.001). Positive results for initial CXR screening were higher in men than in women (9.6% vs. 8.2%) and in older individuals. (16) Cumulative LC incidence rates through the 13-year follow-up were 20.1 per 10 000 person-years in the intervention group and 19.2 per 10 000 person-years in the usual care control group (relative risk: 1.05, 95% CI 0.98-1.12). During the entire 13-year follow-up period, 18% of all LC cases were detected by screening and 12% were interval cancers (non-screen detected cancers diagnosed among those within 12 months of a negative screen and before the next screen). LC histology was similar by group where majority of cancers were adenocarcinomas. Death certificates were used to confirm causes of deaths, and cause-specific deaths were labelled as those with an underlying cause of LC. In total, 1213 LC deaths were observed in the intervention group vs. 1230 in the control group during the 13-year-follow up. The two groups also demonstrated similar cumulative LC incidences (20.1 vs. 19.2
per 10 000 person-years). These results suggest that annual chest radiographic screening for LC did not influence cumulative LC mortality during the 13-year-follow up period.\(^{(134)}\) In conclusion, the PLCO trial provided evidence that there is no significant impact on reducing LC mortality from annual CXR screening.

There are several limitations of the PLCO study, such as the underrepresentation of African Americans and Hispanics and the fact that the study population is representing those of a higher SES than the general population. Strengths of this study included its large sample size, the addition of women and never-smokers, the study groups were similar at baseline and there was overall high adherence to screening in the intervention group and low contamination in the usual care group. Individuals in the intervention group who underwent a diagnostic follow-up procedure also had a low complication rate (0.4%).\(^{(15,19)}\)

2.4.2 The National Lung Screening Trial

A CT scan uses x-rays in combination with sophisticated computers to produce 3-D and cross-sectional images of organs, tissues, bones and blood vessels inside the body. CT has shown to be more effective than CXR in finding the location, shape and size of a lung tumor.\(^{(18,19,26)}\) However, to reduce radiation exposure and scan time, low-dose computed tomography (LDCT) was introduced. LDCT produces clearer images and finds nodules as small as 0.5cm while exposing individuals to less radiation (1.6mSv vs. 7 to 8mSv from conventional CT).\(^{(10,26)}\) Studies have shown LDCT to be more effective in identifying LCs in their early stages than CXR screening,\(^{(18,20,135)}\) however, many studies have not shown a decrease in LC mortality from using LDCT for LC screening.\(^{(136–140)}\)

To understand the effects LDCT has on LC mortality, the NLST was launched on September 18\(^{th}\), 2002. This trial was based on a smaller LC screening pilot study known as the
Lung Screening Study (LSS). The NLST was conducted by LSS screening centers and the American College of Radiology Imaging Network (ACRIN), funded by the NCI Division of Cancer Treatment and Diagnosis. The NLST was a randomized control trial comparing annual LDCT screening with CXR screening for three years in high-risk current and former-smokers. The primary goal of this trial was to assess LC mortality differences between the two study arms, as well as other measures such as LC incidence, overall mortality and screening and treatments related morbidity.\(^{(17)}\)

Participants were recruited and registered at 33 screening sites across the U.S. At each center, potential participants were targeted through a variety of methods such as direct mass mailing, local media broadcasts and presentations to community groups, such as those at clinics and churches. Intensive efforts were made to recruit individuals from minority populations with the help of other organizations such as National Cancer Institute's Cancer Information Service Partnership Program, the American Public Health Association’s Black Caucus of Health Workers, National Cancer Institute's Spirit of Eagles Program, and the American Cancer Society. Eligible participants were those who were 55-74 years of age, had a history of cigarette smoking of at least 30 pack years (packs per day times years smoked) and, if former-smokers, had quit within the last 15 years. Non-eligible participants were those who had any metallic implants or devices in the chest or back, had cancer (other than skin non-melanoma skin cancer or carcinoma in situ) within the previous 5 years, had a history of LC or removal of any part of the lung (excluding needless biopsy), required home oxygen supplement, participated in another cancer screening trial or cancer prevention study, had unexplained weight loss of 15 pounds or more during the prior 12 months, had recent hemoptysis, had pneumonia or acute respiratory infection within the last 12 weeks or had a chest CT examination within the preceding 18 months.\(^{(17,20)}\)
A total of 53,452 individuals were enrolled; 26,722 were randomly assigned to LDCT screening and 26,730 were assigned to screening with CXR. Randomization was stratified by age, sex and screening center and this was done by using a block size of 6 or 8, which was chosen at random. Each participant provided written informed consent and completed a questionnaire on demographics, smoking behavior, medical history and various other topics. Participants underwent three annual screenings, with the first at baseline shortly after randomization. The median duration of follow-up was 6.5 years, with a maximum duration of 7.4 years in each group. All LDCT scanners were certified for use and met the protocol requirements. Multi-detector scanners were utilized for increased spatial resolution of the whole chest and to reduce exposure to an average of 1.5 mSv. In the CXR group, chest radiographs were acquired though screen film radiography or digital equipment and produced an average radiation dose of 0.02 mSv.

Positive results, that were suspicious for LC, were described as any non-calcified nodule ≥4 mm for the LDCT scan and any non-calcified nodule or mass for CXR. For both measures, interpretations of images were first made independently and then in context, meaning old images from the trial or other sources were also analyzed in a comparative manner. NLST radiologists provided recommendations to the health care providers of individuals with positive screens; any information on diagnostic evaluations and other procedures were obtained and kept on record by the NLST screening centers. Following the diagnostic evaluations that were performed in response to positive LC diagnoses, medical information was acquired on associated complications, pathology, tumor staging and histologic features of the LC. Death certificates were also obtained, and the cause of death was determined by an end-point verification team. Any deaths from diagnostic evaluation or treatment for LC were counted as LC deaths.
Overall, the rate of adherence was high in both study arms (95% in LDCT group and 93% in CXR group). There was a higher rate of positive screens in the LDCT study group compared to the CXR group. Since abnormalities suspicious for LC, that were stable across all three rounds, could be classified as negative at the discretion of the radiologist, the rate of positive screens in both study arms was lower in the last round. LDCT identified clinically significant abnormalities, other than those suspicious for LC, 3.6 times higher than CXR (7.5% vs. 2.1%). In total, 1060 LCs were identified in the LDCT group versus 941 in the CXR group (rate ratio = 1.13; 95% CI 1.03-1.23) and fewer stage IV cancers were diagnosed in the LDCT group than the CXR group in the second and third screening rounds. Overall, there was a significant reduction (6.7%) of all-cause mortality from LDCT screening. However, when LC deaths were omitted in this comparison, there was no significant reduction in overall mortality in the LDCT group. When comparing LC specific mortality, there were 356 deaths in the LDCT group and 443 deaths in the CXR group, which is equivalent to rates of 247 and 309 deaths per 100 000 person-years, respectively. These results corresponded to a 20% comparative reduction in LC mortality using LDCT screening (95% CI 7.9-30.7; p = 0.004). Based on the results, the number needed to screen to prevent one LC death was 320, and this is lower than what is currently seen in breast or prostate screening.\(^{(17)}\)

Risks from LDCT screening in this study included false positive results, which was 3-fold higher in the LDCT group than the CXR group and led to unnecessary invasive procedures and possible psychological stress. Although complications from diagnostic procedures were uncommon, the frequency of a major complication occurring during a diagnostic evaluation following a positive screen was 33 per 10 000 and 10 per 10 000 in the LDCT and CXR group respectively. A total of 16 individuals in the LDCT group (10 of which had LC) and 10 in the
CXR group (all of which had LC) died within 60 days after an invasive diagnostic procedure.\(^{(17,26)}\) Over-diagnosis refers to the detection of cancers that would not be clinically apparent or would not impact a patient’s life if left undiagnosed. Over-diagnosis is associated with LDCT; there was an excess of 120 LCs in the LDCT study arm compared to the CXR group and it was estimated that more than 18% of all LC cases detected in the LDCT arm seem to be over-diagnosed.\(^{(145)}\) Over-diagnosed cases after LC screening commonly occur due to screen-detected patients dying from competing causes prior to cancer diagnosis. CCoD is an important cause of over-diagnosis; among those who died from competing causes in the NLST, 78% had a positive LDCT screening test.\(^{(146)}\) Over-diagnosis is difficult to quantify but it does exist and should be considered when discussing the benefits and harms of LDCT screening. Excess radiation exposure is also a risk associated with LDCT which can lead to further cancer or health complications. Radiation exposure was higher in the LDCT screening group than in the CXR screening group in the NLST, however it has been estimated that radiation exposure would potentially cause one death per 2500 people screened in the NLST.\(^{(17,26)}\) Therefore, the benefit in preventing LC deaths in the NLST is much greater than the radiation risk.

Limitations of the NLST include its lack of comparison with standard of care, which is no screening, and the healthy volunteer effect which can bias results such that they are more favorable to the younger, healthier and better educated participants as compared to the general population. Although the NLST had limitations, this trial provided promising results for the use of LDCT to screen early LCs, making early treatment possible and therefore saving lives.\(^{(17)}\)

The NLST so far is the largest randomized study of LC screening in a high-risk population which has shown a mortality reduction benefit of 20% via LDCT screening. Due to these significant findings, the Unites States Preventive Services Task Force (USPSTF)
recommends LC screening for high-risk current and former-smokers if they have quit smoking within the past 15 years and are between 55-80 years of age with a smoking history of at least 30 pack years.\(^{(22)}\) Several organizations have recommended LDCT screening of high-risk people when high quality follow-up and healthcare are available. These recommendations have come from the following organizations: American Association of Thoracic Surgery, American College of Chest Physicians, American Cancer Society, American Lung Association, Cancer Care Ontario, and National Comprehensive Cancer Network. Centers for Medicare and Medicaid also provide LC screening using LDCT for individuals who are insured under their health care program. Most of these recommendations base their definition of high-risk on the NLST criteria or some variant of it.\(^{(10,23–25,147)}\) It is worth mentioning that results from the Dutch-Belgian Randomized Lung Cancer Screening Trial (NELSON study), which is the second largest RCT after the NLST trial, demonstrated a 26% and up to 61% LC mortality reduction, in men and women, respectively, with LDCT screening over a 5.5 year period and 10-year follow-up. Findings from the NELSON and NLST trial support the implementation of LC screening using LDCT among high-risk individuals.\(^{(148)}\)

### 2.5 Risk Prediction Modelling

For screening to be effective in reducing LC mortality, it must be applied to high-risk individuals for that disease. There has been interest in using accurate risk prediction models to identify people at high-risk for LC as it may result in more LC cases being detected, than the NLST or NLST-like criteria, while reducing the number of false positives and over-diagnosed LC cases. Risk prediction models have gained attention as a more effective method to select high-risk individuals using several different predictor variables, other than only smoking history and age as used in the NLST criteria.\(^{(41–43,46,63)}\)
To understand how risk prediction works, the basics of prediction modelling needs to be understood. Prediction models can be used to supplement and guide healthcare professionals reasoning by providing estimated probabilities based on objective information. Prediction models evaluate probability estimates of the occurrence or absence of a future event, during a specified period.\(^{(43)}\) Risk prediction models are mathematical functions which use covariates or predictor variables to provide estimates which can be converted into absolute probabilities of an event (such as LC) over time. Risk stratification can help identify eligible participants for screening, and therefore can help with detection or screening of disease in undiagnosed high-risk subjects which can help with the management of the disease via early treatment. In a clinical setting, risk prediction models can be used to inform patients about their probability of developing a certain disease which can help with the accompanying decision making regarding further management, including additional medical examinations, or lifestyle changes.\(^{(43,149)}\)

Predictor variables, or risk factors, are usually selected based on prior knowledge of these factors associated with the disease or outcome, as well as their predictive performance in previous models. Predictor variables that were formerly found to be statistically significant (\(p<0.05\)) are usually reflected as candidate predictors, however not all significant predictors need to be included in the model. Backward elimination or backward selection is the most commonly used variable selection method in which a model begins with all variables of interest and the least significant effect that does not meet the level for staying in the model is removed. However, prediction modelling is different from hypothesis testing in that predictor variables which are non-significant (\(p>0.05\)) can still improve prediction, and significant predictors can fail to improve prediction; such variable selection must be guided by judgement and clinical relevance.
The focus is not to establish causal relationships but to improve overall predictive performance which should be used to guide prediction modelling.\textsuperscript{(41)}

Logistic regression, a common statistical method, is used to describe the relationship between a dichotomous outcome variable (in which there are only two possible outcomes) and a set of predictor variables.\textsuperscript{(150)} In logistic regression, the natural logarithm (ln) of the odds of the outcome is predicted, also known as the logit of the outcome and this guarantees that estimated probabilities lie between 0 and 1. The regression coefficients from a logistic regression analysis can be used to compute the odds ratio (OR: the relative odds of the outcome occurring given an exposure compared to being unexposed) for one variable while controlling for all the other variables or risk factors, and can be used to compute probabilities or risks.\textsuperscript{(150,151)}

The Brier score can measure the overall predictive performance of a logistic regression model. The Brier score represents the average error in prediction and is calculated by squaring the absolute differences between the observed and predicted probabilities for each individual and then taking the average across that sample. A lower individual Brier score indicates better prediction, and a Brier score of 0.25 indicates random classification.

There are two essential components, other than the Brier score, for assessing predictive performance: calibration and discrimination.\textsuperscript{(41)} Calibration assesses how well the model’s predicted probabilities compare with the actual or observed probabilities. It is sometimes measured by goodness-of-fit tests, which reflects the agreement between the observed and predicted probabilities. Calibration by the Hosmer-Lemeshow (HL) goodness-of-fit test is commonly used in logistic regression. A calibration plot shows predicted probabilities of groups (restricted by ranges of individual probabilities) on the x-axis and the mean observed outcome on the y-axis and this plot is a graphical representation of the HL goodness-of-fit test where a
perfect calibration would lie on a 45-degree line of the plot. The HL test however has several limitations which make it disadvantageous to use, including that it is often non-significant for small sample sizes but almost always statistically significant for larger sample sizes, and overall has limited power to evaluate poor calibration.\(^{137,139,142}\) Another method for assessing calibration involves using a smoothing technique, such as locally weighted scatterplot smoothing (LOWESS), which plots a smoothed curve line showing the relationship between observed and predicted probabilities. This allows the computation of the absolute error between the observed and predicted probability for each participant and can help identify patterns of error. In addition, the median and 90th percentile absolute error, corresponding to the difference between the observed and the model estimated probability, provide a good idea of how well the model is calibrated.\(^{41}\) It is important to have high calibration if risk prediction models are to be used to select individuals for screening.

Discrimination refers to the models’ ability to differentiate between two outcome classes correctly. It is the probability that a participant whom is randomly selected will be correctly classified by the model as either experiencing the outcome (e.g. being diagnosed with LC) or remaining event free. The receiver operator characteristic curve (ROC) with the area under the curve (AUC) is the most common measure of discrimination in logistic regression. The ROC plots the true positive proportion (sensitivity) versus the false positive proportion (1-specificity). Each point on the ROC represents a sensitivity/specificity pair corresponding to a threshold point. Sensitivity is the probability a test result will be positive when the disease is present (true positive) and specificity is the probability a test result will be negative when the disease is not present (true negative). The AUC helps determine how well a parameter can distinguish between two outcomes. A higher AUC represents better discrimination, where an AUC value between 0.8
and 0.9 is excellent and an AUC of 0.5 represents random classification. The concordance or c-statistic is also widely used to measure discrimination for binary outcomes and in survival analysis when Cox models are used. The c-statistic gives the probability that more cases (individuals who will develop the outcome of interest) be detected as high-risk by the model than non-cases. The AUC and c-statistic are mathematically equivalent, and both provide information on how well a model can discriminate cases from control. It has been suggested that prediction models with c-statistic or AUC over 0.8 may be useful for individual level prediction.\(^{(154)}\)

### 2.5.1 Noteworthy Risk Prediction Models

Several previous risk prediction models to estimate LC risk have been published and they differ in the data from which they were developed in terms of design (cohort vs. case-control), country or region of population sampling and inclusion criteria. There is variation in the model type (e.g. logistic, survival analysis), predictor variables used, requirement of direct patient contact or biomarker data and evaluation of predictive performance.\(^{(48)}\) The model by Bach et al.\(^{(44)}\) evaluated LC risk among smokers from the Beta-Carotene and Retinol Efficacy Trial (CARET), which was a randomized controlled study evaluating the impact of beta-carotene and vitamin A supplementation on LC incidence and mortality. This model is based on follow-up data from the CARET study and estimates 1-year LC incidence using Cox-proportional hazards regression. The predictors in the Bach model include age, sex, smoking intensity (cigarettes per day), smoking duration, number of years since smoking cessation and asbestos exposure. The model was run periodically 10 times to estimate a 10-year absolute risk. This model demonstrated a c-statistic, which was internally validated by cross validation, of 0.72 representing moderate discrimination. Calibration was assessed by plotting the observed versus expected probabilities of LC risk by deciles of risk and calibration overall was reported to be
good when internally validated, however, the model had satisfactory calibration when validated externally. Limitations of the model include the measure of asbestos exposure and generalizability. The measurement of asbestos exposure in the CARET study and the model may not be clinically efficient for selecting individuals for LC screening. It was identified using detailed criteria on employment and history of occupational exposure to asbestos and this may not be practical in measuring asbestos exposure. Generalizability is a concern because the model was restricted to a subset of people who are over 50 with a smoking history. When validated in data from other studies, the Bach model demonstrated overall moderate discrimination and poor to moderate calibration.

Cassidy et al. developed a LC risk prediction model based on data from a matched case control study in Liverpool known has the Liverpool Lung Project (LLP). The model used subjects between the ages 20-80 years as in the study, and data from 579 LC cases and 1157 age, sex and smoking status-matched population-based controls. This model predicts 5-year LC risk and includes various predictor variables including occupational asbestos exposure, smoking duration, family history of cancer and prior diagnosis of a malignant tumor or pneumonia, all of which were shown to be significantly related to LC risk (p<0.05). Discrimination was also shown to be high (AUC=0.71), however this is an inflated overestimation due to the inclusion of never-smokers in the sample. There was no direct measurement of calibration reported for the LLP model, however when validated in European and American external datasets, calibration was thought to be poor, especially in regions where important thresholds for LC screening selection may occur. The LLP model also contained several weaknesses such as the complex measurement of asbestos exposure making it clinically impractical, the inclusion of never-smokers, lack of smoking related predictors other than smoking duration, and most
importantly, the fact that cases and controls were matched on smoking status, making it difficult to evaluate the effect of smoking exposure on LC risk.\(^{(45)}\)

The Spitz model\(^{(47)}\) was based on epidemiological data consisting of 1851 LC patients matched with 2001 controls based on race/ethnicity, sex, age and smoking status. Participants were recruited from the Thoracic Center at the University of Texas M.D. Anderson Cancer Center. To decide what risk factors to include in the model, univariate analysis was done to see which variables were statistically significantly associated with LC risk. Among never-smokers, exposure to second hand smoke and family history of cancer was significantly associated with LC. Additional predictor variables that were found to be significantly associated with LC in current and former-smokers were emphysema, dust, age at smoking cessation, hay fever, pack-year for smoking, asbestos, and family history of smoking related cancer. The AUC values were computed for never-smokers, former-smokers and current-smokers and were 0.57, 0.63 and 0.58 respectively, and overall showed poor discrimination. Even though calibration was shown to be adequate due to the non-statistically significant HL goodness-of-fit test statistics (0.777 for never-smokers, 0.712 for former-smokers, and 0.688 for current-smokers), it does not signify good calibration because as previously mentioned the HL statistic has many weaknesses making it an unfit measure for calibration.\(^{(41,47)}\) A big weakness of this model was that the case-control data was matched on age and smoking status, which are strong predictors of LC risk; matching individuals on age and smoking status reduces their effects as risk factors and therefore diminishing their importance in the overall model.\(^{(41)}\) Overall, risk prediction models such as the Spitz and LLP model demonstrate methodological limitations and poor external validation. Since both models were developed using case-control data, there was no accurate measure of true
incidence of LC in the population of interest and therefore estimates of calibration were unreliable.

Several other models have also been produced in special populations to consider racial disparities in LC incidence and risk. Such models include the African American model by Etzel et al. and the Chinese Genetic model by Li et al., however these models indicate low discrimination in their external validation. \(^{(41)}\)

2.5.2 Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial Model Version 2012

The PLCO\(_{M2012}\) model is a logistic regression model developed in the PLCO control group of smokers which predicts the probability of developing LC within 6 years (follow up data up to 6 years was used to make is comparable to the NLST follow-up data). Predictor variables were selected based on predictive performance, therefore the model was not limited to predictors with a p-value less than 0.05. The PLCO\(_{M2012}\) model consists of four smoking and seven non-smoking predictor variables: smoking intensity (average number of cigarettes smoked per day), smoking duration (the number of years smoked), smoking status (current or former-smoker), quit-time in former-smokers, age, COPD, personal history of cancer, family history of LC, education, race/ethnicity and BMI. \(^{(49)}\) In this model, the risk of LC increased with age, having a black race/ethnicity, lower SES status (determined by level of education), lower BMI, self-reported history of COPD, personal history of cancer, family history of LC, current smoking, increased smoking intensity and duration, and in former-smokers, shorter time since quitting.

This model was validated in the PLCO intervention (CXR) group of smokers; discrimination was assessed by AUC, and model calibration was assessed by plotting a smoothed curved line with a LOWESS plot and by the median and 90th percentile absolute error. \(^{(49,159)}\)
AUC for the validation data was 0.797 as compared to 0.689 when the NLST criteria were applied to the validation sample. The model also showed good calibration, where the difference between observed and predicted probabilities of LC risk was less than 0.010 in half the validation sample (median absolute error: 0.009) and less than 0.043 in 90% of the sample (90th percentile absolute error: 0.042). When the NLST criteria was applied to the PLCO intervention group, 14 144 of 37 332 smokers (37.9%) were eligible for screening. For an equal number of individuals to be eligible for screening applying the model, the cut-off risk value was 1.345%, and this threshold allowed for comparison between groups of the same size that were eligible by each criterion. The PLCOM2012 model detected 81 more of the 678 LCs than the NLST criteria with greater sensitivity (83.0% vs. 71.1%, p<0.001) without the loss of specificity (62.9% vs. 62.7%, p=0.54) and had a significantly higher positive predictive value (PPV) (4.0% vs. 3.4%, p<0.001), which demonstrates the probability that subjects with a positive screening test truly have the disease. These findings reveal that using an accurate risk prediction model, such as the PLCOM2012 model, for selecting high-risk individuals for LC screening is more efficient than applying the NLST criteria.

However, there are some limitations of the PLCOM2012 model. Since this model was made in data consisting of individuals between ages 55–74 years, the predictive performance of the model outside this age range should be assessed. The model did show good discrimination and accuracy among all age groups in an Australian sub-cohort of smokers, specifically sensitivity was highest for the 75–79-year age group. Also, the SES of the study population was greater than those of the general population, therefore the model should be tested in other populations with various characteristics to assess generalizability.
The PLCO\textsubscript{M2012} model has many advantages over the NLST or other NLST like criteria. First is the use of a risk threshold for selecting high-risk individuals for screening. In 2014, Tammemägi \textit{et al.}\cite{50} established a PLCO\textsubscript{M2012} risk threshold above which LC mortality rates in the NLST intervention (CT) arm appeared to be constantly lower than those in the NLST control (CXR) arm. This provides a threshold above which there is reliable evidence of a mortality benefit from screening and this was found to be at the 65th percentile risk in the PLCO\textsubscript{M2012} model, which corresponds to a risk of 0.0151. In the 65th to 100th percentile risk range, the risk of LC mortality is 4 times larger than in the 30\textsuperscript{th} to <65\textsuperscript{th} percentile risk range and is statistically significant (6.43 per 10 000 person-years, 95\% CI 1.53 - 11.33, p=0.010), therefore these individuals would benefit most from screening. The number needed to screen (NNS) to prevent one death in this risk range is 255, which is lower than the NNS in the 30th to <65th percentile group (963) and the NNS of 320 reported for the NLST study. This risk threshold also has a high sensitivity, specificity and PPV for 6-year LC incidence and LC mortality since this threshold captures most LC cases and deaths in the PLCO and NLST population. When compared to the USPSTF criteria, the PLCO\textsubscript{M2012} model risk $\geq$0.015 selected 8.8\% fewer individuals for screening ($p$<0.001) while identifying 12.4\% more LC cases. This suggests that using the PLCO\textsubscript{M2012} model with a risk threshold $\geq$0.015 could be more cost-effective over the USPSTF or NLST criteria. \cite{49,50}

Other advantages of this model are that it does not use pack-years smoked to describe smoking history, as opposed to the NLST or USPSTF criteria. The assumption that LC incidence rate is proportional to the product of smoking intensity and smoking duration is incorrect because these smoking variables have different magnitudes of effect on LC incidence.\cite{161} The PLCO\textsubscript{M2012} model separates pack-years into its two components, smoking intensity and smoking duration, to
understand the effect of each predictor. Unlike the NLST criteria, the PLCOM2012 model does not dichotomize its continuous variables; categorizing continuous variables causes valuable information to be lost.\(^{(162)}\) The PLCOM2012 model also considers nonlinear effects associated with certain risk factors. For example, smoking intensity has a nonlinear relationship with LC, the increase in risk becomes smaller as smoking intensity increases. Therefore, representing smoking intensity as a nonlinear variable rather than a linear variable is more accurate and significantly improves prediction.\(^{(49)}\) The PLCOM2012 model contains a total of 11 predictor variables, making it seem more complex than the NLST criteria. However, these variables have been shown to be significantly associated with LC and improve the predictive performance of the model while not requiring complicated measurements to be taken, such as blood tests to look at biomarkers associated with LC. Therefore, this model would be clinically feasible in public health screening programs.\(^{(41)}\)

### 2.6 Competing Causes of Death

It is important to recognize that several competing risks exist for smokers, especially heavy smokers.\(^{(54,163)}\) Evidence shows a higher risk of hepatocellular carcinoma, the most common type of liver cancer, in smokers vs. non-smokers, after controlling for potentially confounding factors.\(^{(164–166)}\) Studies show an association between smoking and colorectal polyps and colorectal cancer. Although adjustment for confounding factors varied to some extent across studies, being a current-smoker and smoking for a longer duration was consistently associated with increased risk of colorectal cancer.\(^{(167,168)}\)

Overall, there is sufficient evidence to suggest a causal relationship between cigarette smoking and increased all-cause mortality, as well as cancer-specific mortality,\(^{(169)}\) where 1 out of 3 cancer deaths are caused by smoking.\(^{(163)}\) Cigarette smoking has been shown to be strongly
associated with COPD,$^{(163,170)}$ and associated to a lesser degree with other respiratory diseases such as asthma, tuberculosis and idiopathic pulmonary fibrosis.$^{(163,171–174)}$ One study looked at the relationship between cigarette smoking and risk of death due to coronary heart disease (CHD), cerebrovascular disease, respiratory disease, LC and other cancers caused by smoking. Statistically significant reductions in risk for all types of mortality were seen with smoking cessation and a 46% reduction in CHD and cerebrovascular disease risk was observed in the first 5 years of quitting.$^{(54)}$ A negative association is also seen between LC survival and the existence of several comorbid illnesses such as congestive heart failure, diabetes, moderate or severe liver disease, dementia, renal disease, and cerebrovascular disease. These associations vary by LC stage, where the survival of localized LC patients has shown to be negatively associated with congestive heart failure (HR = 1.73, 95% CI 1.33–2.25), diabetes with complications (HR = 2.17, 95% CI 1.12–4.18), and moderate or severe liver disease (HR: = 3.74, 95% CI 1.09–12.83). The survival of patients with regional disease is also negatively associated with congestive heart failure (HR = 1.26, 95% CI 1.04–1.52), dementia (HR = 2.33, 95% CI 1.20–4.52), and renal disease (HR = 1.44, 95% CI 1.10–1.88). Compared to individuals with localized or regional LC, the survival of patients with distant disease is negatively associated, to a lesser extent, with congestive heart failure (HR = 1.19, 95% CI 1.05–1.40) and cerebrovascular disease (HR = 1.27; 95% CI 1.08–1.48).$^{(55)}$

Increased age is also associated with many comorbidities, mainly chronic conditions,$^{(175)}$ which negatively influence prognosis of LC.$^{(176)}$ In a competing risks analysis, it was found that with increasing age among those with LC, the risk of surgery increased and non-LC specific causes of death became more significant after surgery.$^{(177)}$ Among LC patients who had undergone lung resection, individuals in the eldest cohort ($\geq$75 years) had a 5-year non-LC
specific cumulative incidence of death of 9.0%, compared to 1.8% in the youngest cohort (<65), and among all participants (≥65 years), the leading cause of death at 1 year was non-LC specific.\(^{(177)}\) Overall, the risk for developing LC, especially among smokers and older individuals, is associated with the presence of co-existing conditions, indicating worse survival.

The risk of death from non-LC or competing factors start elevating around the 35th PLCO\(_{M2012}\) percentile risk and sharply rises at the 65th PLCO\(_{M2012}\) percentile risk, demonstrating that the number of deaths from competing causes is also increasing after the 1.5% risk/6yr threshold. This suggests that the model is predicting non-LC deaths and is causing high-risk individuals who are likely to die from other non-LC related causes to be selected for screening.\(^{(50)}\) As mentioned previously, compared to individuals with no comorbidities, screening individuals with comorbidities who are at high-risk for non-LC death is associated with worse survival as these individuals have an increased risk of complications from downstream diagnostic evaluations and treatment. Early screening, diagnosis and treatment are associated with side-effects such as fatigue, emotional distress, pain and decreased physical function, and these negatively impact the number of QALY gained; among those not healthy enough to withstand diagnostic or treatment procedures, side-effects and risks are magnified and the number of QALY gained is compromised.\(^{(178)}\)

Screening individuals with comorbidities also increases the risk of overdiagnosis, which is a risk or limitation of screening. An individual is defined as being over-diagnosed if they die from other causes before their LC would have affected their well-being. Overdiagnosis can subject individuals to unnecessary treatment, morbidity, follow-up, cost and anxiety.\(^{(145)}\) In the NLST, it was observed that among all LDCT screen-detected tumors, 18.5% (95% CI 5.4%-30.6%) of LC cases detected by LDCT screening were an overdiagnosis.\(^{(145)}\) In another analysis
from the NLST sub-study, there was a 12% reduction in LC specific deaths associated with CT screening in those at highest risk (highest tertile) compared to 23-24% in other lower risk tertiles; this was associated with greater deaths from competing risks such as cardiovascular disease, respiratory disease and other cancers.\(^{(179)}\)

Evaluating which individuals are at high-risk for LC but are at minimal risk for death from other causes might help clinicians decide if screening would be beneficial for those individuals. Developing a model which predicts the probability of dying from non-LC causes within a certain amount of time may help improve the efficiency of the original LC risk prediction model.\(^{(46)}\) A logistic regression model predicting the risk of death in 5 years from competing risks or non-LC deaths has been developed. Predictor variables in this model were selected by observing the characteristics of those in the NLST trial and those in the PLCO trial who are NLST eligible and if these variables had significant OR values (p<0.001) in univariate analysis. In both the PLCO and NLST trials, various diseases were more common in those dying. This model includes sociodemographic variables (age, sex, race/ethnicity, education, marital status), smoking exposures (smoking status, smoking intensity, smoking duration), and comorbidities (BMI, personal history of cancer, heart attack/heart disease, stroke, hypertension, COPD, emphysema, diabetes). Although these predictors have shown to be significantly related to non-LC deaths,\(^{(57)}\) it is important to understand which predictors of the PLCOM\(^{2012}\) model are significantly related to CCoD. Models which assesses the relationship between predictor variables of the PLCOM\(^{2012}\) model and non-LC death will provide information on which components of the PLCOM\(^{2012}\) model are contributing to the selection of individuals at high-risk for non-LC death. These models, along with other approaches to evaluate these predictors, will be discussed in the next section.
3. **Methodology**

This chapter outlines the study design and methodology used to carry out the objectives of this study. The sample data source will be reviewed along with details on data collection and organization. Importantly, analytical strategies used for model building and comparing the two selection methods will be discussed in this section.

**3.1 Study Design**

This study was carried out using a historical cohort, or retrospective study design, in which analysis was conducted on existing follow-up data from the PLCO trial, specifically PLCO smokers’ data (N= 74,207). The PLCO trial was prospective in that exposure measurements were made in disease-free individuals who were subsequently followed to determine the onset of disease at a later point in time. The study population will be discussed shortly.

**3.2 Endpoints**

A primary endpoint of our study was 6-year LC incidence; six-year follow-up data was used to make the model comparable to the NLST study, which had a 6.5-year follow-up. The second outcome of our study was 5-year non-LC death. 5 years of follow-up for competing causes of death was chosen because clinicians are often interested in reporting disease survival in 5-year periods to help with decision making. Hence, 5 years was used in our study in case the competing causes of death model, or any variant of it, will be used in a clinical or public health setting by clinicians. Predictive performance of models are also more powerful and accurate over short terms due to less variability of the study population over time.\(^{(41)}\)
3.3 Study Population

The PLCO trial was a large randomized trial measuring the effectiveness of screening methods for prostate, lung, colorectal and ovarian cancers compared to standard medical care, on LC mortality. The ten screening centers were scattered across the U.S. and each screening center developed an appropriate recruitment plan to recruit between 5,000 and 30,000 individuals from their neighboring areas for a total of ~37,000 males and females in each study arm. Recruitment began in November 1993 and ended in September 2001; participants were recruited from a range of sources, with the most common method being direct mailing. The source of the mailing list included those within the health systems associated with each screening center, free lists provided by organizations and those purchased from commercial enterprises. Community outreach was a method used for minority recruitment for African Americans and Hispanics. Mass media using TV, radio and newspapers was a less common recruitment method.

3.3.1 Inclusion & Exclusion Criteria

Men and women aged 55-74 years who reported no prostate, lung, colorectal, or ovarian cancer, or active treatment for any cancer (except non-melanoma skin cancer), were included in the study. Individuals excluded from the study were those participating in other cancer screening trials or primary prevention trials, had taken finasteride in the preceding 6 months, had surgical removal of the entire prostate, one lung, or the entire colon, or had a colonoscopy, sigmoidoscopy, or barium enema examination in the preceding 3 years. Men who had more than one prostate-specific antigen test were also omitted.

3.3.2 Ethical Approval & Informed Consent

Once participants were recruited and enrolled into the PLCO study, they were asked to sign consent forms. Seven centers used a single consent approach where consent was obtained
prior to randomization. Three centers initially used a dual consent approach in which two consent forms were administered, one before randomization for information on baseline characteristics and the second for the intervention group before screening. Consent forms were approved by the NCI and the National Institute of Health (NIH), and each study center also received approval from their own institutional review board. Participants were educated about the discomforts and risks associated with screening, such as false positive results, and were also informed about any diagnostic and treatment procedures following screening.

3.3.3 Randomization

The Study Management System (SMS) randomly assigned participants into the intervention or control arm using block randomization and participants were stratified by study center, gender and age. A total of 77,445 and 77,456 participants were randomized into the intervention and control study arm respectively. Upon entry, participants completed a self-administered baseline questionnaire regarding details on sociodemographic characteristics, family history of LC, personal medical history, smoking history and cancer screening history. The two study groups were similar; approximately half were women, 64.1% were 55 to 64 years of age, about 45% were never-smokers, 42% former-smokers, 10% current-smokers and 2.4% (in the intervention group) and 4.0% (in the control group) had unknown smoking status. The majority of participants in each trial arm were also of White/non-Hispanic race/ethnicity and college graduates.

The primary endpoint of the PLCO trial was cause-specific mortality throughout the 13 years of follow-up, and some secondary endpoints included cancer incidence, screening compliance, positivity rates, diagnostic follow-up and all-cause mortality. Information concerning the diagnosis of cancer as well as deaths from other causes were acquired through
annual questionnaires and follow-up, and end points were confirmed via the National Death Index (NDI). Additional information on all incident PLCO cancers and deaths were supplemented through the utilization of cancer registries, if such data were available.\(^{(134)}\)

### 3.4 Data Checking

Prior to analysis, data checking, and cleaning was conducted using Stata Software Package (StataCorp. 2015. Stata Statistical Software: Release 14.2. College station, TX: StataCorp LP). The datasets from the PLCO and NLST trials were previously harmonized, meaning common information was collected on the participants of both trials to the best of ability. This allowed for the analysis of associations between various factors using a larger sample size. This was done prior to this study by Information Management Systems, Inc. (IMS) through the National Cancer Institute. Participant information was provided for many variables such as study characteristics (personal identifiers, eligibility status, entry date, randomization groups and many more), demographics, risk factors, mortality status, cancer diagnosis and cancer characteristics; these variables have been appropriately coded into the software. However, in our study all personal identifiers (date of birth or death) had been removed prior to delivery to guarantee anonymity.

The IMS attempted to provide clean data for the PLCO population, therefore minimal cleaning was conducted. Data cleaning and checking consisted of managing outliers, errors, extreme or impossible values and missing data in the data set. Outliers are defined as values 3 times the interquartile range (25\(^{th}\)-75\(^{th}\) percentile). Outliers and extreme values were assessed by Stata commands and boxplots and were excluded from analysis or truncated at a certain value. Errors in the data set were checked for, such as coding errors or unusual and illogical values. Missing values were dealt with by assessing the level of missingness. There can be many reasons
for missing data, such as incomplete questionnaires, incomplete follow-up due to participants dropping out, ambiguous information, or certain variables not applicable to specific participants. Data can be missing in three ways: at random, completely at random and not at random,\(^{\text{181}}\) and the type of missingness in our dataset was unknown. The proportion of missing values, for each variable, were computed on Stata. Participants were missing information on the following variables: smoking status, education level, BMI, history of COPD, personal history of cancer, family history of LC, smoking intensity, smoking duration and smoking quit-time. Missing data on participants eligibility to either criterion was also observed; however, the overall proportion of missing data was within reasonable limits (<10%), therefore observations with missing values were excluded from analysis.

3.5 Analytical Procedures

3.5.1 Descriptive Statistics

All statistical analysis was performed using Stata Software Package. Descriptive statistics for sociodemographic variables and risk factors were analyzed. Statistics for each variable were stratified by LC diagnosis during 6-year follow-up. Means (or medians for non-normally distributed variables), standard deviations and interquartile ranges were calculated for continuous variables, and two sample t-tests and non-parametric test to trend was performed for normally distributed and non-normal continuous variables, respectively, to identify statistically significant differences between individuals with LC and those without. Frequencies, and row and column percentages were calculated for categorical variables, and a Chi-squared (when expected cell count was greater than 5) or Fisher’s exact test (expected cell count equal to or less than 5) was performed to assess for differences between groups. Logistic regression and contingency table analysis provided OR estimates for continuous variables and categorical variables. ORs
estimated for continuous variables described the likelihood of developing LC for a one-unit increase of a specific risk-factor and the ORs for categorical variables described the likelihood of developing LC among those with the risk factor compared to those without it.

3.5.2 Study Objective 1 – To examine whether the PLCO_{M2012} model, at 1.5% and 2% six-year LC risk threshold, identified high-risk individuals for LC incidence better than the NLST criteria

Contingency table analysis (2x2 tables) was utilized to stratify participants into two groups, those eligible for screening by the NLST criteria and those eligible for screening by the model. The proportions of participants in the discordant cells (those selected for screening by one method and not the other) were computed. Contingency tables also provided the absolute number of LC cases detected and the probability of developing LC in 6 years among those selected by each criterion, as well as in the discordant groups. This provided insight into which selection method was more efficient in selecting those for screening, namely, which criterion selected less people for screening while detecting significantly more LC cases.

Sensitivities, specificities, positive predictive values (PPV), false positives and false negatives of each selection method were estimated using contingency table analysis. Sensitivity (true positive rate) identifies the proportion of individuals who truly do have the disease and were criteria-positive and this informed us about the criteria’s ability to correctly classify individuals who subsequently developed LC. Specificity (true negative rate) identifies the proportion of individuals who do not have the disease and were assigned the correct negative criteria classification, this provided information on the criteria’s ability to correctly classify individuals as disease free. The PPV measures the probability that the disease is present in an individual when the test is positive. These accuracy measures provided insight into the
performance of these two different methods when applied to the general population. There is usually a tradeoff between sensitivity and specificity; sensitivity is important because failure to detect LC early leads to fatal results, therefore a selection method with high sensitivity without loss of specificity is optimal.

### 3.5.3 Study Objective 2 – To investigate if PLCOM2012 model predictors selected high-risk individuals who were truly at higher risk for developing LC

The analyses performed in this study objective were essentially the same as the prior objective, however, the second study objective focused on evaluating each individual PLCOM2012 model predictor in the discordant groups directly. Contingency table analysis identified which predictors in the PLCOM2012 model were important in improving the identification of individuals who developed LC compared to the NLST criteria. This was done by first splitting up the PLCO smokers’ population into those eligible by the model (at ≥1.5% and at ≥2%/6yr risk) and not the NLST criteria (PLCOM2012+/NLST criteria−) and those eligible by the NLST criteria, but not the model (NLST criteria+/PLCOM2012−); this step was already completed as part of the first study objective. The proportions of individuals eligible, number of LC cases detected and 6-year LC incidence risk in the high-risk strata for each predictor variable were then estimated for each predictor in the two discordant groups; if these statistics were in favor of the model (PLCOM2012+/NLST criteria−) over the NLST criteria (NLST criteria+/PLCOM2012−), then components of the PLCOM2012 model are truly important predictors of LC risk.

Pearson’s chi square test or Fishers exact test was used to observe statistically significant differences between individuals in the high-risk strata who developed LC in 6-year times and individuals who did not. McNemar’s test was also conducted to find statistically significant differences between the discordant pairs, while providing ORs, which represent the odds of
being eligible for LC screening by the PLCOM\textsubscript{2012} model vs. the NLST criteria in those with the given risk factor, and the odds of developing LC during 6-years among high-risk individuals in the PLCOM\textsubscript{2012}+/NLST criteria– group for all predictor variables. The corresponding 95% confidence intervals were computed using binomial exact methods.

**3.5.4 Study Objective 3 – To assess whether any predictors in the PLCOM\textsubscript{2012} model were strongly associated with non-LC 5-year deaths (deaths from competing causes)**

First, logistic regression was utilized to assess the relationship between the independent variables (PLCOM\textsubscript{2012} predictor variables) and the dependent variable, CCoD; this approach was chosen because the outcome, non-LC death, is dichotomous. Logistic regression provides parameter estimates, or beta coefficients, which are used to predict the log odds (logit) of the dependent variable. The logit is a transformation of the logistic model which allows binary outcomes to be converted into continuous ones. Beta coefficient values are in log-odds units and are difficult to interpret, so they are often converted into odds ratios.

Univariate logistic regression models were made for each predictor variable and produced beta coefficients and odds ratios to observe the magnitude of effect of each variable on non-LC death. This allowed us to see which variables were strongly related to competing causes of death, however, not considering the effect of other predictors. Therefore, a 5-year non-LC death multivariate logistic regression model was developed, allowing for observation of the effect of each variable on the outcome while adjusting for other variables in the model.

**3.5.4.1 Variable selection & handling continuous predictors**

Variables were selected based on biological reasoning and their effect on the model’s predictive performance, rather than an arbitrary p-value cut-off point. This was done by removing variables one by one, from the overall model, and evaluating their relative contribution
on the model’s predictive performance (i.e. by observing how much the ROC-AUC value decreased when variables were omitted). Continuous predictors in the 5-year non-LC death univariate and multivariable models were not categorized, this was done to avoid losing information and statistical power. Multivariable fractional polynomial (MFP) analysis allowed for modelling of non-linear functions. To briefly explain, MFP analysis accounts for non-linear effects of continuous variables by applying appropriate power transformations on these variables, and in this study the suggested transformations from MFP analysis were used in place of the original linear terms in the 5-year non-LC death logistic regression model.

3.5.4.2 Model covariates

PLCO_{M2012} model variables were modified if deemed appropriate. Continuous variables such as age and BMI were centered to the mean (subtracting mean from independent value). Centering values helps reduce collinearity between variables and produces regression estimates for an intercept value when continuous variables are coded as 0, this allowed for better understanding and interpretation of the intercept, which represents the background level of LC incidence in the population. In a model with mean-centered variables, the intercept terms of the model suggest the expectation of the response at the mean of the predictors. The following is the list of predictor variables of the PLCO_{M2012} model that were incorporated into the multivariate CCoD model:

1. Age at baseline is recorded in years and was used as a continuous variable.

2. Race/ethnicity is coded as “White, not Hispanic” =1, “Black, not Hispanic” =2, “Hispanic” =3, “Asian” =4, “American Indian or Alaskan Native” =5 and “Native Hawaiian or Pacific Islander” =6 and was further categorized into two groups: White, not Hispanic/ Other =1 and Black, not Hispanic =2.
3. Education is coded as “less than high school” =1, “high school graduate” =2, “post high school graduate training” =3, “some college” =4, “college graduate” =5, “postgraduate” =6, and “other” =7.
4. BMI at baseline is recorded as body mass divided by the square of the body height and was used as a continuous variable.
5. Family history of LC is coded as “no” =0 and “yes” =1.
6. History of COPD is coded as “no” =0 and “yes” =1.
7. Personal history of any cancer at baseline is coded as “no” =0 and “yes” =1.
8. Cigarette smoking status is coded as “never” =0, “former” =1 and “current” =2.
9. Smoking intensity is recorded as cigarettes smoked per day and was used as a continuous variable.
10. Smoking duration is recorded as number of years spent smoking and was used as a continuous variable.
11. Quit-time is recorded as number of years since smoking cessation and was used as a continuous variable.

3.5.4.3 Impact of each model factor on predictive performance of CCoD and LC incidence

To assess which PLCOM2012 model predictors might be predictive of 5-year non-LC death more than, less than or to the same extent as 6-year LC incidence, the predictors were placed in rank order by how much each variable predicted each outcome. This was achieved through individually omitting predictor variables from the 5-year non-LC death and 6-year LC incidence model. Predictive performance was assessed by measuring the drop in ROC-AUC; predictor variables which resulted in the largest drop in ROC-AUC when removed from either model had the largest effect on predictive outcomes.
3.5.4.4 Model Evaluation

Discrimination, or the ability to classify participants (for each predictor variable) correctly in terms of the outcome (having died within 5 years from competing causes vs. not having died within 5 years from competing causes), was assessed by the ROC-AUC value. The common methods of evaluating calibration were discussed previously in this thesis. In this study, calibration was evaluated in several ways, including calculating the mean absolute error, 50th and 90th percentile absolute difference and the Brier score. The p-value of the Spiegelhalter’s z-statistic was computed alongside the Brier score and this statistic evaluated the extent to which an individual Brier component is extreme, with a larger p-value (>0.05) indicating better calibration. Calibration was also assessed through a calibration plot which provided a visual representation of the observed versus predicted 5-year non-LC death probabilities. A LOWESS curve was plotted against a dashed-line demonstrating a slope of 1.0 (perfect calibration) for reference. Bootstrapping techniques were employed to produce bias-corrected confidence intervals for both the AUC and Brier scores.

Lastly, the original PLCOM2012 model scores were used to predict 5-year non-LC death to further assess if the model was increasing the selection of individuals who are at high-risk of dying from competing causes within 5 years. This was done by running a univariate logistic regression in which the dependent and independent variables were 5-year non-LC death and the 6-year LC probability produced by the PLCOM2012 model, respectively. The ROC-AUC value was calculated to demonstrate the extent to which the original PLCOM2012 model risk estimates predicting non-LC death.

3.5.4.5 Model Diagnostics
Several diagnostic tests were carried out to check for the assumptions of the 5-year non-LC death logistic regression model. Multicollinearity was investigated amongst candidate predictors using correlation matrices and variation inflation factors (VIF), which assessed the degree to which collinearity among the predictor variables reduced or worsened the precision of estimate coefficients.\textsuperscript{(183)} Strongly collinear variables were dealt with as appropriate, either by being combined if possible or excluded from the model, however in this study no variables were strongly correlated. Standardized Pearson residuals of the model were plotted against the model predicted values to check for errors associated with the predictors. Pearson residuals represent the differences between the observed and fitted probabilities. Deviance residuals, another measure of goodness-of-fit which were utilized in this study, are defined as the measure of disagreement between the limits of the observed and fitted log-likelihood functions.\textsuperscript{(150)} The Pregibon leverage statistic was also used to measure the influence of each observation on the regression equation. This is important in identifying influential observations that might have been problematic, and these observations were assessed and handled as necessary.
4. Results

4.1 Ethics Approval

The PLCO study dataset was obtained through NCI’s Cancer Data Access System after being approved by U.S. NIH Officials and Brock University through a Data Transfer Agreement. This study also has been reviewed and received ethic clearance through the Brock University Research Ethic Board.

4.2 Study Population Characteristics

The distributions of predictor and sociodemographic variables among the 74,207 ever-smokers in the PLCO screening trial are presented in Table 2. Double sided P-values and 95% confidence intervals (CI) are presented alongside all effect estimates. A higher proportion of PLCO smokers were male (57.7%), with a mean age of 62.4 years, from a White/other non-Black background (94.6%), post high-school graduates (69.7%), with a mean BMI of 27.4 kg/m$^2$, had no personal history of cancer (95.4%), no family history of LC (88.1%) and no self-reported history of COPD (91.1%). The mean smoking intensity in this population was 24.7 cigarettes smoked per day and the average smoking duration was 27.6 years. 80.1% of participants were former-smokers with a mean smoking quit-time of 16.2 years; overall a higher proportion of participants were eligible by the NLST criteria (38.3%), as compared to the PLCOM$^{2012}$ risk prediction model at $\geq$1.5%/6yr (35.0%) and $\geq$2.0%/6yr (27.1%) threshold risk.

4.2.1 Characteristics of PLCO smokers associated with a Lung Cancer Diagnosis

Overall, 1.8% of the study population developed LC throughout 6 years, with a higher proportion of diagnoses made among males (1.9% vs. 1.6% of females) (Table 2, column 3). There were significant differences between participants with LC and without LC for all PLCOM$^{2012}$ predictor variables (p<0.001). The average age of LC diagnosis was 64.6 years,
compared to 62.3 years for those without a LC diagnosis. Compared to individuals without LC, individuals with LC, on average, had a lower BMI (26.4 kg/m² vs. 27.4 kg/m²). The following characteristics were associated with a LC diagnosis: African American ethnicity (2.8% vs. 1.7% of White/other non-Black ethnicity), having a high-school diploma or a lower level of education (2.4% vs. 1.5% with post high-school training), having a personal history of cancer (2.6% vs. 1.7% with no history of cancer), family history of LC (3.0% vs. 1.6% with no family history of LC), history of COPD (4.0% vs. 1.5% with no history of COPD), and being a current-smoker versus a former-smoker (3.9% vs. 1.2%). Compared to those without a LC diagnosis, individuals with LC smoked more cigarettes per day (29.9 vs. 24.6), for a longer duration (40.1 vs. 27.3 years), and in former-smokers, had a shorter time since quitting (7.0 vs. 16.4 years). Compared to the NLST criteria, a greater proportion of PLCO smokers eligible by the PLCOM2012 risk prediction model, at ≥1.5% and ≥2.0%/6yr risk, developed LC during 6-year follow-up (3.4% for NLST vs. 4.1% and 4.7% for the model at ≥1.5% and ≥2.0%/6yr risk, respectively).

4.2.2 Impact of Participant Characteristics on Lung Cancer Risk in Univariate Analysis

Crude univariate odds ratios are presented for each variable (Table 2, last column), estimating the odds of developing LC during 6-year follow-up for those with a specific risk factor as compared to those without that risk factor. Logistic regression analysis identified a significant univariate association between all variables and having a LC diagnosis. Overall, the odds of developing LC were greatest among participants eligible by the PLCOM2012 risk prediction model at ≥1.5%/6yr risk (OR = 8.15, 95% CI 7.08-9.38), followed by the model at ≥2.0%/6yr risk (OR = 7.14, 95% CI 6.31-8.07) and both were significant (p<0.001). In descending order of magnitude, dichotomous categorical predictor variables with the strongest risk of LC were current-smoking status, previous history of COPD, family history of LC, African
American ethnicity, an education less than or equal to high-school and personal history of cancer. A negative association was seen with smoking quit-time for former-smokers and with BMI, representing a decrease in LC risk for every one-year and kg/m$^2$ increase for smoking quit-time and BMI respectively.
Table 2. Characteristics of the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207) and univariate logistic regression odds ratios predicting lung cancer.

<table>
<thead>
<tr>
<th>Variables*</th>
<th>Participants without LC† (n=72 900)</th>
<th>Participants with LC† (n=1 307)</th>
<th>Total† (N=74 207)</th>
<th>P-value</th>
<th>Crude Univariate Odds Ratio§ (95% CI, p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sociodemographic</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age, mean, median (IQR), years</td>
<td>62.3, 62 (58-66)</td>
<td>64.6, 65 (60-69)</td>
<td>62.4, 62 (58-66)</td>
<td>&lt;0.001‡</td>
<td>1.08 per-yr (1.07-1.09, &lt;0.001)</td>
</tr>
<tr>
<td>Sex (n [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>42 005 [98.1%]</td>
<td>809 [1.9%]</td>
<td>42 814 [57.7%]</td>
<td>0.002§</td>
<td>1.19 Male (1.07-1.34, 0.002)</td>
</tr>
<tr>
<td>Female</td>
<td>30 895 [98.4%]</td>
<td>498 [1.6%]</td>
<td>31 393 [42.3%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race/Ethnicity (n [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>3 867 [97.2%]</td>
<td>111 [2.8%]</td>
<td>3 978 [5.4%]</td>
<td>&lt;0.001§</td>
<td>1.66 Black (1.35-2.03, &lt;0.001)</td>
</tr>
<tr>
<td>White/Other</td>
<td>69 033 [98.3%]</td>
<td>1 196 [1.7%]</td>
<td>70 229 [94.6%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education (n [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-school graduate or less</td>
<td>21 966 [97.6%]</td>
<td>540 [2.4%]</td>
<td>22 506 [30.3%]</td>
<td>&lt;0.001§</td>
<td>1.63 ≤High-school (1.46-1.83, &lt;0.001)</td>
</tr>
<tr>
<td>Post high-school</td>
<td>50 934 [98.5%]</td>
<td>767 [1.5%]</td>
<td>51 701 [69.7%]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Medical history</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body Mass Index, mean (SD, IQR), kg/m²</td>
<td>27.4 (4.9, 24.2-29.9)</td>
<td>26.4 (4.4, 23.4-28.8)</td>
<td>27.4 (4.9, 24.2-29.9)</td>
<td>&lt;0.001‡</td>
<td>0.95 per-kg/m² (0.94-0.97, &lt;0.001)</td>
</tr>
<tr>
<td>Personal history of Cancer (n [%])</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>3 359 [97.4%]</td>
<td>90 [2.6%]</td>
<td>3 449 [4.7%]</td>
<td>&lt;0.001§</td>
<td>1.53 (1.22-1.90, &lt;0.001)</td>
</tr>
<tr>
<td>No</td>
<td>69 541 [98.3%]</td>
<td>1 217 [1.7%]</td>
<td>70 758 [95.4%]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued on the following page)
### Family history of LC (n [%])

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8 567 [97.0%]</td>
<td>262 [3.0%]</td>
</tr>
<tr>
<td></td>
<td>8 829 [11.9%]</td>
<td>65 378 [88.1%]</td>
</tr>
</tbody>
</table>

- **History of COPD (n [%])**
  
<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 362 [96.0%]</td>
<td>268 [4.0%]</td>
</tr>
<tr>
<td></td>
<td>6 630 [8.9%]</td>
<td>67 577 [91.1%]</td>
</tr>
</tbody>
</table>

### Exposure History

#### Smoking status (n [%])

<table>
<thead>
<tr>
<th></th>
<th>Current</th>
<th>Former</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14 190 [96.1%]</td>
<td>576 [3.9%]</td>
</tr>
<tr>
<td></td>
<td>14 766 [19.9%]</td>
<td>731 [1.2%]</td>
</tr>
</tbody>
</table>

- **Smoking intensity, mean, median (IQR), cigarettes/day**
  
<table>
<thead>
<tr>
<th></th>
<th>24.6, 20 (10-30)</th>
<th>29.9, 30 (20-40)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.7, 20 (10-30)</td>
<td>27.3, 28 (16-39)</td>
</tr>
</tbody>
</table>

#### Smoking duration, mean, median (IQR), years

<table>
<thead>
<tr>
<th></th>
<th>27.3, 28 (16-39)</th>
<th>40.1, 42 (35-47)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27.6, 28 (19-39)</td>
<td>7.0, 1 (0-11)</td>
</tr>
</tbody>
</table>

#### Smoking quit-time in former smokers, mean, median (IQR), years

<table>
<thead>
<tr>
<th></th>
<th>16.4, 15 (3-28)</th>
<th>7.0, 1 (0-11)</th>
</tr>
</thead>
</table>

### Eligibility

- **NLST eligible (n [%])**
  
<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>27 438 [96.6%]</td>
<td>958 [3.4%]</td>
</tr>
<tr>
<td></td>
<td>28 396 [38.3%]</td>
<td>45 811 [61.7%]</td>
</tr>
</tbody>
</table>

<0.001§ 1.88 (1.64-2.16, <0.001)  
<0.001§ 2.70 (2.35-3.10, <0.001)  
<0.001§ 3.26 Current (2.91-3.64, <0.001)  
<0.001‡ 1.02 per-cig/day (1.017- 1.023, <0.001)  
<0.001‡ 1.09 per-yr. (1.08-1.09, <0.001)  
<0.001‡ 0.93 per-yr. (0.93-0.94, <0.001)  
<0.001§ 4.55 (4.01-5.16, <0.001)

(Continued on the following page)
<table>
<thead>
<tr>
<th>PLCOM2012 eligible at ≥1.5%/6yr (n [%])</th>
<th></th>
<th>&lt;0.001§</th>
<th>8.15 (7.08-9.38, &lt;0.001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>24 906 [95.9%]</td>
<td>1 057 [4.1%]</td>
<td>25 963 [35.0%]</td>
</tr>
<tr>
<td>No</td>
<td>47 994 [99.5%]</td>
<td>250 [0.5%]</td>
<td>48 244 [65.0%]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PLCOM2012 eligible at ≥2.0%/6yr (n [%])</th>
<th></th>
<th>&lt;0.001§</th>
<th>7.14 (6.31-8.07, &lt;0.001)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>19 147 [95.3%]</td>
<td>938 [4.7%]</td>
<td>20 085 [27.1%]</td>
</tr>
<tr>
<td>No</td>
<td>53 753 [99.3%]</td>
<td>369 [0.7%]</td>
<td>54 122 [72.9%]</td>
</tr>
</tbody>
</table>

**Abbreviations:** IQR, Interquartile range; LC, Lung Cancer; COPD, Chronic Obstructive Pulmonary Disease.

*The mean, standard deviation (for normally distributed), median (for non-normally distributed) and IQR (25th-75th percentile) are shown for continuous variables and the proportion, and row and column percentages are shown for categorical variables.

†Row percentages are provided in brackets for categorical variables for participants without lung cancer (second column) and with lung cancer (third column), i.e. the percentages of participants for that row (with or without a specific risk factor) who did not develop lung cancer and who did during follow-up; Column percentages provided in brackets for categorical variables for total participants (fourth column), i.e. the percentage of participants for that column (N=74 207) with specific characteristics.

‡P-value from nonparametric test of trend across ordered groups.

§P-value from Fishers exact test.

¶ Unadjusted odds ratios calculated for dichotomous variables (yes vs. no) and represent the odds of developing lung cancer for those with the risk factor (classified as “yes”) versus those without the risk factor; odds ratios calculated using univariate logistic regression for continuous variables and represent the odds of developing lung cancer for one-unit increase of risk-factor.
4.3 Study Objective 1: NLST versus PLCO\textsubscript{M2012} Model Risk $\geq 1.5\%$ & $\geq 2.0\%/6\text{yr}$ Criteria for Selecting LC Screenees

Results for study objective 1 are presented in Tables 3 and 4. Compared to the NLST criteria, the PLCO\textsubscript{M2012} model at $\geq 1.5\%$ and $\geq 2.0\%/6\text{yr}$ risk selected fewer individuals for LC screening (35.0% and 27.1%, respectively, vs. 38.3%). However, a greater proportion of LC cases were detected by the PLCO\textsubscript{M2012} model at $\geq 1.5\%/6\text{yr}$ risk (80.9% vs. 73.3%, $p<0.001$, Table 3a), while fewer LC cases were detected by the model at $\geq 2.0\%/6\text{yr}$ risk (71.8% vs. 73.3%, Table 3b), but this was not significant ($p = 0.05$). The overall 6-year LC incidence probability was also higher in participants eligible for screening by the model at $\geq 1.5\%/6\text{yr}$ risk (4.1%) and at $\geq 2.0\%/6\text{yr}$ risk (4.7%), compared to the NLST criteria (3.4%).

When the NLST and PLCO\textsubscript{M2012} ($\geq 1.5\%/6\text{yr}$ risk) criteria were applied to PLCO smokers (N=74 207), 4 929 participants (6.6%) were classified as positive (selected for LC screening) by the PLCO\textsubscript{M2012} model and not the NLST criteria (PLCO\textsubscript{M2012}+/NLST criteria−) and 7 362 participants (9.9%) were selected for screening by the NLST criteria and not the PLCO\textsubscript{M2012} model (NLST criteria+/PLCO\textsubscript{M2012}−) (Table 3a, cells b and c). These discordant cells provide a meaningful comparison of the two selection methods. Although 33% more individuals were selected for screening in the NLST criteria+/PLCO\textsubscript{M2012}− group, compared to the PLCO\textsubscript{M2012}+/NLST criteria− group, the probability of developing LC in 6 years was about 4 times higher (0.8% vs. 3.2%) in the latter. Therefore, individuals selected for LC screening by the NLST criteria, whom were missed by PLCO\textsubscript{M2012} model, were at low-risk for developing LC and the PLCO\textsubscript{M2012} model was effective at selecting less individuals for LC screening while significantly identifying more LC cases.
Observing the model at ≥2.0%/6yr risk (Table 3b), the PLCOM2012+/NLST criteria– group had a greater 6-year LC incidence risk than those eligible by the NLST criteria only (3.9% vs. 1.2%), but due to significantly fewer individuals in this group, there was a decrease in LC cases diagnosed within 6-year time as compared to the NLST criteria+/PLCOM2012– group (8.6% vs. 10.2%), and this was not a significant decrease (p = 0.202).

Increasing the PLCOM2012 model risk threshold from ≥1.5% to ≥2.0%/6yr resulted in 22.6% fewer individuals being selected for LC screening (25 963 vs. 20 085), 11.3% fewer LC cases detected (1 057 vs. 938) and an overall 1.15-fold (95% CI 1.05-1.25, p = 0.002) increase in 6-year LC incidence risk (4.7% vs. 4.1%). Compared to individuals in the PLCOM2012+ (≥1.5%/6yr risk)/NLST criteria– group those in the PLCOM2012+ (≥2.0%/6yr risk)/NLST criteria– group had a 1.24 times greater risk of developing LC (3.9% vs. 3.2%).

Table 3a. Proportion of participants eligible, 6-year lung cancer risk and lung cancer cases detected by NLST criteria and PLCOM2012 risk ≥1.5%/6yr criterion status in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207).

<table>
<thead>
<tr>
<th>Table 3a</th>
<th>NLST criteria +</th>
<th>NLST criteria -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PLCOM2012 risk ≥1.5% +</strong></td>
<td>28 396 (row % = 38.3%)</td>
<td>45 811 (row % = 61.7%)</td>
<td>74 207 (cell % = 100%)</td>
</tr>
<tr>
<td>Proportion eligible, n (%)</td>
<td>21 034 (cell % = 28.3%)</td>
<td>4 929 (cell % = 6.6%)</td>
<td>25 963 (column % = 35.0%)</td>
</tr>
<tr>
<td>6-yr LC incidence risk</td>
<td>4.3%</td>
<td>3.2%</td>
<td>4.1%</td>
</tr>
<tr>
<td>No. of LC cases</td>
<td>900</td>
<td>157</td>
<td>1 057</td>
</tr>
<tr>
<td><strong>PLCOM2012 risk ≥1.5% -</strong></td>
<td>7 362 (cell % = 9.9%)</td>
<td>40 882 (cell % = 55.1%)</td>
<td>48 244 (column % = 65.0%)</td>
</tr>
<tr>
<td>Proportion eligible, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-yr LC incidence risk</td>
<td>0.8%</td>
<td>0.5%</td>
<td>0.5%</td>
</tr>
<tr>
<td>No. of LC cases</td>
<td>58</td>
<td>192</td>
<td>250</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>28 396 (row % = 38.3%)</td>
<td>45 811 (row % = 61.7%)</td>
<td></td>
</tr>
<tr>
<td>Proportion eligible, n (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-yr LC incidence risk</td>
<td>3.4%</td>
<td>0.8%</td>
<td>1.8%</td>
</tr>
<tr>
<td>No. of LC cases</td>
<td>958</td>
<td>349</td>
<td>1 307</td>
</tr>
</tbody>
</table>

Abbreviations: LC, Lung cancer. Bold indicates informative cells in which disagreement exists between the two classification criteria. *Row percentages represent the percentage of smokers whom were eligible or not eligible for lung cancer screening by the NLST criteria; Column percentages represent the percentage of smokers whom were eligible or not eligible for lung cancer screening by the PLCOM2012 risk prediction model; Cell percentages represent the percentage of smokers eligible for lung cancer screening by both methods (a), eligible by the PLCOM2012 model and not the NLST criteria (b), eligible by the NLST criteria and not the PLCOM2012 model (c), and not eligible by both methods (d).
Table 3b. Proportion of participants eligible, 6-year lung cancer risk and lung cancer cases detected by NLST criteria and PLCO_{M2012} risk ≥2.0%/6yr criterion in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207).

<table>
<thead>
<tr>
<th>Table 3b</th>
<th>NLST criteria +</th>
<th>NLST criteria -</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLCO_{M2012} risk ≥2.0% + Proportion eligible, n (%)</td>
<td>17 213 (cell % = 23.3%)</td>
<td>4.8%</td>
<td>825</td>
</tr>
<tr>
<td>6-yr LC incidence risk No. of lung cancers</td>
<td>2 872 (cell % = 3.9%)</td>
<td>3.9%</td>
<td>113</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>20 085 (column % = 27.1%)</td>
<td>4.7%</td>
<td>938</td>
</tr>
<tr>
<td>PLCO_{M2012} risk ≥2.0% - Proportion eligible, n (%)</td>
<td>11 183 (cell % = 15.1%)</td>
<td>1.2%</td>
<td>133</td>
</tr>
<tr>
<td>6-yr LC incidence risk No. of lung cancers</td>
<td>42 939 (cell % = 57.9%)</td>
<td>0.6%</td>
<td>236</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>54 122 (column % = 72.9%)</td>
<td>0.7%</td>
<td>369</td>
</tr>
<tr>
<td>Total</td>
<td>28 396 (row % = 38.3%)</td>
<td>3.4%</td>
<td>958</td>
</tr>
<tr>
<td>6-yr LC incidence risk No. of lung cancers</td>
<td>45 811 (row % = 61.7%)</td>
<td>0.8%</td>
<td>349</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>74 207 (cell % = 100%)</td>
<td>1.8%</td>
<td>1 307</td>
</tr>
</tbody>
</table>

*Row percentages represent the percentage of smokers whom were eligible or not eligible for lung cancer screening by the NLST criteria; Column percentages represent the percentage of smokers whom were eligible or not eligible for lung cancer screening by the PLCO_{M2012} risk prediction model; Cell percentages represent the percentage of smokers eligible for lung cancer screening by both methods (a), eligible by the PLCO_{M2012} model and not the NLST criteria (b), eligible by the NLST criteria and not the PLCO_{M2012} model (c), and not eligible by both methods (d).

Abbreviations: LC, Lung cancer.
Bold indicates informative cells in which disagreement exists between the two classification criteria.
Sensitivity was greatest for the model at ≥1.5%/6yr risk (80.9% and 19.1%, respectively) as compared to the NLST criteria (sensitivity = 73.3%) and these differences were significant (p<0.001) (Table 4). Therefore, when compared to the NLST criteria, the model at ≥1.5%/6yr risk was more accurate in correctly identifying LC cases while missing the least number of LC cases. Specificity was highest for the model at ≥2.0%/6yr risk as compared to the NLST criteria (73.7% vs. 62.4%, p<0.001), demonstrating that the model was accurate in identifying participants without LC as being low-risk, resulting in fewer false positives. PPV was highest for the PLCO_{M2012} model at ≥2.0%/6yr risk (4.7%) versus the model at ≥1.5%/6yr risk (4.1%), and the NLST criteria (3.4%), and these differences were significant (p<0.001). Therefore, using a higher model risk threshold resulted in selecting individuals for screening who had the greatest
likelihood of developing LC within 6 years. Overall, the PLCOM2012 model at a risk threshold of 1.5%/6yr performed better than the NLST criteria in all accuracy measures.

Table 4. Accuracy of lung cancer classification by each criterion in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=72 407).

<table>
<thead>
<tr>
<th></th>
<th>NLST criteria</th>
<th>PLCOM2012 ≥1.5% risk</th>
<th>PLCOM2012 ≥2.0% risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>73.3%</td>
<td>80.9%</td>
<td>71.8%</td>
</tr>
<tr>
<td>Specificity</td>
<td>62.4%</td>
<td>65.8%</td>
<td>73.7%</td>
</tr>
<tr>
<td>PPV</td>
<td>3.4%</td>
<td>4.1%</td>
<td>4.7%</td>
</tr>
<tr>
<td>False Positive rate (%)</td>
<td>37.6%</td>
<td>34.2%</td>
<td>26.3%</td>
</tr>
<tr>
<td>False Negative rate (%)</td>
<td>26.7%</td>
<td>19.1%</td>
<td>28.2%</td>
</tr>
</tbody>
</table>

Abbreviations: PPV, Positive predictive value.

4.4 Study Objective 2: PLCOM2012+/NLST criteria− vs. NLST criteria+/PLCOM2012 – for selecting LC Screenees by PLCOM2012 Model Predictors

For study objective 2, predictor variables were split up into their high-risk strata which includes participants who were ≥62 years old, from African American descent, had a high-school degree or a lesser education, were underweight/normal, had a previous history of cancer, family history of LC, history of COPD, were current-smokers, smoked ≥20 cigarettes per day, for a duration of ≥28 years, and if former-smokers, had quit smoking ≤15 years.

Tables 5a and 5b demonstrate the differences in PLCOM2012 model predictors between the discordant cells, PLCOM2012+/NLST criteria− and NLST criteria+/PLCOM2012 – at ≥1.5% and at ≥2.0%/6yr PLCOM2012 model risk, respectively. For each predictor variable, the number of participants eligible by each criterion (n), number of LC cases, and the risk of developing LC in 6 years are presented. This provides information on which components of the model accurately identify high-risk individuals while detecting more LC cases.

Compared to the NLST criteria+/PLCOM2012 – group (at both risk thresholds), individuals in the PLCOM2012+/NLST criteria− group were, on average, older, and a greater proportion were
African Americans. Compared to those selected exclusively by the NLST criteria, a greater proportion of participants in the PLCOM2012+ (≥1.5%/6yr risk)/NLST criteria– group had a high-school level of education or less (46.2% vs. 20.6%), and a lower average BMI (26.4 vs. 29.4 kg/m²) (Table 5a). When looking at the PLCOM2012 model at ≥2.0%/6yr risk (Table 5b), there were fewer individuals with an education less than or equal to high-school and with a BMI ≤25 kg/m² selected for screening by the model, as compared to the NLST criteria. However, relative to the total number of individuals eligible for screening by each criterion, a higher proportion of these individuals were selected for screening in the PLCOM2012+/NLST criteria– group.

A greater percentage of individuals in the PLCOM2012+/NLST criteria– group, for both model thresholds, had a history of cancer, family history of LC and history of COPD, when compared to selection by the NLST criteria. A greater proportion of PLCOM2012+/NLST criteria– eligible individuals were current smokers, who, on average, smoked fewer cigarettes per day, for a longer duration of time, and, if were former-smokers, had a greater quit-time.

These results demonstrate the effectiveness of the PLCOM2012 model, over the NLST criteria, in selecting a greater proportion of high-risk individuals for each risk factor, except for smoking intensity, duration and quit-time; although the PLCOM2012 model selected less individuals in the high-risk strata for these variables, individuals selected by the model had a higher 6-year LC risk compared to the NLST criteria, and this will be explained below. Screening only individuals at high-risk translates into improved efficiency and cost-effectiveness of LC screening, which is discussed in the next chapter.
4.4.1 Differences in 6-year Lung Cancer Incidence Risk in Discordant groups by Risk Predictor Status

The PLCO\textsubscript{M2012} model at $\geq1.5\%/6\text{yr}$ risk detected a greater number of LC cases among individuals in the high-risk strata for each predictor variable compared to the NLST criteria+/PLCO\textsubscript{M2012} – group (Table 5a). A greater number of LC cases were also detected in those eligible exclusively by the model at $\geq2.0\%/6\text{yr}$ risk (Table 5b), compared to the NLST criteria+/PLCO\textsubscript{M2012} – group for all predictor variables, except for smoking intensity, duration and quit-time. However, when considering the total number of individuals eligible for screening in the high-risk strata for each predictor variable (Table 5b, n), the LC incidence proportion, also known as the 6-year LC risk, was greater in the PLCO\textsubscript{M2012}+ ($\geq2.0\%/6\text{yr}$ risk)/NLST criteria– group as compared to the NLST criteria+/PLCO\textsubscript{M2012} – group. The LC incidence proportion is estimated by dividing the number of LC cases developed in the high-risk strata by the total number of individuals (n) eligible for screening in that strata. For example, in Table 5b, when observing individuals $\geq62$ years old in the PLCO\textsubscript{M2012}+/NLST criteria– group, 4.0% developed LC (110 LC cases/2766 individuals in high-risk strata), as compared to 1.4% among individuals in the high-risk age group in the NLST criteria+/PLCO\textsubscript{M2012} – group.

Increasing the PLCO\textsubscript{M2012} model threshold from 1.5% to 2.0%/6yr also resulted in a greater 6-year LC-incidence risk for all predictor variables; the greatest risk increase was seen in current-smokers, where current-smokers eligible by the PLCO\textsubscript{M2012} at $\geq2.0\%/6\text{yr}$ risk had a 1.2 times increased risk of developing LC in 6 years, compared to those eligible by the PLCO\textsubscript{M2012} at $\geq1.5\%/6\text{yr}$ risk (5.1% vs 4.1%).
4.4.2 In the High-Risk Strata of each Model Risk Predictor: What are the Odds of 1) Being selected for Screening and 2) Being Diagnosed with Lung Cancer by the PLCOM2012 Model vs. NLST Criteria?

When observing the model at ≥1.5%/6yr risk (Table 5a), all PLCOM2012 risk factors were associated with an increase in odds of being eligible for LC screening by the PLCOM2012+/NLST criteria– versus the NLST criteria+/PLCOM2012 – group, except for smoking intensity, smoking duration and smoking quit-time where negative associations were seen, and all associations were significant (p<0.001). A similar pattern was seen for the model at ≥2.0%/6yr risk and in addition to the smoking variables, negative associations were also seen for high-school graduates (or those with a lesser education) and in individuals with a BMI ≤25 kg/m². However, these crude OR values are based on the absolute number of individuals selected for screening; when considering the sample size in the discordant groups, a greater proportion of these high-risk individuals in the PLCOM2012+/NLST criteria– group were selected for LC screening as compared to the NLST criteria.

The odds of being eligible for LC screening by the model at both 1.5% and 2.0%/6yr risk thresholds, versus the NLST criteria, was highest for individuals with a personal history of cancer, followed by those with a family history of LC, African Americans, individuals ≥62 years of age, with a history of COPD, high-school graduates or individuals with a lesser education, those with a BMI ≤25 kg/m², current smokers, individuals who smoked for ≥28 years, and smoked 20 or more cigarettes a day. The odds of being eligible for LC screening was the smallest for former-smokers with a smoking quit-time less than or equal to 15 years meaning that these individuals were less likely to be selected for LC screening in the PLCOM2012+/NLST criteria– group as compared to the NLST criteria+/PLCOM2012– group (at ≥1.5% and ≥2.0%/6yr
risk). When the high-risk strata of each model risk predictor were considered separately, the odds of detecting LC in individuals in the PLCO_{M2012}+ (≥1.5%/6yr risk)/NLST criteria– group was greater for all risk-factors, compared to the NLST criteria+/PLCO_{M2012} – group (≥1.5%/6yr risk) (Table 5a, last column). In descending order of magnitude, an increase in odds of developing LC in the PLCO_{M2012}+ (≥1.5%/6yr risk)/NLST criteria– group was also seen in individuals with a family history of LC, African Americans, individuals with a history of COPD, individuals ≥62 years old, current-smokers, individuals with a BMI ≤25 kg/m², high-school graduates or those with a lesser education, individuals who smoked for ≥28 years, and smoked ≥20 cigarettes per day, and in former-smokers with a smoking quit-time of ≤15 years.

Observing the model at ≥2.0%/6yr risk, the OR for identifying LC among individuals in the PLCO_{M2012}+/NLST criteria– group when compared to the NLST criteria+/PLCO_{M2012} – group (≥2.0%/6yr risk) was greater for all risk-factors, except for smoking intensity, duration, and quit-time (Table 5b, last column). In the high-risk predictor strata, in descending order of magnitude, an increase in odds of developing LC in the PLCO_{M2012}+ (≥2.0%/6yr risk)/NLST criteria– group was seen in individuals with a family history of LC, followed by African Americans, individuals ≥62 years old, with a history of COPD, a BMI ≤25 kg/m², current-smokers, and in high-school graduates or those with a lesser education. Decreased odds ratios for developing LC were seen for individuals who smoked for ≥28 years, and smoked ≥20 cigarettes/day, and in former-smokers with a smoking quit-time of ≤15 years. However, the 6-year risk of LC incidence, which is a more valid measure of efficiency, was greater for all high-risk individuals in the PLCO_{M2012}+ (≥2.0%/6yr risk)/NLST criteria– group.
<table>
<thead>
<tr>
<th>Predictors</th>
<th>PLCO\textsubscript{M2012}+/NLST criteria\textsuperscript{−} (n=4 929)</th>
<th>NLST criteria+/\textsuperscript{−} PLCO\textsubscript{M2012} (n=7 362)</th>
<th>Difference in LC risks (p-value) \textsuperscript{‡}</th>
<th>Crude Odds Ratio\textsuperscript{§} (95% CI, p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD), years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥62 years (n [%], no. LC cases, 6-yr LC risk)</td>
<td>68.3 (4.0)</td>
<td>58.3 (3.2)</td>
<td>&lt;0.001</td>
<td>3.74 Eligible, ≥62 yr. (3.50-3.99, &lt;0.001) \textsuperscript{(1)}</td>
</tr>
<tr>
<td>&lt;62 years (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(4 592 [93.2%], 149, 3.2%)</td>
<td>(1 228 [16.8%], 13, 1.1%)</td>
<td></td>
<td>11.46 LC, ≥62 yr. (6.50-22.03, &lt;0.001) \textsuperscript{(2)}</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(618 [12.5%], 21, 3.4%)</td>
<td>(141 [1.9%], 1, 0.7%)</td>
<td>0.099</td>
<td>4.38 Eligible, Black (3.64-5.30, &lt;0.001) \textsuperscript{(1)}</td>
</tr>
<tr>
<td>White/Other (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(4 311 [87.5%], 136, 3.1%)</td>
<td>(7 221 [98.1%], 57, 0.8%)</td>
<td></td>
<td>21.00 LC, Black (3.38-868.45, &lt;0.001) \textsuperscript{(2)}</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-school graduate or less (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(2 276 [46.2%], 73, 3.2%)</td>
<td>(1 519 [20.6%], 17, 1.1%)</td>
<td>&lt;0.001</td>
<td>1.50 Eligible, ≤High-school (1.40-1.60, &lt;0.001) \textsuperscript{(1)}</td>
</tr>
<tr>
<td>Post high-school (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(2 653 [53.8%], 84, 3.2%)</td>
<td>(5 843 [79.4%], 41, 0.7%)</td>
<td></td>
<td>4.29 LC, ≤High-school (2.51-7.77, &lt;0.001) \textsuperscript{(2)}</td>
</tr>
<tr>
<td>Body mass index, mean (SD), kg/m\textsuperscript{2}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underweight/Normal (≤25 kg/m\textsuperscript{2}) (n [%], no. LC cases, 6-yr LC risk)</td>
<td>26.4 (4.3)</td>
<td>29.4 (5.4)</td>
<td>&lt;0.001</td>
<td>1.36 Eligible, ≤25kg/m\textsuperscript{2} (1.27-1.46, &lt;0.001) \textsuperscript{(1)}</td>
</tr>
<tr>
<td>Overweight/Obese (&gt;25 kg/m\textsuperscript{2}) (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1 889 [38.3%], 76, 4.0%)</td>
<td>(1 388 [18.9%], 11, 0.8%)</td>
<td></td>
<td>6.91 LC, &gt;25kg/m\textsuperscript{2} (3.65-14.42, &lt;0.001) \textsuperscript{(2)}</td>
</tr>
<tr>
<td>Personal history of Cancer\textsuperscript{*}</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(542 [10.6%], 15, 2.8%)</td>
<td>(80 [1.2%], 0, 0%)</td>
<td>0.238</td>
<td>6.78 Eligible (5.35-8.68, &lt;0.001) \textsuperscript{(1)}</td>
</tr>
<tr>
<td>No (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(4 387 [89.0%], 142, 3.2%)</td>
<td>(7 282 [98.9%], 58, 0.8%)</td>
<td></td>
<td>15.0 LC (2.31-631.47, &lt;0.001) \textsuperscript{(2)}</td>
</tr>
<tr>
<td>Family history of Lung Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1 329 [27.0%], 46, 3.5%)</td>
<td>(236 [3.2%], 2, 0.9%)</td>
<td>0.037</td>
<td>5.63 Eligible (4.90-6.49, &lt;0.001) \textsuperscript{(1)}</td>
</tr>
<tr>
<td>No (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(3 600 [73.0%], 111, 3.1%)</td>
<td>(7 126 [96.8%], 56, 0.8%)</td>
<td></td>
<td>23.00 LC (6.02-195.60, &lt;0.001) \textsuperscript{(2)}</td>
</tr>
</tbody>
</table>
### History of COPD

<table>
<thead>
<tr>
<th></th>
<th>Yes (n [%], no. LC cases, 6-yr LC risk)</th>
<th>No (n [%], no. LC cases, 6-yr LC risk)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>(800 [16.2%], 38, 4.8%)</td>
<td>(7103 [96.5%], 56, 0.8%)</td>
<td>0.002</td>
</tr>
</tbody>
</table>

### Smoking status

<table>
<thead>
<tr>
<th></th>
<th>Current (n [%], no. LC cases, 6-yr LC risk)</th>
<th>Former (n [%], no. LC cases, 6-yr LC risk)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1231 [25.0%], 51, 4.1%)</td>
<td>(6419 [87.2%], 52, 0.8%)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

### Smoking intensity*, mean (SD)

<table>
<thead>
<tr>
<th></th>
<th>≥20 cigarettes/day (n [%], no. LC cases, 6-yr LC risk)</th>
<th>&lt;20 cigarettes/day (n [%], no. LC cases, 6-yr LC risk)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>(3341 [67.8%], 92, 2.8%)</td>
<td>(0, -----, -----)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

### Smoking duration, mean (SD), years

<table>
<thead>
<tr>
<th></th>
<th>≥28 years (n [%], no. LC cases, 6-yr LC risk)</th>
<th>&lt;28 years (n [%], no. LC cases, 6-yr LC risk)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>(3778 [76.6%], 129, 3.4%)</td>
<td>(1045 [14.2%], 7, 0.7%)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

### Smoking quit-time, mean (SD), years

<table>
<thead>
<tr>
<th></th>
<th>≤15 years (n [%], no. LC cases, 6-yr LC risk)</th>
<th>&gt;15 years (n [%], no. LC cases, 6-yr LC risk)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1848 [37.5%], 68, 3.7%)</td>
<td>(0, -----, -----)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Abbreviations**: LC, Lung Cancer; n, number of people eligible; no., number of lung cancer cases; SD, standard deviation; COPD, Chronic Obstructive Pulmonary Disease.

*Personal History of Lung Cancer; 1 lung cancer added to zero cell in NLST criteria+/PLCO\textsubscript{M2012}− group for odds ratio calculation of identifying lung cancer. Smoking intensity measured as the average number of cigarettes smoked per day.
†PLCO\textsubscript{M2012}+/NLST criteria−, eligible by the model (≥1.5%/6yr risk) but not the NLST criteria; NLST criteria+/PLCO\textsubscript{M2012}−, eligible by the NLST criteria but not the model (≥1.5%/6yr risk).
‡Comparing incidence of lung cancer in the high-risk strata between the two criteria: PLCO\textsubscript{M2012}+ versus NLST+; P-value from Fishers exact test.
§Two odds ratios are presented; 1) For those in the high-risk strata of the predictor, the odds of being eligible for screening by the PLCO\textsubscript{M2012} model (≥1.5%/6yr risk) versus NLST criteria in those with the given risk factor, 2) For those in the high-risk strata of the predictor, the odds of developing lung cancer among those eligible and diagnosed with lung cancer by the PLCO\textsubscript{M2012} model (≥1.5%/6yr risk) versus NLST criteria with the given risk factor; P-value by McNemar’s test.
Table 5b. Distribution of Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207) in discordant groups by PLCO<sub>M2012</sub> risk factors at ≥2.0%/6yr PLCO<sub>M2012</sub> model risk.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>PLCO&lt;sub&gt;M2012&lt;/sub&gt;/NLST criteria – † (n=2 872)</th>
<th>NLST criteria+/ PLCO&lt;sub&gt;M2012&lt;/sub&gt; – † (n=11 183)</th>
<th>Difference in LC risks (p-value) ‡</th>
<th>Crude Odds Ratio § (95% CI, p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD), years</td>
<td>69.1 (3.6)</td>
<td>58.8 (3.5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥62 years (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(2 766 [96.3%], 110, 4.0%)</td>
<td>(2 468 [22.1%], 35, 1.4%)</td>
<td>&lt;0.001</td>
<td>1.12 Eligible, ≥62 yr. (1.06-1.18, &lt;0.001) (1)</td>
</tr>
<tr>
<td>&lt;62 years (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(106 [3.7%], 3, 2.8%)</td>
<td>(8 715 [77.9%], 98, 1.1%)</td>
<td></td>
<td>3.14 LC, ≥62 yr. (2.13-4.74, &lt;0.001) (2)</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(414 [14.4%], 15, 3.6%)</td>
<td>(296 [2.6%], 3, 1.0%)</td>
<td>0.030</td>
<td>1.40 Eligible, Black (1.20-1.63, &lt;0.001) (1)</td>
</tr>
<tr>
<td>White/Other (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(2 458 [85.6%], 98, 4.0%)</td>
<td>(10 887 [97.3%], 130, 1.2%)</td>
<td></td>
<td>5.00 LC, Black (1.41-26.95, 0.005) (2)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-school graduate or less (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1 382 [48.1%], 50, 3.6%)</td>
<td>(2 601 [23.3%], 44, 1.7%)</td>
<td>&lt;0.001</td>
<td>0.53 Eligible, ≤High-school (0.50-0.57, &lt;0.001) (1)</td>
</tr>
<tr>
<td>Post high-school (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1 490 [51.9%], 63, 4.2%)</td>
<td>(8 582 [76.7%], 89, 1.0%)</td>
<td></td>
<td>1.14 LC, ≥High-school (0.742-1.744, 0.536) (2)</td>
</tr>
<tr>
<td>Body mass index, mean (SD), kg/m&lt;sup&gt;2&lt;/sup&gt;</td>
<td>26.2 (4.2)</td>
<td>28.9 (5.2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Underweight/Normal (&lt;25 kg/m&lt;sup&gt;2&lt;/sup&gt;) (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1 141 [39.7%], 56, 4.9%)</td>
<td>(2 431 [21.7%], 26, 1.1%)</td>
<td>&lt;0.001</td>
<td>0.47 Eligible, ≤25kg/m² (0.44-0.50, &lt;0.001) (1)</td>
</tr>
<tr>
<td>Overweight/Obese (&gt;25 kg/m&lt;sup&gt;2&lt;/sup&gt;) (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1 731 [60.3%], 57, 3.3%)</td>
<td>(8 752 [78.3%], 107, 1.2%)</td>
<td></td>
<td>2.15 LC, ≤25kg/m² (1.33-3.57, &lt;0.001) (2)</td>
</tr>
<tr>
<td>Personal history of Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(374 [13.0%], 12, 3.2%)</td>
<td>(174 [1.6%], 2, 1.1%)</td>
<td>0.244</td>
<td>2.15 Eligible (1.79-2.59, &lt;0.001) (1)</td>
</tr>
<tr>
<td>No (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(2 498 [87.0%], 101, 4.0%)</td>
<td>(11 009 [98.4%], 131, 1.2%)</td>
<td></td>
<td>6.00 LC (1.34-55.20, 0.008) (2)</td>
</tr>
<tr>
<td>Family history of Lung Cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(905 [31.5%], 36, 4.0%)</td>
<td>(538 [4.8%], 6, 1.1%)</td>
<td>0.002</td>
<td>1.68 Eligible (1.51-1.88, &lt;0.001) (1)</td>
</tr>
<tr>
<td>No (n [%], no. LC cases, 6-yr LC risk)</td>
<td>(1 967 [68.5%], 77, 3.9%)</td>
<td>(10 645 [95.2%], 127, 1.2%)</td>
<td></td>
<td>6.00 LC (2.50-17.42, &lt;0.001) (2)</td>
</tr>
</tbody>
</table>
### History of COPD

<table>
<thead>
<tr>
<th>Smoking status</th>
<th>Yes</th>
<th>No</th>
<th>p-value</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>(559 [19.5%], 31, 5.5%)</td>
<td>(554 [5.0%], 11, 2.0%)</td>
<td>0.002</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>2 313 [80.5%], 82, 3.5%</td>
<td>10 629 [95.0%], 122, 1.1%</td>
<td></td>
<td>2.82</td>
</tr>
</tbody>
</table>

### Smoking status

<table>
<thead>
<tr>
<th>Smoking status</th>
<th>Current</th>
<th>Former</th>
<th>p-value</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>(830 [28.9%], 42, 5.1%)</td>
<td>(2 247 [20.1%], 36, 1.6%)</td>
<td>&lt;0.001</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>2 042 [71.1%], 71, 3.5%</td>
<td>8 936 [79.9%], 97, 1.1%</td>
<td></td>
<td>1.17</td>
</tr>
</tbody>
</table>

### Smoking intensity*

<table>
<thead>
<tr>
<th>Smoking intensity*</th>
<th>≥20 cigarettes/day</th>
<th>&lt;20 cigarettes/day</th>
<th>p-value</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>27.5 (18.9)</td>
<td>29.5 (13.0)</td>
<td>&lt;0.001</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>1 879 [65.4%], 67, 3.6%</td>
<td>11 183 [100%], 133, 1.2%</td>
<td></td>
<td>0.50</td>
</tr>
</tbody>
</table>

### Smoking duration, mean (SD), years

<table>
<thead>
<tr>
<th>Smoking duration, mean (SD), years</th>
<th>≥28 years</th>
<th>&lt;28 years</th>
<th>p-value</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>37.0 (10.0)</td>
<td>10 033 [89.7%], 126, 1.3%</td>
<td>&lt;0.001</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>2 412 [84.0%], 98, 4.1%</td>
<td>(1 150 [10.3%], 7, 0.6%)</td>
<td></td>
<td>0.78</td>
</tr>
</tbody>
</table>

### Smoking quit-time, mean (SD), years

<table>
<thead>
<tr>
<th>Smoking quit-time, mean (SD), years</th>
<th>≤15 years</th>
<th>&gt;15 years</th>
<th>p-value</th>
<th>Relative Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n [%], no. LC cases, 6-yr LC risk)</td>
<td>13.1 (10.3)</td>
<td>11 183 [100%], 133, 1.2%</td>
<td>&lt;0.001</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>1 154 [40.2%], 50, 4.3%</td>
<td>[0, -----, -----]</td>
<td></td>
<td>0.38</td>
</tr>
</tbody>
</table>

### Abbreviations:
- LC, Lung Cancer; n, number of people eligible; no., number of lung cancer cases; SD, standard deviation; COPD, Chronic Obstructive Pulmonary Disease.
- *Smoking intensity measured as the average number of cigarettes smoked per day.
- †PLCO_M2012+/ NLST criteria, eligible by the model (≥2.0%/6yr risk) but not the NLST criteria; NLST criteria+/ PLCO_M2012 –, eligible by the NLST criteria but not the model (≥2.0%/6yr risk).
- ‡Comparing incidence of lung cancer in the high-risk strata between the two criteria: PLCO_M2012+ versus NLST+; P-value from Fishers exact test.
- §Two odds ratios are presented: 1) For those in the high-risk strata of the predictor, the odds of being eligible by the PLCO_M2012 model (≥2.0%/6yr risk) versus NLST criteria in those with the given risk factor, 2) For those in the high-risk strata of the predictor, the odds of developing lung cancer in those eligible by the PLCO_M2012 model (≥2.0%/6yr risk) versus NLST criteria and with the given risk factor; P-value by McNemar’s test.
4.5 Study Objective 3: Univariate & Multivariable Models for 5-year non-LC deaths

The following section will make comparisons between the 6-year LC incidence model and the 5-year non-LC death model. The variable competing causes of death (CCoD) represents individuals who died from non-LC causes, which included other smoking cancers, smoking diseases and any other cause of death. Of 74 207 PLCO smokers, 2 964 (4.0%) died from CCoD in 5 years. To evaluate the effect of each predictor variable on the outcome, CCoD, univariate and multivariable models were developed. Univariate models predicting 6-year LC risk are also presented for comparison. For each univariate logistic regression, the odds ratio, area under the curve (AUC), pseudo-$r^2$ and regression coefficients (for variables with a non-linear relationship with the log odds of the outcome) are provided in Table 6, as well as the odds ratios and regression coefficients for the multivariable 5-year non-LC death and 6-year LC risk models.

Overall, all PLCOM2012 model predictor variables were significantly associated with the 5-year risk of death from competing causes ($p<0.001$), except for family history of LC ($p = 0.396$), which was removed from the final multivariable model due to biological reasoning and its negligible effect on prediction, which will be explained shortly. The multivariable CCoD model showed that the risk of dying from non-LC causes within 5 years increased with increasing age, lower BMI (<15kg/m$^2$) or high BMI (>30 kg/m$^2$), and lower education level. Furthermore, having a previous history of any cancer or COPD was associated with an increased risk of death from competing causes. Being a current-smoker, and smoking a greater number of cigarettes per day, or for a longer duration and having quit smoking for a shorter period, as a former-smoker, increased one’s risk of dying from non-LC causes.

Backwards selection and biological reasoning were used to select variables for inclusion into the final 5-year non-LC death model. Backwards selection was utilized by assessing ROC-AUC values and biological reasoning was used to explore which variables appeared to have a
biologic relationship with non-LC death. This selection process started with the complete multivariable model with all predictor variables, and the ROC-AUC value for this overall model was 0.7116, demonstrating fair discrimination. When family history of LC was excluded from the model, the difference in AUC was negligible (0.7114) meaning that family history of LC did not have a large effect in the overall predictive performance of the model. The odds ratio for family history of LC approached the null (OR = 1) in multivariable and univariate analysis, suggesting no strong association between family history of LC and risk of non-LC death. Family history of LC has been shown to be associated with the risk of developing COPD, which increases one’s risk of death. However, family history of LC is a stronger predictor of LC risk and is overall weakly associated with non-LC specific mortality, and therefore was not included in the final CCoD multivariable model; the overall AUC of this model was 0.7114. After removing age, the AUC dropped to 0.6858, which was the largest drop, hence age was kept in the final model. Removing smoking intensity and smoking duration decreased the AUC value to 0.7057 and 0.7070, respectively. When BMI and history of COPD were removed from the model, the AUC dropped to 0.7077 and 0.7079, while removing race/ethnicity and cigarette smoking status dropped the AUC to 0.7082 and 0.7094. Removal of personal history of cancer and education decreased the AUC value to 0.7096 and 0.7099, respectively. These risk factors through backwards selection process demonstrated to be good predictors. The smallest decrease in AUC was seen when smoking quit-time was removed from the model (AUC = 0.7109). Smoking quit-time was kept it in the model because it has shown to have an overall effect on health benefits; the longer an individual has stopped smoking, their chance of death from other diseases, such as cardiovascular diseases, respiratory diseases, cerebrovascular disease and other cancers, decreases.
### Table 6. Univariate and multivariable logistic regression models for 6-year lung cancer incidence and 5-year risk of dying from competing causes* in Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74 207) using PLCO\textsubscript{M2012} model predictors.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>Lung Cancer Incidence</th>
<th>Competing Causes of Death</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate</td>
<td>Univariate</td>
</tr>
<tr>
<td></td>
<td>OR (95% CI, p-value)</td>
<td>OR (95% CI, p-value)</td>
</tr>
<tr>
<td>Age, per 1 year</td>
<td>1.08 (1.07-1.09, &lt;0.001)</td>
<td>1.06 (1.05-1.08, &lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.626 Pseudo-R\textsuperscript{2} = 0.0182</td>
<td>(\beta = 0.062)</td>
</tr>
<tr>
<td>Race/Ethnicity, African American vs. Other</td>
<td>1.66 (1.36-2.02, &lt;0.001)</td>
<td>1.70 (1.38-2.09, &lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.516 Pseudo-R\textsuperscript{2} = 0.0017</td>
<td>(\beta = 0.528)</td>
</tr>
<tr>
<td>Education, per 1 level increase</td>
<td>0.84 (0.81-0.87, &lt;0.001)</td>
<td>0.93 (0.90-0.97, &lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.577 Pseudo-R\textsuperscript{2} = 0.0073</td>
<td>(\beta = -0.072)</td>
</tr>
<tr>
<td>Body mass index, per kg/m\textsuperscript{2}</td>
<td>0.95 (0.94-0.97, &lt;0.001)</td>
<td>0.97 (0.95-0.98, &lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.558 Pseudo-R\textsuperscript{2} = 0.0045</td>
<td>(\beta = -0.034)</td>
</tr>
<tr>
<td>Personal history of Cancer, yes vs. no</td>
<td>1.53 (1.23-1.90, &lt;0.001)</td>
<td>1.35 (1.08-1.68, 0.008)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.511 Pseudo-R\textsuperscript{2} = 0.0010</td>
<td>(\beta = 0.298)</td>
</tr>
<tr>
<td>Family history of LC* yes vs. no</td>
<td>1.88 (1.64-2.16, &lt;0.001)</td>
<td>1.73 (1.50-1.99, &lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.541 Pseudo-R\textsuperscript{2} = 0.0055</td>
<td>(\beta = 0.546)</td>
</tr>
<tr>
<td>History of COPD*, yes vs. no</td>
<td>2.67 (2.35-3.09, &lt;0.001)</td>
<td>1.45 (1.26-1.67, &lt;0.001)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.559 Pseudo-R\textsuperscript{2} = 0.0127</td>
<td>(\beta = 0.370)</td>
</tr>
<tr>
<td>Smoking status, current vs. former</td>
<td>3.26 (2.92-3.64, &lt;0.001)</td>
<td>1.27 (1.10-1.48, 0.001)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.623 Pseudo-R\textsuperscript{2} = 0.0304</td>
<td>(\beta = 0.242)</td>
</tr>
<tr>
<td>Smoking intensity§</td>
<td>(\beta = -2.44 \text{ (p&lt;0.001)})</td>
<td>(\beta = -2.25 \text{ (p&lt;0.001)})</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.617 Pseudo-R\textsuperscript{2} = 0.0184</td>
<td>(\beta = 0.400 \text{ (p&lt;0.001)})</td>
</tr>
</tbody>
</table>

\(\text{AUC}\): Area Under the Curve, \(\text{Pseudo-R}\textsuperscript{2}\): Pseudo-R\textsuperscript{2}.
<table>
<thead>
<tr>
<th>Smoking duration, per 1 year</th>
<th>1.09 (1.08-1.09, &lt;0.001)</th>
<th>1.05 (1.03-1.06, &lt;0.001)</th>
<th>β = 0.063*(p&lt;0.001)</th>
<th>1.03 (1.02-1.03, &lt;0.001)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC: 0.767</td>
<td>β = 0.047</td>
<td>AUC: 0.641</td>
<td>β = 0.026</td>
</tr>
<tr>
<td></td>
<td>Pseudo-R² = 0.0922</td>
<td></td>
<td>Pseudo-R² = 0.0306</td>
<td></td>
</tr>
<tr>
<td>Smoking quit-time, per 1 year</td>
<td>β = -1.349*(p&lt;0.001)</td>
<td>0.99 (0.97-1.00, 0.046)</td>
<td>β = -0.489*(p&lt;0.001)</td>
<td>β = 0.005*(p = 0.001)</td>
</tr>
<tr>
<td></td>
<td>AUC: 0.711</td>
<td>β = -0.015</td>
<td>AUC: 0.582</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pseudo-R² = 0.0561</td>
<td></td>
<td>Pseudo-R² = 0.0967</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>AUC: 0.8002</td>
<td></td>
<td>AUC: 0.7114</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pseudo-R² = 0.1197</td>
<td></td>
<td>Pseudo-R² = 0.0677</td>
<td></td>
</tr>
</tbody>
</table>

**Abbreviations:** OR, Odds ratio calculated using univariate and multivariable logistic regression; LC, Lung Cancer; COPD, Chronic Obstructive Pulmonary Disease.

*6-year lung cancer incidence logistic regression model refers to the PLCO M2012 risk prediction model; multivariable model for competing causes of death refers to logistic regression model with the dependent variable as death from competing causes in 5 years and the independent variables are the 11 predictors in the PLCO M2012 risk prediction model.

†Age was centered at 62 years, education was centered at level 4 for the multivariable lung cancer incidence logistic regression model, body mass index was centered at 27 kg/m², smoking intensity was centered at 20 cigarettes per day, smoking duration was centered at 27 years, and smoking quit time was centered at 10 years.

‡Education was measured in six ordinal levels: less than high-school graduate (level 1), high-school graduate (level 2), some training after high school (level 3), some college (level 4), college graduate (level 5), and postgraduate or professional degree (level 6).

§Predictors have a nonlinear association with the outcome (lung cancer incidence or competing causes of death) and are therefore transformed. For this reason, odds ratio is not directly interpretable, and the beta coefficient is provided instead. β, beta coefficient represents the change in log-odds of the outcome with a one-unit change in the continuous independent variable or in the transformed variable in the case of nonlinear associations. The transformations were as followed:

- Education in the multivariable competing causes of death model = X^-2-.0745558807, where X = one of the six education levels (1-6).
- Body mass index in the univariate competing causes of death model = (X/10) ^-2-.1371742112 (for BMI1) and (X/10) ^-1-.3703703704 (for BMI2), where X= body mass index (kg/m²).
- Body mass index in the multivariable competing causes of death model = (X/10) ^-1-.3703703704 and (X/10) ^-1*ln(X/10)-.367871027, where X = body mass index (kg/m²).
- Smoking intensity in the univariate and multivariable lung cancer incidence model = (X/10) ^-.5-.7071067812, where X = number of cigarettes smoked per day.
- Smoking intensity in the univariate and multivariable competing causes of death model = ln(X/10)-.6931471806.
- Smoking duration in the univariate competing causes of death model = (X/10) ^2-.7.29, where X = number of years spent smoking.
- Smoking quit-time in the univariate lung cancer incidence model = [(X+1)/10] ^-.5-1.048808848, where X = number of years since smoking cessation. Smoking quit-time in the univariate competing causes of death model = [(X+1)/10] ^3-.1.331.

¶The backwards selection process led family history of lung cancer to be excluded from the final multivariable competing causes of death model.
PLCOM2012 model variables are ranked by predictive performance for the two outcomes, 6-year LC incidence and 5-year non-LC death (Table 7). The ROC-AUC value next to each predictor demonstrates the overall discrimination of the model once that variable was removed; the larger the drop in the ROC-AUC value, the greater the contribution of the variable on the model’s predictive performance. This allows comparison of relative importance of predictors in the PLCOM2012 model and the competing causes of death model. Age, smoking intensity, smoking duration and BMI are the four leading PLCOM2012 model predictors for both LC incidence and competing causes of death; the ranking of variables for both outcomes were congruent. Results from the univariate and multivariable models are summarized in detail on the following pages.

Table 7. PLCOM2012 model predictors ranked, from greatest to least (1-10) predictive ability, by decrease in AUC values* in multivariable lung cancer incidence and competing causes (non-lung cancer) death models.

<table>
<thead>
<tr>
<th>6-year LC incidence multivariable model (Full model AUC=0.8002)</th>
<th>5-year non-LC death multivariable model (Full model AUC=0.7114)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Smoking Intensity (0.7855)</td>
<td>1) Age (0.6858)</td>
</tr>
<tr>
<td>2) Age (0.7951)</td>
<td>2) Smoking Intensity (0.7057)</td>
</tr>
<tr>
<td>3) Smoking Duration (0.7953)</td>
<td>3) Smoking Duration (0.7071)</td>
</tr>
<tr>
<td>4) Body Mass Index (0.7980)</td>
<td>4) Body Mass Index (0.7077)</td>
</tr>
<tr>
<td>5 &amp; 6) Education† &amp; History of COPD (0.7986)‡</td>
<td>5) History of COPD (0.7079)</td>
</tr>
<tr>
<td>7) Race/Ethnicity (0.7990)</td>
<td>6) Race/Ethnicity (0.7082)</td>
</tr>
<tr>
<td>8) Smoking Status (0.7991)</td>
<td>7) Smoking Status (0.7094)</td>
</tr>
<tr>
<td>9) Personal History of Cancer (0.7994)</td>
<td>8) Personal History of Cancer (0.7096)</td>
</tr>
<tr>
<td>10) Smoking Quit-time (0.8001)</td>
<td>9) Education† (0.7099)</td>
</tr>
<tr>
<td></td>
<td>10) Smoking Quit-time (0.7109)</td>
</tr>
</tbody>
</table>

Abbreviations: AUC, Area Under the Curve; LC, Lung Cancer; COPD, Chronic Obstructive Pulmonary Disease.

*Brackets show the AUC value for whole model without the predictor variable and variables with a higher reduction in AUC represents greater predictive performance.
†Education was measured in six ordinal levels: less than high-school graduate (level 1), high-school graduate (level 2), some training after high school (level 3), some college (level 4), college graduate (level 5), and postgraduate or professional degree (level 6).
‡Education and history of COPD in the 6-year LC incidence model had similar reduction in magnitude of AUC when removed from the LC incidence model.
4.5.1 Evaluation of PLCO\textsubscript{M2012} Model Predictors in Univariate and Multivariable Analysis

4.5.1.1 Age

Age was coded as a continuous variable and centered at 62 years. Examining the association between age and LC risk, the odds of developing LC with every year decreased from 1.08 (95% CI 1.07-1.09, p<0.001) to 1.06 (95% CI 1.05-1.08, p<0.001), after adjusting for other predictor variables (Table 6, columns 1,2). For the CCoD model, the odds ratio for age decreased from 1.10 (95% CI 1.10-1.11, p<0.001, column 3) to 1.09 (95% CI 1.08-1.11, p<0.001, column 4), after adjusting for other predictor variables. This represents 1.09 greater odds or 9\% higher odds of dying from competing risks when age increases by one year. The ROC-AUC value of the univariate CCoD model for age was greater than the univariate 6-year LC incidence model (AUC: 0.649 vs. 0.626). Removing age from the multivariable 5-year non-LC death model resulted in the greatest reduction in AUC (AUC\textsubscript{Full model} = 0.7114 vs. AUC = 0.6858 without age). This shows that age contributed most to the overall prediction of 5-year non-LC death compared to other predictors, whereas age was ranked as the second most important predictor of 6-year LC incidence risk (Table 7).

4.5.1.2 Race/Ethnicity

Race/ethnicity was categorized into White/other and African Americans. In univariate analysis, the OR reported in the 6-year LC incidence model for African Americans was 1.76 (95% CI 1.55-2.01, p<0.001), and 1.66 in the CCoD model (OR = 1.66, 95% CI 1.36-2.02, p<0.001). However, in multivariable analysis, compared to White and other races combined, African Americans had 69\% higher odds of dying from competing causes within 5 years (OR = 1.69, 95% CI 1.47-1.94) and 70\% higher odds of developing LC in 6 years (OR =1.70, 95% CI 1.38-2.09), after adjusting for other variables (Table 6, both p<0.001).
The ROC-AUC value for the univariate 5-year non-LC death model was slightly greater than the univariate 6-year LC incidence model (AUC: 0.518 vs. 0.516), however were close in value. The ROC-AUC value also decreased to a slightly greater extent when race/ethnicity was removed from the 5-year non-LC model than when removed from the 6-year LC incidence model (ΔAUC Non-LC death = 0.0032 vs. ΔAUC LC-risk = 0.0012, Table 7). Overall race/ethnicity was significantly associated with both outcomes.

4.5.1.3 Education

Education was coded as a continuous variable due to the consistent change in effects seen in increasing levels of education in the univariate models. In the univariate model, there was a 16% decrease, or 0.84-fold lower odds, in 6-year LC incidence risk associated with every one-level increase in education (OR = 0.84, 95% CI 0.81-0.87, p<0.001). When examining the association between education and CCoD, a one-level increase in education status was associated with 0.88-fold lower odds (95% CI 0.86-0.90, p<0.001) of dying from competing causes in 5 years. (Table 6, columns 1,3).

Education had a linear association with LC risk in the multivariable 6-year LC incidence model and with every one-level increase in education, PLCO smokers have a 0.93 lower risk (6.9% decrease) of developing LC (OR = 0.93, 95% CI 0.90-0.97, p<0.001, column 2). In contrast, education was found to have a non-linear relationship with risk of death from competing causes risk through MFP analysis. A transformed variable of education was used in the final multivariable model and due to this nonlinear association, the odds ratio is not directly interpretable, and the nonlinear relationship is best demonstrated graphically (Figure 1). The association between education and risk of death from competing causes is negative; with every increase in education level, individuals have a reduced probability of dying from non-LC causes.
in 5 years. This effect is seen more sharply from level 1 (individuals who have less than a high-school degree) to level 2 (individuals who finish with a high-school degree).

The ROC-AUC value was lower for the univariate 5-year non-LC death model than for the univariate 6-year LC incidence model (0.555 vs. 0.577, Table 6). Adjusting for other variables, education was ranked higher in predictive ability in the 6-year LC incidence model compared to the 5-year non-LC death model (Table 7), suggesting that education level contributed more to the predictive performance of the LC incidence model compared to the non-LC death model.

Figure 1. Non-linear relationship between education (6 levels) and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207), adjusted for other predictors in the PLCO_M2012 risk prediction model.

Education was measured in six ordinal levels: less than high-school graduate (level 1), high-school graduate (level 2), some training after high school (level 3), some college (level 4), college graduate (level 5), and postgraduate or professional degree (level 6).

4.5.1.4 Body Mass Index

BMI was coded as a continuous variable and centered at 27kg/m². The relationship between BMI and 6-year LC incidence was linear; with every one-unit increase in BMI, individuals had significantly lower odds (OR = 0.97, 95% CI 0.95-0.98, p<0.001) of developing LC in 6 years, adjusting for other predictors (Table 6, column 2). Contrarily, in the univariate
and multivariable CCoD models, MFP analysis transformed BMI into two variables due to the nonlinear association observed between BMI and risk of 5-year non-LC death (Figure 2). Due to this nonlinear relationship, the odds ratios are not directly interpretable, and Figure 2 provides a meaningful interpretation by illustrating the relationship between BMI and non-LC death risk, when adjusting for other predictor variables. Underweight individuals with very low BMI (<20kg/m²) had an increased risk of death from non-LC causes, and this risk drastically decreased for normal weight individuals (18.5-25kg/m²), and gradually rose again for individuals classified as overweight (25-30kg/m²) and increased at a greater rate for obese individuals (>30 kg/m²).

The ROC-AUC was greater for 6-year LC incidence than 5-year non-LC death in univariate analysis (0.558 vs. 0.535, Table 6). Although the drop in AUC when BMI was removed from the 5-year non-LC death model was slightly greater than the AUC difference when BMI was removed from the 6-year LC incidence model (ΔAUC_{Non-LC death} = 0.0037 vs. ΔAUC_{LC-risk} = 0.0022, Table 7), BMI had the same ranking for both models meaning it was similarly predictive of both outcomes.

**Figure 2.** Non-linear relationship between body mass index and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207), adjusted for other predictors in the PLCO_M2012 risk prediction model.
4.5.1.5 Personal History of Cancer

Personal history of cancer was coded as a dichotomous variable and univariate analysis showed that PLCO smokers with a history of cancer (compared to those with no history of cancer) had 1.73 times greater odds of dying from non-LC causes during 5 years (95% CI 1.51-1.99, p<0.001, Table 6 column 3), which was greater than the odds of developing LC during 6 years (OR = 1.53, 95% CI 1.23-1.90, p<0.001, column 1). After adjusting for other predictor variables, individuals with a history of cancer, compared to those with no history, had 57% increased risk of non-LC death (OR = 1.57, 95% CI 1.36-1.81, p<0.001, column 4), but remained greater than the OR for the multivariable 6-year LC incidence model (OR = 1.35, 95% CI 1.08-1.68, p = 0.008, column 2).

Univariate ROC-AUC comparison shows that personal history of cancer had slightly higher AUC for 5-year non-LC death (AUC = 0.515) than 6-year LC incidence (AUC = 0.511) (Table 6). When adjusting for other variables, this difference in predictive performance between the two models decreased (Table 7). Overall, personal history of cancer was significantly associated with both outcomes.

4.5.1.6 Family History of Lung Cancer

Individuals with a family history of LC had 1.88 times greater odds of developing LC in 6 years compared to those with no family history (OR = 1.88, 95% CI 1.64-2.16, p<0.001). In univariate analysis, family history of LC had an odds ratio of 0.95-fold (95% CI 0.85-1.07, p = 0.396) for non-LC causes of death (Table 6). This relationship approached the null value and did not approach significance and therefore family history of LC was not included in the multivariable 5-year non-LC death model. In univariate analysis, the ROC-AUC value for family history of LC represented very poor discrimination (0.503), as compared to the univariate 6-year
LC incidence model, which demonstrated a greater AUC (0.541) (Table 6). Family history of LC was found to be a weak predictor of non-LC death during backwards section variable and is shown to be a much greater predictor of 6-year LC incidence, which makes sense as family history of LC has shown to increase an individual’s risk of LC through possible environmental or genetic factors or both.

4.5.1.7 History of Chronic Obstructive Pulmonary Disease

The odds ratio for developing LC in 6-years among those with COPD was 2.70 (95% CI 2.35-3.09, p<0.001) and for dying from non-LC causes was 2.34 (95% CI 2.12-2.58, p<0.001) (Table 6, columns 1,3). In adjusted analysis, individuals with COPD (compared to those without COPD) had a 45% increased risk for 6-year LC incidence (OR =1.45, 95% CI 1.26-1.67, p<0.001, column 2), compared to a 63% increased risk of dying from non-LC causes (OR = 1.63, 95% CI 1.47-1.81, p<0.001, column 4).

In univariate analysis, history of COPD had a greater ROC-AUC for 6-year LC incidence (0.559) than for 5-year non-LC death (0.547) (Table 6). In multivariable analysis, when history of COPD was eliminated from both models, the overall ROC-AUC difference was slightly greater for the 5-year non-LC model (ΔAUC \text{Non-LC death model} = 0.0035 vs. ΔAUC \text{LC-risk model} = 0.0016, Table 7). COPD was ranked similarly between both models, meaning there was similar importance in predictive performance between both outcomes.

4.5.1.8 Smoking Status

Cigarette smoking status was categorized into former-smokers and current-smokers. In univariate analysis, current-smokers had 3.26 times greater odds of developing LC (OR = 3.26, 95% CI 2.92-3.64, p<0.001) and 1.67 times greater odds of dying from competing causes (OR = 1.67, 95% CI 1.54-1.81, p<0.001) when compared to former-smokers. After adjusting for other
variables however, the OR for both modes were similar (LC incidence OR = 1.27, 95% CI 1.10-1.48, CCoD OR = 1.25, 95% CI 1.12-1.39) and both ORs were significant (p<0.05, Table 6).

Smoking status demonstrated greater AUC in the univariate 6-year LC incidence model (ROC-AUC = 0.623) compared to the univariate non-LC death model (ROC-AUC = 0.546) and this difference decreased in multivariable analysis, however smoking status appears to be a more important predictor of LC incidence than non-LC death after adjusting for other predictor variables (ΔAUC Non-LC death = 0.002 vs. ΔAUC LC-risk = 0.0011, Table 7).

4.5.1.9 Smoking Intensity

Smoking intensity (number of cigarettes smoked per day) had a nonlinear association with 6-year LC incidence and 5-year non-LC death in both univariate and multivariable MFP analysis and was transformed in all models. Figure 3 shows the positive nonlinear relationship between smoking intensity and probability of CCoD in 5 years. Individuals who smoked a greater number of cigarettes per day had an increased risk in non-LC deaths and this risk began to level off as smoking intensity increased.

The ROC-AUC value for the univariate 6-year LC incidence risk model (0.617) was greater than the ROC-AUC value for the univariate 5-year non-LC death model (0.562) (Table 6). When adjusting for other variables in multivariable analysis, smoking intensity continued to be a strong predictor of LC risk (Table 7).
Figure 3. Non-linear relationship between smoking intensity and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207), adjusted for other predictors in the PLCO_M2012 risk prediction model.

4.5.1.10 Smoking Duration

Smoking duration had a positive association with the risk of developing LC in PLCO smokers; the risk of developing LC drastically increases with increasing duration of smoking. With every one-year increase in duration smoked, there was 1.09-fold increase in odds of developing LC within 6 years (OR = 1.09, 95% CI 1.08-1.09, p<0.001). Smoking duration was transformed through MFP in univariate analysis for the 5-year non-LC death model (Table 6). Figure 4 shows the unadjusted relationship between number of years spent smoking and 5-year risk of death from non-LC causes. There is a positive exponential association between the number of years smoked and risk of dying from non-LC causes. The probability of dying from non-LC causes increases sharply for individuals who smoked ≥40 years.
Figure 4. The unadjusted non-linear relationship between smoking duration and risk of death from competing causes among former and current-smokers in the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (N=74 207).

In adjusted analysis, the OR for smoking duration in the 6-year LC incidence model was 1.05 (95% CI 1.03-1.06, p<0.001, column 2) and this was greater than the adjusted OR for smoking duration in the competing death model (OR per-year = 1.03 95% CI 1.02-1.03, p<0.001, column 4). This translates into a 1.03-fold increase in odds of dying from CCoD with every one-year spent smoking. The relationship between smoking duration and risk of non-LC death was a regular logit curve and was not transformed in MFP analysis). In univariate analysis, the ROC-AUC value for smoking duration represented fair discrimination (0.767) for 6-year LC risk and poor discrimination for 5-year non-LC death (0.641) (Table 6). Smoking duration was also more predictive of LC incidence than non-LC death in multivariable analysis (Table 7).

4.5.1.11 Smoking Quit-time

MFP analysis showed the relationship between smoking quit-time and 6-year LC risk to be non-linear in univariate analysis, and therefore smoking-quit-time was transformed in the model (Table 6, column 1) The risk of developing LC in 6 years significantly decreased in the
first 20 years of smoking cessation and then gradually continued to fall. Adjusting for other predictor variables, the relationship became linear and with every year an individual has remained free from smoking cigarettes, there was 0.99 lower odds of developing LC in 6 years (OR = 0.99, 95% CI 0.97-1.00, p = 0.046, column 2).

The association between smoking quit-time and risk of 5-year non-LC death was found to be nonlinear through MFP analysis and therefore smoking quit-time was transformed in univariate and multivariate analysis. In multivariable analysis, unexpectedly, smoking quit-time displayed a positive relationship with risk of death from non-LC causes, however this was due to the collinearity between quit-time and smoking duration, and therefore was not illustrative of the real relationship between smoking quit-time and 5-year non-LC death risk. Figure 5 depicts the negative univariate relationship between smoking quit-time and 5-year non-LC death in PLCO smokers. Greater number of years spent abstinent of cigarette smoking was associated with a decrease probability of dying from non-LC causes in 5-years. This decrease in risk was most prominent in the first 10 years of smoking cessation and continued to decrease exponentially.

Univariate analysis demonstrated that smoking quit-time was more predictive of 6-year LC incidence than for 5-year non-LC death (AUC = 0.711 vs. 0.582, Table 6) However, multivariable analysis demonstrated the opposite; adjusting for other variables, smoking quit-time showed a slightly greater drop in ROC-AUC when removed from competing causes of death model than when removed from the LC incidence risk model, although this difference was small ( ΔAUC Non-LC death = 0.0005 vs. ΔAUC LC-risk = 0.0001, Table 7).
4.5.2 Competing Causes of Death Model Performance

Statistics used to evaluate the 5-year non-LC death model’s performance are presented in Table 8. Discrimination was fair when the model was applied to PLCO smokers (AUC = 0.7114, 95% CI 0.7017-0.7201), demonstrating that the non-LC model was reasonable at discriminating between individuals who would or would not die of non-LC causes in 5 years. The Brier score, average squared difference between the observed and predicted risk of outcome (5-year non-LC death), was 0.0373 (95% CI 0.0362-0.0385) which demonstrates good overall model performance. The Spiegelhalter’s z-statistic was also non-significant (p = 0.447) indicating good calibration of the 5-year non-LC model.

The mean absolute difference between the observed and predicted risks in the 5-year non-LC death model was very low (~0.001). The median (50th percentile) and 90th percentile absolute differences between observed and predicted risk probabilities were 0.001 and 0.002, respectively. This means that the differences between observed and predicted probabilities of
non-LC death risk was less than 0.001 in half of the sample and less than 0.002 in 90% of the sample, which indicates good calibration.

**Table 8. Predictive performance of multivariable competing causes of death model in Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial smokers (N=74207).**

<table>
<thead>
<tr>
<th>Model performance</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrimination</td>
<td></td>
</tr>
<tr>
<td>Area under the curve (95% CI) *</td>
<td>0.7114 (0.7017-0.7201)</td>
</tr>
<tr>
<td>Calibration</td>
<td></td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.0012149</td>
</tr>
<tr>
<td>50th and 90th percentile absolute differences</td>
<td>0.00093, 0.00186</td>
</tr>
<tr>
<td>Spiegelhalter’s z-statistic (p-value)</td>
<td>0.1329 (p = 0.4471)</td>
</tr>
<tr>
<td>Overall prediction</td>
<td></td>
</tr>
<tr>
<td>Brier score (95% CI) *</td>
<td>0.0373 (0.0362-0.0385)</td>
</tr>
</tbody>
</table>

*Bias-corrected confidence intervals using bootstrapping of 1000 samples

The calibration plot, seen in Figure 6, shows that the model was predicting non-LC death relatively accurately. Although the model starts to overestimate after predicted risk of ≥20%, figure 7, which displays a plot of individuals who developed or did not develop non-LC death within 5 years, shows that majority of individuals (scatter diagram “1”) had risks below 20%. Therefore, the non-LC death model was correctly predicting the risk of dying from non-LC causes among PLCO smokers, many of which had risks lesser than 20%.

**Figure 6. Calibration plot: Observed vs. predicted 5-year non-lung cancer death probabilities.**
The PLCO\textsubscript{M2012} model risk score was also significantly associated with 5-year non-LC death (p<0.001) in univariate logistic regression analysis. Figure 8 represents a LOWESS plot showing the relationship between the PLCO\textsubscript{M2012} model score and 5-year probability of non-LC death in the smokers’ population. As the PLCO\textsubscript{M2012} model’s probability increases, the probability of dying from competing causes also rises. When PLCO\textsubscript{M2012} model risk was used to predict 5-year non-LC death, the corresponding ROC-AUC value was 0.670 (95% CI 0.660-0.680) representing fair discrimination, indicating that factors predicting LC also predict CCoD.
4.5.3 Competing Causes of Death Model Assumption & Diagnostics

Collinearity diagnostics were run for all variables in final multivariable CCoD model. The correlation matrix showed a negative correlation between smoking duration and quit-time, however, there was no problematic degree of multicollinearity observed as variance inflation factor (VIF) values were all under 10. The highest VIF value were for smoking duration (4.21) and smoking quit-time (3.17) and the mean VIF was 1.64.

Pearson, deviance and Pregibon leverage residuals for the 5-year CCoD model when applied to PLCO smokers demonstrated that very few observations showed any deviation from the rest of the sample and these observations did not hold much importance to the overall model’s performance individually, as observed by the leverage plot. By not including these observations, the final logistic regression estimate will not be much different from the model that does include these observations, therefore these observations were not removed.
5. **Discussion**

This study demonstrated the superiority of using risk prediction modelling, specifically the PLCO M2012 LC risk prediction model, over the NLST criteria, in identifying individuals who are at high-risk for developing LC. Furthermore, findings based on the overall ROC-AUC values for the CCoD and LC incidence risk models show that the PLCO M2012 model appears to be more predictive of LC incidence than non-LC mortality. However, most of the PLCO M2012 model predictors increase the selection of individuals with comorbidities, and therefore increasing their risk of mortality from competing causes, and this will be discussed in greater detail below.

5.1 **PLCO M2012 Model Superior in Selecting High-Risk Individuals for LC Screening**

5.1.1 **The PLCO M2012 Model at ≥1.5%/6yr Risk Threshold**

The first aim of this study was to determine if the PLCO M2012 risk prediction model was more effective, than the NLST criteria, in identifying high-risk individuals for LC screening and detecting more LC cases. Contingency table analyses demonstrated that the model with a risk threshold of ≥1.5%/6yr was statistically and clinically more efficient than the NLST criteria, because it led to a smaller number of individuals being screened, identified significantly more LCs, and had a higher PPV, meaning individuals selected for screening by the model had a higher 6-year risk of developing LC. These findings were comparable to previous studies stratifying PLCO smokers by NLST versus PLCO M2012 (≥1.5%/6yr risk) eligibility. When validated in different cohorts of ever-smokers, the PLCO M2012 at this threshold selected a greater number of high-risk individuals for screening as compared to the NLST or NLST-like criteria. The PLCO M2012 risk prediction model demonstrated many advantages over the NLST criteria, such as having high prediction, discrimination, and good calibration in different validation datasets, as well as selecting individuals at high-risk for LC, making it more efficient.
Previous studies comparing the accuracy of the PLCO\textsubscript{M2012} model (≥1.5%/6yr risk) against the NLST/NLST-like criteria showed the model to have significantly improved sensitivity (83.0% vs. 71.1%) and PPV (4.0% vs. 3.4%) without loss of specificity (62.9% vs. 62.7%).\textsuperscript{(51)} Our study demonstrated similar performance; sensitivity, specificity and PPV were significantly improved for the model at ≥1.5%/6yr risk (p<0.001), translating into improved accuracy.

5.1.2 The PLCO\textsubscript{M2012} Model at ≥2.0%/6yr Risk Threshold

Increasing the model threshold risk resulted in fewer individuals selected for LC screening compared to the NLST criteria, less false positives, and an overall smaller sample size. The PLCO\textsubscript{M2012} model at ≥2.0%/6yr risk has shown to select a smaller number of U.S ever-smokers for screening than the USPSTF criteria,\textsuperscript{(187)} however this may vary by study population; in the Australian 45 and Up Study, the PLCO\textsubscript{M2012} model (≥2.0% risk) selected a greater number of individuals for screening than the NLST-like, or USPSTF, risk criteria among a sub-cohort of ever-smokers (55-79 years).\textsuperscript{(166)}

The PPV of the PLCO\textsubscript{M2012} model was greatest at ≥2.0%/6yr risk threshold compared to the model at ≥1.5%/6yr risk or the NLST criteria, and this is consistent with previous literature.\textsuperscript{(160,186)} The PPV of a test is an important measure of accuracy because clinicians and researchers are interested in knowing how likely an individual, who is eligible by a selection criterion, is to develop LC.\textsuperscript{(52,160,186)} PPV varies depending on the disease prevalence in the population and since LC prevalence was low among PLCO smokers, PPV values for the NLST criteria and PLCO\textsubscript{M2012} model were small in magnitude. The PPV also gives information on the 6-year risk of developing LC which was greatest for individuals who were eligible by the PLCO\textsubscript{M2012} model (≥2.0%/6yr risk). This means that even though the model (at ≥2.0%/6yr risk) selected fewer participants for screening with an overall smaller absolute number of LCs, these
participants had the highest risk for developing LC, translating into a greater proportion of LC
deaths averted per number of people screened. These findings suggest that using the PLCO_{M2012}
model (≥2.0%/6yr risk) to select high-risk individuals for LC screening might improve cost-
effectiveness as resources would be targeted towards individuals with the highest LC risk. This is
important because for policymakers to implement population-based LC screening, cost-
effectiveness must be considered. Within cost-effective analysis, the benefits of minimizing the
number of false-positives and individuals required to be screened is important. False positives
are associated with unnecessary diagnostic evaluations that consist of further imaging and
invasive procedures.\textsuperscript{[17,188]} These procedures can be costly, time consuming and can subject
people to psychological stress. It is worth mentioning that the risk for harms from LC screening
and false-positive results have been minimized with appropriate management of screening
results, such as using a Lung Imaging Reporting and Data System (Lung-RADS).

Increasing the PLCO_{M2012} model risk threshold decreased the likelihood of capturing true
positives (LC cases), therefore reduced sensitivity. The consequences of missing potential LCs
are severe and not screening potential high-risk individuals could result in delayed diagnosis
which in turn can result in treatable early-stage cancers to develop into higher stage
malignancies, negatively influencing prognosis. However, previous studies evaluating the model
at ≥2.0%/6yr risk in different populations show the model to detect more LC cases than the
NLST or NLST-like criteria, while having improved sensitivity, specificity and PPV.\textsuperscript{[52,160,186]}

5.2. Components of the PLCO_{M2012} model select individuals with highest 6-year LC risk

The second aim of this study was to determine which components of the PLCO_{M2012}
model were increasing the selection of high-risk individuals. In other words, which model
predictors picked up the greatest proportion of LCs among smokers in the high-risk strata for
each variable who were eligible exclusively by the PLCO_{M2012} or the NLST criteria.
This study verifies that \( \text{PLCO}_{2012} \) model predictors are important risk factors for LC. Among smokers selected for screening by the model vs. the NLST criteria, individuals with a family history of LC had a greater likelihood of developing LC than those with a history of COPD or personal cancer, demonstrating that family history of LC has a strong association with LC. When looking at 6-year LC risk among individuals eligible by the \( \text{PLCO}_{2012} \) model, smokers with a history of COPD had the highest risk of developing LC in 6 years. Overall, the higher proportion of LCs associated with each risk factor was consistent with what was seen in previous studies, especially looking at ever-smokers.\(^{(41,46,48,49,51,52)}\)

Interpreting and evaluating risk is more intuitive than odds; compared to the NLST criteria, the \( \text{PLCO}_{2012} \) model (\( \geq 1.5\% \) and \( \geq 2.0\%/6\text{yr} \) risk) selected a greater proportion of individuals (except for smoking duration, intensity and quit-time), with the greatest 6-year LC incidence risk, in the high-risk strata for each risk factor, meaning that individuals were more likely to develop LC if eligible by the model.

Findings demonstrate that the NLST criteria selects some individuals who are at relatively low risk for developing LC, while failing to include a statistically significant large proportion of LC cases as compared to the \( \text{PLCO}_{2012} \) model. The NLST criteria has previously been criticized for selecting individuals at low-risk for developing LC and excluding those at high-risk for developing LC.\(^{(185,189)}\) For optimal cost-effectiveness, individuals with a sufficiently high risk of developing LC need to be identified so that the benefit-to-harm ratio of screening can be maximized. If LC screening is offered to low-risk individuals, then very few cases of LC will be detected and there would be no significant reduction in LC mortality. In the NLST, the number of individuals who would need to be screened to prevent one LC death decreased from 5276 for the CT group among 20% of participants at lowest risk (0.15-0.55%) to 161 among the 20% of participants at highest risk (\( \geq 2.0\%/6\text{yr} \)).\(^{(189)}\) Therefore, a greater number of low-risk
individuals would be needed to screen to prevent one death from LC and this would make LC screening less cost-effective. By utilizing the PLCO\textsubscript{M2012} model over the NLST criteria, resources would be targeted towards high-risk individuals who would benefit the most from screening. This could help optimize public health services and decrease medical costs associated with LC screening (intervention costs, costs of follow-up and costs associated with false-positive results), and lead to greater potential lives saved from screening.

5.3 Predicting Competing Causes of Death

The 5-year non-LC death model demonstrated fair to adequate discrimination in our dataset (AUC = 0.7114). This study revealed that the PLCO\textsubscript{M2012} model estimated 6-year LC incidence risk was predictive of 5-year non-LC death, meaning the model is selecting individuals at high-risk for LC but also at risk of dying from competing risks. Interestingly, the relative rank order of important predictors for LC and non-LC death was remarkably similar, except for family history of LC. Only 4 out of the 10 model predictors are discussed below; the four leading predictors of both outcomes were age, smoking intensity, smoking duration and BMI. The two models are not directly comparable due to differences in individuals, follow-up periods and outcomes, nevertheless, relative comparisons may still reveal pertinent points.

5.3.1 Age

Age was significantly associated with both risk of 6-year LC incidence and 5-year non-LC death. Age was one on the two leading predictors for both models, therefore had similar ranking between the two. Older age is linked to numerous comorbidities such as heart disease, myocardial infarction, hypertension, cerebrovascular disease, peripheral vascular disease, respiratory disease (including COPD), diabetes and many more; all of which increase the risk of death in the elder population.\textsuperscript{(175,176)} Several of these comorbidities are also prevalent among LC patients which contributes to reduced LC survival.\textsuperscript{(175)} One study evaluating the impact of
comorbid conditions on patient survival found that majority of LC diagnoses were among individuals >65 years of age (highest age group), and each stage of LC (localized, regional and distant) was associated with worse survival when comorbidities were present.\(^{(55)}\)

The benefit of LC screening is directly related to the benefit of treatment of early-stage LC. High rates of age-related comorbidities which increase the risk of non-LC death in older individuals limits the benefit of screening and increases the harms from screening. Such harms include subsequent invasive procedures, operative risks and false-positive results which are higher in the elder population.\(^{(190)}\) This is important because the PLCO\(_{M2012}\) model selects an older population for screening compared to the NLST criteria, and therefore selects individuals with comorbidities whom are more likely to die from competing risks. Screening these individuals is expected to lead to fewer life-years gained and QALY gained than screening younger individuals.\(^{(185)}\) Recommendations to use the NLST, or NLST-like criteria, over risk prediction modelling partly come from the notion that risk prediction models results in selecting individuals who are older, with poorer health and more comorbidities, who would not benefit from screening.\(^{(128)}\) However, previous analysis among PLCO smokers found that excluding individuals at low-risk (<1.5% risk by the PLCO\(_{M2012}\) model) in the NLST criteria+/ PLCO\(_{M2012}\)- group, resulted in a sample with similar mean age as the PLCO\(_{M2012}\)+/ NLST criteria – group, demonstrating that the lower mean age in those eligible by the NLST criteria was greatly driven by the young age occurring in those who were at low risk.\(^{(185)}\) Therefore, while screening younger individuals would lead to potentially more life years gained with fewer comorbidities, it would also result in fewer individuals being selected for screening whom are at high-risk for LC, resulting in potential LC cases missed and no significant benefit in mortality reduction.

As expected, age was a significant predictor of 6-year LC incidence risk (p<0.001). Despite the concerns with screening older individuals, recent studies have shown that
the benefits of LDCT screening, in terms of detecting more LC cases and preventing LC death, can extend beyond the current age criterion (55-77 years).\(^{(191)}\) Although the ranking of predictors based on AUC values showed age to be a stronger independent predictor of 6-year LC incidence risk than smoking duration, smoking duration has been shown to be the most predictive factor of LC risk among smokers in other studies.\(^{(71)}\) Overall, since age was one of the top predictors of both LC incidence and non-LC death, assessing competing risks from other causes is imperative when making decisions related to screening in older populations.

5.3.2 Smoking Intensity and Duration

Smoking is the primary risk factor for LC and has shown consistently to be a strong predictor of LC risk, hence it was interesting to find that smoking intensity and duration were two of the top 3 leading predictors of non-LC death and demonstrated similar ranking in predictive performance of LC incidence and non-LC death. Heavy tobacco use is associated with various comorbidities, which, when severe enough, are associated with greater risk of harm from screening and reduced ability to treat screen-detected cancers.\(^{(192)}\) These comorbidities include other respiratory diseases, especially COPD, cardiovascular diseases, and other cancers (mouth, pharynx, esophagus, and stomach), all of which increase the risk of non-LC specific death.\(^{(54,192)}\) In a previous study, a longer duration of smoking was associated with an increased risk of mortality due to COPD, any cancer and LC. A longer duration of smoking cessation was associated with a decreased risk of all-cause and CVD mortality.\(^{(169)}\) Smoking a greater number of cigarettes per day (≥20 cigarettes/day) has been shown to increase the risk of mortality from infectious diseases, neoplasms, CVD, respiratory diseases and digestive system disorders.\(^{(193)}\)

The association between LC risk and smoking intensity and smoking duration observed in this study was consistent with previous studies. \(^{(49)}\) Even though the PLCO\(_{M012}\) model selected a smaller proportion of participants with greater smoking intensity (≥20 cigarettes/day) and
duration of smoking (≥28 years) than the NLST criteria, it detected more LC cases among these individuals. Therefore, smokers detected by the model were at greatest risk for LC and could also have been at greater risk of non-LC death. Studies have shown that smoking, especially at an older age (60-70 years), increases the risk of dying from LC to a greater extent than dying from cardiovascular disease, stroke or other cancers in 10 years.\(^{163,194}\) Although smoking is associated with reduced survival, current-smokers are the most targeted populations for LC screening as they are at highest risk for developing LC compared to those who do not smoke.

5.3.3 Body Mass Index

BMI had a strong nonlinear U-shaped relationship with CCoD among PLCO smokers; the risk of death from competing causes was lowest among normal weight (18.5-24.9 kg/m\(^2\)) and overweight (25-29.9 kg/m\(^2\)) individuals and greatest among underweight (<18.5 kg/m2) and obese (≥30 kg/m\(^2\)) individuals. A similar association was observed in other studies among different populations consisting of smokers and nonsmokers.\(^{57,195}\) Compared to smokers with a normal BMI, smokers who are overweight have an greater risk of dying from all-cause mortality, mainly from cardiovascular disease.\(^{196}\) Obesity, as well as being underweight increases an individual’s risk of all-cause mortality. Obesity is linked to a range of chronic comorbidities such as cardiovascular diseases, hypertension and type 2 diabetes mellitus,\(^{197}\) while being underweight indicates poorer nutrition, bone loss or the presence of an illness which has detrimental effects to physical well-being, and this negative effect worsens with age.\(^{198,199}\)

Smokers tend to be leaner than non-smokers for several reasons,\(^{200,201}\) and current-smokers face multiple consequences to physical health, therefore having increased mortality due to both smoking and smoking-related weight loss.\(^{199,202,203}\) An OR of 0.97 (95% CI 0.95-0.98, p<0.001) was reported for the association between BMI and LC incidence risk which is similar
to the BMI and LC risk association reported in previous studies\(^{(49,52)}\) indicating that increased BMI has a protective effect on LC risk, even after adjusting for smoking status.\(^{(95,200)}\)

5.3.4 **CCoD negatively impacts Lung Cancer Screening Outcomes**

In this study, PLCO\(_{M2012}\) model predictors, except for family history of LC, that were important in the selection of high-risk individuals for LC screening were also strongly associated with an increase in risk of dying from competing, or non-LC, causes in a brief period of time. This suggests that risk prediction models, such as the PLCO\(_{M2012}\), select an ample proportion of individuals who might not benefit from screening in the first place. For individuals to benefit from a screening program, they should be healthy enough to undergo screening, diagnostic and treatment procedures and should not die from something else before getting diagnosed and treated for LC. The main benefit, or outcome, of LC screening is reduction in LC mortality through early diagnosis and treatment of LC. However, studies show that deaths from competing causes surpass LC incidence rates and LC deaths among individuals at highest risk for LC (PLCO\(_{M2012}\) ≥2.0% risk threshold).\(^{(189)}\) Therefore, competing risks which increase an individual’s risk for CCoD will result in a smaller reduction in LC-specific mortality, as individuals will be dying from other causes before they acquire, or die from, LC.

CCoD negatively impacts the quality of adjusted life-years gained (QALY). QALY is an informative measure of LC screening effectiveness which considers the number of life-years gained due to early detection of a disease as well as the quality of those extra years. By receiving early treatment for a disease, an individual may have a higher life expectancy, however, if they spend those years with a disability, chronic pain and/or anxiety, then the overall benefit of screening, in terms of QALY’s gained, appears insignificant. Utilizing the PLCO\(_{M2012}\) model for screening selection may result in a decrease in life-years gained and QALY’s gained because the model is selecting individuals who are older, with comorbidities who are dying from non-LC
causes in a short period of time, making LC screening seem less beneficial. Individuals at high risk for non-LC death are also subject to significant harms from screening and other downstream evaluations because co-existing illnesses render them to become less tolerable to these procedures, increasing their risk of complications as well as emotional stress and anxiety.\textsuperscript{(26,192)} It is worth mentioning that although the benefit of screening, when measured in life-years or QALY gained, is lower than when measured by number of deaths averted, the PLCO\textsubscript{M2012} model has shown to still perform better than the NLST or NLST-like criteria.\textsuperscript{(204)} This study’s findings highlight the importance of accounting for non-LC specific mortality as a competing event when assessing LC risk among high-risk individuals, especially in the elderly population; possible ways to examine CCoD will be discussed shortly.

5.4 Study strengths

This study has several strengths including: 1) The use of a prospective cohort study that allows for the calculation of incidence, which is ideal for risk prediction modeling,\textsuperscript{(51)} and a large sample size which allows for the examination of multiple outcomes, while increasing power and providing precise effect estimates. The PLCO dataset is also of high-quality with rigorous follow-up procedures ensuring that baseline information, as well as the end points, are accurately documented throughout the study period.\textsuperscript{(15)}

2) The methodological process of this study involved advanced statistical techniques such as multivariable fractional polynomial analysis, which allowed for the examination of non-linear relationships of continuous variables, and bootstrapping techniques, which were used to produce bias-corrected confidence intervals for predictive performance measures of the final 5-year non-LC death model, resulting in improved internal validation.

3) To measure the predictive ability of the PLCO\textsubscript{M2012} model predictors in the non-LC death and LC incidence model, ROC-AUC values were used rather than p-values. P-values provide
information on which variables have a significant association with the outcome of interest but provide no information on how well variables improve the prediction of the outcome. Since our study aimed to measure predictive performance, and not causal relationships, ROC-AUC values were more appropriate to use.\(^{(42)}\) By utilizing ROC-AUC values and biological reasoning as compared to a p-value cut-point, useful predictors were included in and uninformative variables were excluded from the 5-year non-LC death model.

4) This study is novel in several ways; this is the first study, to the best of our knowledge, to directly compare each specific predictor between individuals who are eligible exclusively by either the NLST criteria or the PLCOM\(_{2012}\) risk prediction model. No other study, to the best of our knowledge, assessed the effectiveness of the model at a ≥2% six-year risk in selecting high-risk individuals for each individual predictor or has investigated the association between each PLCOM\(_{2012}\) model predictor variables and non-LC mortality.

5.5 Study limitations

Possible study limitations include the following: 1) Decreased generalizability as the PLCO population was not completely representative of the general U.S. population. Although participants in the PLCO Cancer Screening Trial were selected to include individuals from different backgrounds in an effort to be representative of the whole U.S. population, the ever-smokers subset seemed to be disproportionately white.\(^{(15,180)}\) In our study, 5.4% of ever-smokers were African Americans, however, in the U.S population approximately 17.6% of ever-smokers are African American adults\(^{(6)}\); this suggests that the study population is underrepresenting minorities, especially those at high risk for LC. Participants in this study were also more likely to have attained post high-school education. However, in the general U.S. population, cigarette smoking is higher among those with a GED (general education development) certificate than those with post-high-school education or a higher level of education.\(^{(6)}\) The PLCO cancer
screening trial has been described previously as having low generalizability due to the study population not being racially diverse and having a higher SES than the U.S. population (180); the PLCO_{M2012} model was created using U.S. data and the results of this study are exclusively generalizable to the U.S. population, however, the PLCO_{M2012} model has been shown to perform well when validated using Canadian, Australian, German and UK data. (57,166,167,193)

2) Although the percentage of missingness in our study was low, multiple imputation was not used to estimate missing data, instead missing values were excluded from analysis and this can result in the loss of valuable information and bias results. However, in this study, there were no variables with a large proportion of missing observations, therefore the amount of bias introduced due to missing data, and the effect on prediction, is presumed to be minor.

3) One major limitation in this study was the inability to directly compare the effect estimates and predictive performance of the CCoD and LC incidence model. The two models assessed different outcomes with different follow-up periods among different group of smokers, making the models incomparable.

4) Logistic regression was used in this study, producing odds ratios, which may not be the best measure of association between risk factors and LC incidence or non-LC mortality. Relative risk is a more accurate measure of association for cohort studies. (206) However, if the outcome is rare (<10%) in the population or population-based sampled is used then the OR can be used to estimate relative risk, and the incidences of both outcomes in our study were low and sampling was roughly pseudo-population-based.

5) Our results do not provide any information on how higher-risk selection can affect the likelihood of being over-diagnosed. The number of over-diagnosed or incidental LC cases were not evaluated in our study, therefore the extent to which the PLCO_{M2012} model had a higher overdiagnosis rate among PLCO smokers, compared to the NLST criteria, was unknown.
5.6 Implications

When implementing a high-quality LC screening program, identifying an eligible population is one of the first and most important steps. This study proposes that the PLCO\textsubscript{M2012} model is an efficient and effective method to identify individuals who are at high-risk for developing LC. This can have several implications to the public healthcare system. For instance, utilizing the PLCO\textsubscript{M2012} model over the NLST criteria in organized screening programs to select individuals for screening would be more economical because resources would be targeted towards individuals who benefit the most from screening, in terms of early diagnosis and treatment, and this translates into the PLCO\textsubscript{M2012} model being more cost-effective. However, cost-effectiveness analysis involves multiple factors which were not measured in this study, such as costs associated with screening, diagnosis and treatment, and patients’ quality of life, as well as the costs associated with curative and non-curative treatment mortality, and side effects.\textsuperscript{(207,208)}

Although cost-effective analysis was not performed in this study, compared to the NLST criteria, results were in favor of the PLCO\textsubscript{M2012} model at $\geq2.0\%/6\text{yr}$ risk to be cost-effective as it selected less individuals for screening, with decreased false positive results and a higher PPV, although with lower sensitivity. Overall, the PLCO\textsubscript{M2012} model accurately predicts LC incidence risk while being relatively easy to use; predictor variables can be measured noninvasively, and calculation of risk is neither time consuming nor costly.

Another important reason for screening a high-risk population is to maximize benefits and reduce the exposure of risks associated with screening. More LC cases detected by the model than the NLST criteria translates into early diagnosis and treatment and therefore more LC deaths averted. Utilizing the PLCO\textsubscript{M2012} model to select individuals for screening results in a greater reduction in LC mortality which also means a reduction in the number of individuals needed to be screened to prevent one death from LC (NNS). The number of LY and QALY
gained is also an important measure of the benefits of screening; although these measures were not evaluated in this study, the PLCO\textsubscript{M2012} selected individuals who had competing risks which put them at elevated risk for non-LC mortality, potentially attenuating LY and QALY’s gained among these individuals. However, as mentioned previously, many individuals eligibly by the NLST criteria are at too low-risk to benefit from screening, making the NLST criteria less beneficial and less cost-effective. Previous studies have shown that screening high-risk individuals (fourth and fifth quintiles of the PLCO\textsubscript{M2012} model risks) results in significantly lower cost per quality-adjusted life year gained, compared to screening those at lower risk.\textsuperscript{(209)}

It is also important to consider the clinical benefits and risks of LC screening when determining an optimal risk threshold. PLCO\textsubscript{M2012} $\geq 1.5\%/6\text{yr}$ risk has been determined as an appropriate threshold for selecting individuals for LC screening.\textsuperscript{(42,52)} However, different thresholds may be suitable when the model is applied to different populations.\textsuperscript{(58,168)} Decision curve analysis considers the impact of false positive and false negative results and can graphically provide the net benefit of one or more models over a range of model risk thresholds.\textsuperscript{(207,208)} The net benefit represents the difference between the expected benefit of screening, if an individual has the disease, and the harms that would follow a false-positive test. A decision curve comparing the PLCO\textsubscript{M2012} model and the NLST criteria in the PLCO intervention arm smokers found that the model’s benefits exceeded the benefits of applying the NLST criteria at all risk thresholds under 9.0%.\textsuperscript{(155)} Many decision analyses require information not readily available in validation studies and these analyses were beyond the scope of this study.

One important implication of the PLCO\textsubscript{M2012} risk prediction model is that it considers racial/ethnic differences when estimating LC incidence risk, whereas the NLST criteria ignores this disparity.\textsuperscript{(51)} Addressing these healthcare differences has social implications; screening programs should target recruitment among high-risk African Americans, reducing the burden of
LC in this population. Early LC screening implementation helps to further investigate what factors influence screening behavior in high-risk populations and this knowledge can be used to design proper patient outreach programs. Adopting the PLCO\textsubscript{M2012} risk prediction model for LC screening eligibility can help increase awareness of screening options in underrepresented and geographically dispersed patient populations and foster dialogue between patients and clinicians about screening eligibility.

**Implications in Canada**

LC is a burden to the Canadian public health system and similar trends in LC incidence rates exist in Canada as they do in the U.S.\(^5\) The presence of racial/ethnic LC disparities should be investigated in Canada. Research is required on Indigenous populations whose high-risk levels put them in a position to benefit from LC screening. An ample proportion of members from the First Nation community develop LC at a younger age (<55 years), hence, the performance of prediction models which include predictors for Indigenous peoples’ risks should be evaluated in these populations to assess whether screening younger Indigenous high-risk individuals is cost-effective.\(^{168}\) The PLCO\textsubscript{M2012} model has worked well using Canadian data, and has shown to be efficient and affordable to the Canadian health care system.\(^{52,209}\) A cost-effective analysis performed in preparation for CCO’s “Lung Cancer Screening Pilot for People at High-Risk” found that using the PLCO\textsubscript{M2012} model at \(\geq 2.0\%\) risk threshold would identify the same number of individuals being screened as a preferred microstimulated NLST-like model-based criteria which was affordable at \(<\$50,000\) per life-year saved, while having significantly higher sensitivity and PPV.\(^{209}\)

Implementation of the PLCO\textsubscript{M2012} risk prediction model in Canadian settings could help prevent many deaths per year. Utilizing the model in various setting such as in general practitioner offices, walk-in clinics, hospitals or screening centers could lead to an accurate
selection of high-risk individuals for screening, preventing unnecessary exposure to harms from screening and waste of medical resources. LC screening using LDCT in not yet part of the standard of care in Canada’s health care system and there is no consensus on whether risk prediction modeling should be used for LC screening eligibility. The Canadian Associations of Radiologists recommend the use of validated risk prediction models to determine a high-risk population for screening, specifically screening patients who have a 1.5% or higher risk of developing LC over the next 6 years.\textsuperscript{(210)} LC screening guidelines, however, are in favor of using the NLST or NLST-like criteria to select individual for screening; the Canadian Task Force on Preventative Health Care recommends LC screening using LDCT once each year for 3 years in individuals defined as high-risk using the NLST criteria.\textsuperscript{(211)}

Before the PLCOM\textsubscript{2012} model can be established into the public health system, screening trials using risk modelling based selection must be conducted. To our knowledge, there are no organized LC screening programs in Canada. Organized screening provides certain benefits such as ensuring that appropriate populations are screened with appropriate follow-up of abnormal findings, and providing quality monitoring, reporting and management of findings. Organized LC screening will also ensure that CT scans use the appropriate low-dose of radiation and that scans are interpreted by trained radiologists.\textsuperscript{(212)} Some provinces have begun LC screening strategies, such as initiating pilot studies, to better understand how to properly implement provincial-wide organized LC screening. For example, Ontario’s “LC Screening Pilot for People at High-risk” is currently underway, and the purpose of this pilot is to determine how to best implement organized LC screening in Ontario through the evaluation of key screening components such as recruitment, navigation, LDCT scanning, retention, follow-up, cancer stage at diagnosis and treatment.\textsuperscript{(53)} Results are already promising in that a high proportion of early stage LCs are being detected; evaluation of this pilots’ results will inform the design and
implementation of a provincial LC screening program. Findings from various LC screening trials that are in favor of using risk prediction modelling help organizations make recommendations to use the PLCOM2012 model in public health practice. This can help convince or guide policymakers in the direction of recommending the use of model-based risk assessment for determining who should be offered screening, as well to provide funding to local health organizations to conduct further screening trials and to implement the model into clinical practice.

Challenges with implementing risk prediction models

Identifying and inviting only eligible individuals into a population-based LC screening program can be difficult; recruitment strategies would need to target individuals who would benefit the most from screening. In attempt to invite appropriate individuals for population-based LC screening programs, LC screening of high-risk individuals could be publicized through various means such as telephone interviews, advertisements, radio, social media, cancer-related charity events or through general practices. Utilizing mobile LC screening units in underserved high-risk (PLCOM2012 model at ≥1.5% risk) populations, from a community-based ‘Lung Health Check’ pilot in Manchester, U.K., show promising results in reducing disparities in screening and late-stage LC diagnoses. Increasing screening availability in marginalized communities translates into possibly more treatable early-stage LCs detected and access to proper treatment, as well as better quality of care, among individuals in these communities.

Patient decision aids should become widely available for the public to increase awareness of the benefits and harms of LC screening, as well as information on the factors that contribute to an increased risk of LC; this will help improve the public’s understanding of the rationale behind using model risk assessment for selecting high-risk individuals for LC. Decision aids have been made available such as infographics, online toolkits, factsheets and FAQs. Publicly
available online risk calculators have also been developed and can aid in the decision-making process for individuals in the general population to help facilitate discussion with their primary health care providers to discuss if LC screening is appropriate for them. Informed decision-making is an important part of the doctor-patient relationship and is crucial when implementing a LC screening program. It allows patients and their providers to make health care decisions together while considering numerous factors like important clinical evidence on the topic of LC screening and its harms and benefits, the patients values and preferences and the patients current health status. For optimal shared decision-making, primary health care providers should be knowledgeable about LC screening and model-risk based selection and should have good understanding and knowledge about the evidence based and clinical practice guidelines of both methods.

A challenge when utilizing the PLCO\textsubscript{M2012} model to aid in the decision-making process is the collection of information on risk-factors in order to identify high-risk individuals for screening. Reliable information needs to be recorded and this can be difficult as many individuals may not know their smoking history or information on other non-measurable predictor variables. Accurate ways to record reliable information on individual’s medical history is important and should be stressed when utilizing the PLCO\textsubscript{M2012} model or similar models.

Advocating the efficacy of the PLCO\textsubscript{M2012} model to health care providers is important. General practitioners, as well as researchers, can get educated through various means such as keeping updated with publications, attending conferences, workshops and different educational seminars. Hospital based research institutes can also be targeted; hospitals home many research programs with multidisciplinary experts, scientists and clinicians who can work side by side to evaluate the PLCO\textsubscript{M2012} risk prediction model in LC screening studies. This can aid in
convincing policymakers to adopt model-based risk assessment as a standard practice of LC screening eligibility.

Is the PLCO\textsubscript{M2012} Model Selecting Individuals who will die from Non-LC Causes?

Individuals selected for screening by the PLCO\textsubscript{M2012} risk prediction model are more likely to be older with comorbidities and screening these individuals is expected to lead to fewer life-years and QALYs gained.\textsuperscript{(185)} This suggests that individuals selected by the model might not live long enough to experience the benefits of early LC detection.

From observing our 6-year LC incidence and 5-year non-LC death models, and comparing the drop in AUC values, predictor variables used in the PLCO\textsubscript{M2012} model are associated with and predict non-LC death. This is important because for LC screening to be effective, patients must be healthy enough to undergo treatment of early-stage LC and must not be susceptible to competing causes of death that would substantially lessen the benefits of screening. If certain individuals possess traits that have been shown to be highly predictive of both LC incidence and non-LC mortality, they should be further evaluated for their risk of non-LC mortality before being given screening.

Due to the similar relative rank order of important predictors for LC and non-LC death (except for family history of LC) observed in this study, recommendations to modify or remove PLCO\textsubscript{M2012} model predictors to decrease the selection of individuals at high risk for non-LC death cannot be made. One option is to assess the ratio between the risk of dying from non-LC death over the risk of developing LC. A ratio such as this would require a model that accurately predicts non-LC mortality, which has not yet been developed and validated. To optimize predictive ability, such a model would need to include important predictors beyond those in the PLCO\textsubscript{M2012} model, and other variables which provide information on the severity of comorbidities or functional status of high-risk individuals.\textsuperscript{(192)} At the same time, using non-LC
mortality risk prediction models raises concerns, including the issue of investigating what threshold cut-off point should be used to determine when LC screening should not be offered. This raises the ethical question of whether it is reasonable to withhold LC screening from individuals at high risk for competing causes of death. Although models predicting non-LC death may help improve patient selection for LC screening, they may only deliver limited guidance; decisions on LC screening may come down to clinician and patient communication weighing the risks and benefits of screening. It is important that competing risks are recognized and properly examined among individuals selected for screening. Non-LC mortality risk prediction models can be used before patients are enrolled in LC screening programs and might be useful as they can make it easier for health care providers to understand and convey estimated risk of death from competing risks. Currently, to the best of our knowledge, there are no non-LC mortality risk prediction models developed which have been externally validated in different datasets and are being used in clinical practice. The development of such models can help determine the impact of competing risks on LC screening outcomes, as well as aid in the doctor-patient decision making process.

5.7 Conclusion

LC is a fatal disease with a high mortality rate when diagnosed at an advanced stage, therefore modelling-based selection of high-risk individuals for screening, as compared to NLST or NLST-like criteria-based selection, could lead to additional deaths prevented through early diagnoses. The PLCOM2012 risk prediction model performed significantly better than the NLST criteria in predicting 6-year LC risk with higher accuracy, and this translates into more efficient selection of high-risk individuals for LC screening. If implemented in the public health system, the PLCOM2012 model is expected to also decrease racial/ethnic healthcare disparities. Various populations should be investigated for LC incidence risk and incorporated into the model; as
mentioned, Canada has many Indigenous communities in various areas who might benefit from LC screening.

This study suggests that selecting individuals for LC screening using a stricter model risk threshold would be more cost-effective than the NLST criteria or the model at $\geq 1.5\%$ risk. This means that utilizing risk prediction models, such as the PLCO$\text{M2012}$, can make LC screening programs appear more appealing for policymakers. Considering the budget and resources available for LC screening are important for determining what risk threshold would be affordable to the healthcare system, while maximizing life-years gained and increasing quality-adjusted life years (QALY).

Our analysis showed that certain model predictors, such as age, smoking history and BMI predict non-LC death, suggesting that the PLCO$\text{M2012}$ model is increasing the selection of individuals at risk for non-LC death. However, the PLCO$\text{M2012}$ model is not predicting non-LC death to a greater extent than LC incidence. Changes to the PLCO$\text{M2012}$ model, by dropping or adding variables, to reduce the impact of CCoD cannot be made at this time given our existing data. In fact, risk prediction models, using additional predictors beyond the ones in the PLCO$\text{M2012}$ model, to predict non-LC mortality should be developed to aid in clinical decisions regarding if individuals with certain characteristics, which put them at risk of non-LC mortality, should undergo LC screening.

Future work is also needed to convince institutions to accept the model-based risk approach in selecting individuals for LC screening to save more lives from LC and to help guide clinical decision-making with patients who may not benefit from LC screening due to coexisting illnesses or risk factors which put them at risk of mortality.
References


53. Cancer Care Ontario. Lung Cancer Screening Pilot – Cancer Care Ontario [Internet]. 2017


80. Etzel CJ, Amos CI, Spitz MR. Risk for smoking-related cancer among relatives of lung


89. Neugut AI, Robinson E, Lee WC, Murray T, Karwoski K, Kutcher GJ. Lung cancer after


early lung cancer with low-dose spiral CT: prevalence in 817 asymptomatic smokers.


146. Marcus; DD. Causes of death among lung cancer patients in the national lung screening trial: competing causes of death as a hallmark of overdiagnosis. Bmj Evidenced-based


183. Steyerberg EW. Clinical Prediction Models A Practical Approach to Development,


192. Patricia Rivera M, Tanner NT, Silvestri GA, Detterbeck FC, Tammemägi MC, Young RP, et al. Incorporating coexisting chronic illness into decisions about patient selection for


