The University Of Sheffield.

VALIDATING AND UPDATING LUNG CANCER PREDICTION MODELS

Eoin Gray

Supervisors:
M. Dawn Teare
John Stevens

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

The University of Sheffield
Faculty of Medicine, Dentistry and Health
School of Health and Related Research (ScHARR)

January 19, 2018

Contents

Part I

Glossary, Abbreviations, Chapter Summary and Abstract

Acknowledgements

I would like to thank my supervisors Dr. Dawn Teare and Dr. John Stevens, for their guidance, direction and advice throughout my thesis.

I would like to thank the International Lung Cancer Consortium, for their continual support throughout and without whose data this project would not be possible.

I would like to thank the Roy Castle Lung Cancer Foundation who funded this project including national and international conferences, and publications on this research.

Finally I would like to thank Elizabeth Boyes and my parents, Patricia and Michael Gray, who have offered unconditional support throughout this project.

List of Figures

List of Tables

Publications

Gray, Eoin P. et al. "Risk Prediction Models For Lung Cancer: A Systematic Review". Clinical Lung Cancer 17.2 (2016): 95-106. Web.

Glossary

$\mathcal{A}$

Area Under the ROC Curve: The area under the ROC Curve (AUC or AUROC) *see: Receiving Operating Characteristic* summarises a medical classification tool by assessing all the pairs of individuals, one who is diseased and one who is disease free, to determine the probability of correctly assigning a higher risk or weighting to the individual with disease.

$\mathcal{B}$

$\mathcal{C}$

Calibration: The calibration is an assessment of how successfully the model predicts risks in individuals by comparing the predicted and observed incidence rates. This is commonly measured through the Hosmer-Lemeshow test.

Cancer: Cancer is a growth in a part of the body caused by uncontrolled cell division due to cell mutation.

Computer Tomography Scan: A computerized tomography (CT) scan is a series of X-ray images taken from different angles. These images are then combined to create detailed images of an organ in the body.

95% Confidence Interval: There is 95% confidence the true value for the measure is contained between the upper and lower bound of the interval.

Confounding: When measuring the association between a health outcome and a variable, confounding is an additional variable that is correlated to both the health outcome and the variable that may exaggerate the level of association.

$\mathcal{D}$

Discrimination: The discrimination is a measure for all participants in the dataset how successfully the model assigns a higher risk to cases in comparison to controls. This is commonly measured by the area under the receiver operating characteristic curve (AUROC or AUC).

$\mathcal{E}$

Endobronchial ultrasound: An endobronchial ultrasound is a medical procedure to diagnose lung cancer. It creates an image of the lungs through ultrasound during a bronchoscopy to identify any abnormalities.

$\mathcal{F}$

First Degree Relatives: First degrees relative to an individual are their parents, siblings, and children.

Five(5)-year survival rate: The 5-year survival is the percentage of participants that are alive 5 years after diagnosis for a disease.

$\mathcal{G}$

$\mathcal{H}$

$\mathcal{I}$

$\mathcal{J}$

$\mathcal{K}$

$\mathcal{L}$

Lobectomy: Lobectomy is surgical removal of an organ such as a lung.

Low Dose Computer Tomography: This is a CT scan (see Computer Tomography Scan) that uses less ionizing radiation. This is commonly used in lung cancer screening as Low Dose Computer Tomography (LDCT) reduces the risk of other forms of cancer from increased radiation exposure.

$\mathcal{M}$

$\mathcal{N}$

Non-Small Cell Lung Cancer: Non-Small Cell Lung Cancer (NSCLC) is one of the two main lung cancer subgroups; it is the most common type of cancer of lung occurring in approximately 87% of patients. There are three subgroups of NSCLC which are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma.

$\mathcal{O}$

$\mathcal{P}$

Pneumonectomy: Pneumonectomy is the surgical removal part of an organ such as a lung.

- ***Prediction Rules:*** The prediction rules are an assessment of the screening programmes screening criteria to identify high risk participants for screening. This can be simple criteria such as age, or using a prediction model and identifying high risk participants whose risk exceeds a specified value. There are many measures to evaluate the prediction rules which commonly evaluate the proportion of cases successfully classified as high risk and the proportion of controls that were classified as low risk.

$\mathcal{Q}$

Quality-adjusted life year: The quality-adjusted life year (QALY) is a measure of the improvement or lowering of the standard of living in a patient with a disease. The measure considered the quality and the duration of life lived with the condition to calculate a QALY. An alleviation of symptoms and prolonged life would improve a QALY score with one QALY equating to one year with no symptoms.

$\mathcal{R}$

Receiver Operating Characteristic: The receiver operating characteristic (ROC or ROC Curve) is a plot of the sensitivity and (1 - specificity) at each risk threshold. It measures the performance between these two measures and evaluates the ability of a model to distinguish between diseased and disease free groups. This is commonly presented alongside the AUC.

Risk Prediction Model: A risk prediction model estimates the probability of a condition presenting in an individual over a specified duration using predictors that are specific to the individual.

$\mathcal{S}$

Screening Programme: A screening programme is testing participants once or periodically to identify a disease or health outcome, such as lung cancer, in individuals who appear healthy. A selective screening programme uses a defined criteria, such as an age range, to identify a population subgroup that would most benefit from screening.

Sensitivity: The sensitivity is a measure of the performance for a selective screening criteria to identify high risk participants. The sensitivity is the proportion of cases for a disease that would be correctly classified as high risk if developing the disease and sent for screening, defined as the true positive rate.

Small Cell Lung Cancer: Small Cell Lung Cancer is one of the two main lung cancer subgroups occurring in approximately 12% of individuals. This is a more aggressive strain of cancer that spreads rapidly.

Smoking Pack Years: Pack years is a simple calculation to determine one' smoking history and can be calculated by $\frac{CigarettesPerDay}{20} \times SmokingDuration$

Specificity: The specificity is a measure of the performance for a selective screening criteria to identify high risk participants. The specificity is the proportion of non-cases for a disease that would be correctly classified as low risk of developing the disease and declined for screening, defined as the true negative rate.

$\mathcal{T}$

$\mathcal{U}$

$\mathcal{V}$

Validation: Validation is an assessment of a model, its ability to generate accurate risks (see calibration) and distinguish between cases and controls (see discrimination and prediction rules). A validation can be internal, in the same population as the model was created; temporal, in the same population the model was devised but participants excluded from the model building stage; or external, a new environment.

$\mathcal{W}$

$\mathcal{X}$

$\mathcal{Y}$

$\mathcal{Z}$

Abbreviations

AUC: Area Under the Receiver Operating Characteristic
COPD: Chronic Obstructive Pulmonary Disease
CPD: Cigarettes per Day
CT: Computer Tomography
BMA: Bayesian Model Averaging
EBUS: Endobronchial ultrasound
ELCAP: Early Lung Cancer Action Project
LDCT: Low-dose computed tomography
NELSON: Dutch-Belgian lung cancer screening trial
NLST: National Lung Screening Trial
NSCLC: Non-Small Cell Lung Cancer
SCLC: Small Cell Lung Cancer
TNM: Tumour, Lymph nodes, Metastasis staging scale
UKLS: United Kingdom Lung Screening
95% CI: 95% Confidence Interval

Chapter Summary

In Chapter One the lung cancer statistics are presented, which includes diagnosis and survival rates. The main risk factors associated with elevating or minimising lung cancer risk are presented. The process to identify, diagnose and stage is presented and how different stages influences the treatment options available to cure lung cancer. Finally, the chapter presents selective screening trials that have been implemented to improve early lung cancer diagnosis rates and reviews their success and limitations of the different screening trials.

Chapter Two explores why prediction models are developed to estimate the likelihood of a disease presenting in an individual within a time period. The chapter then focuses on how prediction models can be used to identify high risk individuals who may benefit from periodic screening. Prior to prediction models being considered to identify high risk individuals, there needs to be confidence that the model is reliable. Therefore, the models need to be rigorously evaluated to create confidence in their performance; the different methods available to evaluate a prediction model are presented.

Chapter Three conducts a systematic review of lung cancer prediction models. This identifies all published lung cancer prediction models and reviews how they performed in validation studies. This assesses the model's potential as a selective screening tool and identifies the models with most promise for further evaluation in this study.

Chapter Four presents datasets that were obtained from the International Lung Cancer Consortium. The raw datasets were collected and the chapter details how the data was prepared so lung cancer prediction models could be applied in the dataset. A review of methods to deal with or impute missing information was conducted and an appropriate method identified.

Chapter Five reviews the prepared datasets that were collected. This identified any specific design feature within the datasets during the participant recruitment, and how this could influence the prediction models that are applied and validated in the dataset.

In this first section of the validation of the lung cancer prediction model, Chapter Six validates and compares the models. The chapter analyses their potential to accurately estimate an individual's risks and assign a higher risk to individuals with lung cancer rather than disease free individuals. Additionally, the chapter reviews the model's prediction rules to assess how they would perform as a selective screening tool.

The second stage of the external validation is reported in Chapter Seven where the models are validated in a more restrictive population that is more representative for lung cancer screening. All the models are reviewed in this same target population to allow a direct comparison between the models. The prediction rules are evaluated to review how each of the models would perform as a selective screening tool.

Prediction models can be updated or multiple models aggregated to create a new model which may have an improved performance. Therefore, Chapter Eight identifies and reviews methods to update a prediction model and assesses how the methods performed when they have previously been implemented. The practicality of the methods for lung cancer prediction models is reviewed and appropriate methods are identified to be applied in a later chapter.

Chapter Nine introduces the lung cancer prediction models which will be updated. The original models that would be subsequently updated were validated in the dataset and the results presented. This will provide a baseline performance that we shall compare the updated models to, to assess whether the model updating methods create a model with an increased potential as a selective screening tool.

Chapter Ten applies methods to update a single prediction model. The updated models were presented

and validated to assess if an improved prediction model was devised. The chapter also assesses the potential and limitations of the different methods based on their performance when applied to the lung cancer prediction model.

Chapter Eleven employs methods to aggregate multiple prediction models. The new lung cancer prediction models are presented and validated to assess if an improved prediction model could be devised. The chapter provides an opportunity to review how the different methods perform, using lung cancer prediction models as an example, to identify robust methods that can subsequently be applied to new lung cancer prediction or different types of prediction models.

The thesis concludes with a discussion, this presented the key findings within the thesis and how this embedded in the current environment and previous research. Based on the results obtained recommendations were made on the future research that should be conducted that will hopefully propel lung cancer prediction models being used as a selective screening tool and aim to improve lung cancer diagnosis and survival rates. The limitations of the project were discussed but in context how the results obtained should be interpreted and how this can be rectified in future studies.

Thesis Abstract

Lung cancer is a global disease that affect millions of individuals worldwide [1]. Additionally, the disease is beset with a poor 5-year survival rate, a direct consequence of a low early stage diagnosis rate [1]. In an attempt to improve lung cancer prognosis, individuals at high risk of developing lung cancer should be identified for periodic screening.

Prediction models are devised to predict an individual's risk of developing a disease over a specified time period. These can be used to identify high risk individuals and be made publically available to allow individuals' to be conscience of their own risk. While prediction models have multiple uses it is imperative the models demonstrate a good standard of performance consistently when reviewed.

The project conducted a systematic review, analysing previously published lung cancer prediction models. The review identified that there had been inadequate reporting of the existing models and when these models have been validated this had not been consistent across different publications. As a consequence models have not been consistently considered as a selective screening tool.

The project then validated the prediction models using datasets from the International Lung Cancer Consortium. The validation identified the leading models which will allow a more targeted focus on these models in future research. This could culminate in the model being implemented as a clinical utility.

The final stage reviewed methods to update a single prediction model or aggregate multiple prediction models into a meta-model. A literature review identified and evaluated the different methods, discussing how different methods can be successful in different scenarios. The methods were also reviewed for their suitability updating selected lung cancer prediction models, and appropriate methods were identified. These were then applied to create updated lung cancer models which were validated to assess which methods were successful at improving the performance and robustness of lung cancer prediction models. As lung cancer research develops, particularly into researching genetic markers that may explain lung cancer risk, these factors could be incorporated into already successful prediction models using appropriate model updating methods that were identified in our research.

Part II

Thesis

CHAPTER 1

Introduction to Lung Cancer and Screening Programmes

1.1 Introduction

Lung cancer can be a deadly disease that affects people worldwide, accounting for one out of four cancer related mortalities [1]. This chapter provides an overview and context into lung cancer including statistics, disease progression and staging, the importance of early diagnosis, and screening strategies. This chapter will provide a background into lung cancer and evaluate the effectiveness of the current screening criteria to identify lung cancer in patients.

1.2 Objectives

To provide a thorough introduction on lung cancer this chapter will;

1. Present essential information and terminology;
    - Describe the key risk factors.
    - Present the stages of disease progression.
    - Discuss how lung cancer can be diagnosed.
    - Introduce the available treatment options for each stage of the disease.
2. Present key statistics including disease prevalence and survival rates stratified by lung cancer stage.
3. Present screening programmes that have been implemented for lung cancer;
    - Discuss why screening programmes have been implemented.
    - Present the screening programmes methodologies.
    - Summarise their effectiveness.
    - Discuss how screening programmes could be improved.
4. Identify how further research could improve early diagnosis and survival rates for lung cancer.

1.3 Introduction to Lung Cancer

Tumours present in a body when cells begin to uncontrollably divide, the tumour can be malignant and affect the surrounding tissue and organs and is then classified as cancer [2]. Cancer can occur throughout the body and is defined by the organ where the tumour originated. Lung cancer is a malignant tumour formed in the lungs or the cells that line the air passages.

There are two main types of lung cancer which are classified by the types of cells from which the cancer originates. The most common form of lung cancer is non-small cell, this occurs in 80% of cases, and can be further subdivided into squamous cell carcinoma, adenocarcinoma, or large-cell carcinoma. The second form, small-cell lung cancer, is less common but is more aggressive often spreading more rapidly across the lungs [3].

1.4 Risk Factors of Lung Cancer

There are many risk factors which elevate lung cancer risk; these can be lifestyle choices or environmental factors. The main risk factor is a positive smoking history and is present in around 85% of new lung cancer cases [4, 5]. A systematic review estimated the magnitude of an increased risk due to a smoking history, reporting an odds ratio of 8.96 (95% CI $[6.73, 12.1]$) for current smokers, and 3.85 (95% CI $[2.77, 5.34]$) in former smokers [4]. The statistically significant odds ratio demonstrates how a positive smoking history increases ones' risk of developing lung cancer; despite this 1 in 5 UK adults currently smoke [6] and there are 1 billion smokers worldwide [7].

Lung cancer risk is also increased from occupational exposures, and in the UK 20.5% of males with lung cancer and 4-5% of females reported a positive exposure [8]. A positive exposure can be linked to one of several types of harmful substances, including asbestos, wood, dust, and radon; all of which can elevate lung cancer risk. Asbestos exposure has been shown to increase risk of lung cancer and longer durations of exposure further elevates risk [4].

A family history of lung cancer in first degree relatives has also been linked to an increased risk of lung cancer. A meta-analysis of 40 studies reported an odds ratio of 1.72 (95% CI $[1.56, 1.88]$) in people who had at least one first degree relative diagnosed with lung cancer [4]. This suggests lung cancer can be influenced by genetic factors that could prevent or increase susceptibility of developing lung cancer. Further research is being conducted to identify the genetic markers associated with lung cancer including specific blood-based biomarkers such as pro-surfactant protein B [9], micro RNAs [10] - [13], and cytokinesis-blocked micronucleus assay [14].

Prior lung diseases have also been confirmed to report an elevated risk. These diseases can damage the cells in the lungs and cell mutations and tumours can develop as the body replaces the damaged cells. The main lung diseases; pneumonia, chronic bronchitis, Chronic Obstructive Pulmonary Disease (COPD), emphysema, and tuberculosis were all associated with lung cancer risk when evaluated in a meta-analysis [15]. This was supported by additional research which showed a significant elevated risk for chronic bronchitis with an odds ratio of 1.47 (95% CI $[1.29, 1.68]$), tuberculosis (1.48 (95% CI $[1.17, 1.87]$)), and pneumonia (1.57 (95% CI $[1.22, 2.01]$)) [16]. The study also demonstrated prior lung conditions elevated risk independently of tobacco use, with never-smokers also reporting a significantly increased risk of developing lung cancer [16]. However, reverse causality should be considered as lung cancer can weaken the immune system and increase the chance of pneumonia and tuberculosis infections [15].

Finally, lung cancer risk is associated with an increased age [17] with lung cancer rarely presenting in patients under 40. Although there could be confounding, as while lung cancer increases with age this could also be linked to a developing smoking history present in 85% of new lung cancer cases.

1.5 Diagnosing and Staging Lung Cancer

Diagnosing and staging lung cancer is important; determining the lung cancer stage allows physicians to make an informed decision about the available treatment options. Staging determines how far the disease has spread and the size of the affected area. Cancers are categorised into four different stages (Stage 1 – Stage 4) with a higher staging indicating that the disease has progressed further and affects a larger area. A tumour, lymph nodes, metastasis (TNM) staging scale is used for lung cancer [18]. This categorises the cancer by the size of the tumour, whether the cancer has spread to the surrounding lymph nodes, and

whether the tumour has metastasized to other parts of the body [19]. A detailed summary how the TNM scale can stage lung cancer is provided on-line [18].

There are a variety of methods available to diagnose and stage lung cancer. A CT scan is the most common test [20], and is often conducted as the initial assessment to ascertain whether a tumour is visible on the lungs. A bronchoscopy is another available method that examines the inside of the airways by placing a narrow, flexible tube down the throat [20]. This is an invasive procedure that can affect the patients and they will be required to not eat or drink after the procedure until the anaesthetic has worn off. Endobronchial ultrasound (EBUS) can also be used to diagnose lung cancer; this is like a bronchoscopy with an ultrasound. The probe creates ultrasound pictures of the lung tissue and lymph glands to determine the cancer stage [20]. Finally, a percutaneous lung biopsy can be used which requires a thick needle injected through the skin and muscle of the chest to extract a cell sample [20] or staged through keyhole surgery. However, these tests are invasive and cause discomfort to the patient.

1.6 Lung Cancer Treatment

Treatment options for lung cancer depend on the type of lung cancer, the stage at diagnosis, and the health of the patient. Early diagnosis means more treatment options are available such as surgery. However, as the cancer spreads too much of the lung would need to be removed, making this a non-viable option. The health of the patient is also important as risky procedures or major surgery may not be conducted if the patient is deemed too unhealthy. Removing a proportion of the lung would be inadvisable in patients with respiratory problems.

Diagnosing non-small cell lung cancer at stage 1 will normally result in surgery removing part of the lung (a lobectomy) or the entire lung (a pneumonectomy) [21]. By stage 3, if scans show the cancer cells have spread to the middle of the chest and close to the heart, then radiotherapy would be recommended instead of surgery as the cancer is too close to vital organs to operate safely [21]. By a stage 4 treatment options aim to control the cancer for as long as possible and attempt to shrink the cancer to alleviate symptoms [21].

For the less common small-cell lung cancer chemotherapy and radiotherapy of the lung are the most likely treatment options. This form of cancer can spread to the brain so radiotherapy of the brain is also conducted. While surgery could be performed for small-cell lung cancer the aggressive nature of this strain means the cancer has often spread to an advanced stage so this is not a viable option [21]. Therefore most treatment options aim to improve quality of life and alleviate symptoms.

1.7 Lung Cancer Statistics

Lung cancer is a global concern. In the US in 2016 there will be an estimated 224,390 new cases of lung cancer and 158,080 mortalities [1]. In the UK there are approximately 44,500 people diagnosed each year [22]. Lung cancer is one of the leading cancers with the second highest incidence rates for both males and females; behind prostate cancer in males and breast cancer in females [1]. However, it is the leading cause of deaths from cancer and accounts for approximately 26% of all cancer related death in both genders [1]. As a leading cause in cancer incidence and mortality rates there is a greater focus in attempting to improve the poor lung cancer prognosis.

Lung cancer risk increases with age with few incidences in individuals under 40. Cancer will develop in 0.7% of individuals aged 50 – 59 years but this increases to 2% in 60 – 69 years and 6.4% in people over 70 of years [1]. The increased age at diagnosis may also contribute to the high mortality rates as some treatment options, such as surgery, may not be practical due to poorer health and respiratory difficulties.

Lung cancer is commonly diagnosed at a later stage possibly because the major symptoms are masked by a smoking history and no external symptoms to alert the patient. Therefore, patients do not report symptoms to their physicians until more sinister symptoms develop at later stages. Only 16% of lung cancers are diagnosed at stage 1, 27% at stage 2, and 57% are diagnosed at stage 3 or 4 [1]. The lack of

early diagnosis rates are a major concern as late stage diagnosis for lung cancer has one of the poorest survival rates for all cancers except liver and pancreatic cancer. The 5-year survival rate in patients whose lung cancer is diagnosed at Stage 3 or 4 is only 4% [1], in contrast the 5-year survival for a stage 1 diagnosis is 56% [1]. The low survival rate for late diagnosis is due to the limited treatment options available. Clearly, lung cancer would benefit from screening programmes to improve early diagnosis rates. The evidence supports improving early stage diagnosis rates which will improve survival rates and lower mortality rates as more patients have extensive treatment options available. Unfortunately, despite the compelling evidence this has not been observed and 5-year survival has only improved from 12% to 18% between 1975 and 2011 [1].

1.8 Lung Cancer Screening Programmes

To improve early diagnosis rates for lung cancer several screening trials have been assessed. These studies identified high risk populations, using different criteria, and the participants were periodically screened. There have been four main lung cancer screening trials, namely the United Kingdom Lung Screening (UKLS), National Lung Screening Trial (NLST), Dutch-Belgian lung cancer screening trial (NELSON), and Early Lung Cancer Action Project (ELCAP). The studies periodically screened high risk participants using low-dose computed tomography (LDCT) because of its improved ability in comparison to x-rays to detect small lung tumours at an earlier stage [23, 24]. However, it is imperative only participants who would benefit from screening are selected, as with LDCT there can be anxiety and a fear of radiation exposure, which lowered the willingness of patients to be screened [25].

The lung cancer screening trials are introduced.

1.8.1 NLST Screening Programme

The NLST programme evaluated whether LDCT, in comparison to x-rays, could diagnose cancer at an early stage and reduce lung cancer mortality in high risk individuals [26]. If the NLST programme demonstrated a cost-effective screening regimen that detected a high proportion of individuals with lung cancer while limiting unnecessary screening then the trial would be considered for implementation on a national scale funded by the government and managed care organizations [27].

In the NLST programme high risk participants were identified using an age and smoking history criteria; participants were considered eligible for screening if they were current smokers aged 55-74 years with a minimum 30 pack year smoking history; former smokers were also included if they met the same criteria and had quit smoking within the previous 15 years [26]. The eligible high risk participants were randomised with 26,722 selected for LDCT and another 26,732 selected for chest radiography and all participants received annual screening for 3 years [26].

The trial reported a high positive screening rate with 39.1% of participants in the LDCT arm and 16% in the radiography arm requiring further testing after screening reported abnormalities [26]. Unfortunately, a high proportion of the additional testing was not necessary with 96.4% of participants in the LDCT group and 94.5% in the radiography group having a false positive result [26]. This resulted in 7% in the LDCT arm and 4% in the radiography arm having unnecessary invasive follow-up treatment [28]. The trial did demonstrate the improvement using LDCT to screen participants in preference to the chest radiography; indeed the LDCT group significantly lowered the lung cancer mortality rate by 20% (95% CI $[6.8, 26.7]$) in comparison to the radiography group [26]. There have been cost-effectiveness analysis studies of the NLST programme and the results varied between studies. The cost-effectiveness of the study ranged from a promising \$2,500 per life-year saved to a less promising \$269,000 per quality-adjusted life-year (QALY) saved [27]. The results are quite variable because of variability in the volume of unnecessary disease free individuals screened. The results vary across the US $\$100,000$ to US $\$160,000$ per QALY that is relatively deemed acceptable [29, 30]. The NLST criteria reported a reduced lung cancer mortality, with a 10-year survival between $18-25\%$ [27] which is a large improvement upon the current 5-year survival rate. Studies

indicate that the NLST programme would need to screen 320 patients using LDCT to identify one lung cancer [27], while this is a high proportion of unsuccessful screening this improves upon an approximate yearly prevalence rate of 0.1% for lung cancer [31]. Although, improving screening technology with quality assurance [32] may reduce unnecessary follow-up screening of disease free individuals improving the lung cancer capture rate and cost-effectiveness of the screening programme.

The trial showed annual LDCT screening of high risk groups could be beneficial [33]. Based on the NLST results the "American Cancer Society (ACS), the American College of Chest Physicians (ACCP), the American Society of Clinical Oncology (ASCO), and the National Comprehensive Cancer Network have now updated their clinical practice guidelines so that those who qualify for screening using the NLST criteria may be eligible for annual screening" [27]. However, since the NLST results reported limitations in some studies there are arguments "to maximize screening efficiency, results from more robust lung cancer prediction models, which include variables beyond age and smoking history, may be helpful to guide selection of patients for screening" [27].

1.8.2 UKLS Screening Programme

The UKLS programme evaluated LDCT to screen a population who were at high risk of developing lung cancer [34]. The trial assessed whether lung cancer screening could be successfully implemented to improve lung cancer diagnosis rates. Participants were screened using LDCT as evidence indicated this was the leading procedure to identify early stage lung cancer [35].

The high risk participants were aged 50 to 75 years and had a risk exceeding 5% of developing lung cancer in the next 5 years based on the Liverpool Lung Project Model [34]. This was the only trial that considered additional factors other than age and smoking history to identify their high risk population for screening. The study randomised half the participants to receive LDCT screening in comparison to a control arm who did not receive screening [34]. The UKLS study initially considered 88,897 participants which was reduced to 2,848 (12%) who were classified as high risk from the prediction model [34].

The trial screened 1,994 participants annually for 2 years, with 42 participants (2.1%) correctly diagnosed with lung cancer [36]. This is a higher diagnosis rate than the approximate UK lung cancer prevalence rate [31] demonstrating that the trial successfully targeted screening to a high risk population. Further, 28 (66.7%) participants were detected as being at Stage I and 8 (19%) at Stage II [36], a dramatic improvement on the current early diagnosis rates for lung cancer. This allowed 91.6% of the early stage diagnosed participants to receive surgery [36]. The trial demonstrated some promising results costing £8466 per QALY gained [36] with net treatment costs per person of £60 [36] These results were promising in the context of US \$100,000 to US \$100,000 per QALY [29, 30], and the cost of treatment is reasonable below US \$500 [37]. This highlights the fact lung cancer screening can be beneficial, cost-effective, and affordable if targeted to the correct population. Although, the success of a trial needs to be reproducible because an increase in unnecessary screening of disease free individuals will reduce the effectiveness of the screening programme [38].

The screening programme demonstrated the potential of prediction models to identify a high risk population for screening.

1.8.3 NELSON Screening Programme

The NELSON study selected participants for LDCT screening if they were male, aged 50-75 years, and were former-smokers who had either quit smoking within the previous 10 years after smoking more than 15 cigarettes a day for at least 25 years, or were current smokers who smoked more than 10 cigarettes daily for over 30 years [40]. These high risk participants were screened at baseline and after 1, 2, 4, and 6 years [40].

The criteria identified 7,557 participants for screening. At the baseline screening, 196 (2.6%) of participants had a positive result although 64.3% of these were determined to be a false positive [40]. The trial is currently ongoing, with 3 years remaining in the study, with current results reporting the lung cancer

detection rate was 0.9% [40], an improvement in comparison to a blanket screening approach where lung cancer has a 0.1% prevalence rate.

1.8.4 ELCAP Screening Programme

The ELCAP programme was designed to evaluate the effectiveness of annual LDCT screening in comparison to the current diagnosis procedures where a high proportion of lung cancers are diagnosed at an advanced stage [41]. The study was targeted to individuals over 60 years with a minimum 10 pack year smoking history [41] who received one round of LDCT screening.

A true positive lung cancer diagnosis was found in 2.7% (27 of 1,000) of eligible participants [41], an improvement upon the capture rate in a blanket screening approach. The study found that annual screening in the target population detected over 80% of lung cancers at Stage I [42, 43] in comparison to over 70% commonly diagnosed at an advanced stage in current diagnosis treatment [41]. The study concluded that LDCT could substantially improve early detection of lung cancer, decreasing mortality rates, provided the correct target population was identified [41]. The screening programme increased survival by 0.1 years at a cost of approximately $230 per increase based on costs for treatment and screening to identify lung cancer [41]. The cost per life-year saved for one round of screening in the sample population was $2,500 [41]. These are reasonable results based on evidence suggesting screening programmes are effective if the cost is limited to US $100,000 to US $160,000 per QALY [29, 30]. The study also determined the approximate costs for treatment after diagnosis in USA would be $20,100, $23,000, $31,800-$32,700, and $25,900 for each increasing stage at diagnosis [41] demonstrating how an earlier diagnosis will lower treatment cost. Stage IV showed a slight decrease but this is because most treatment options are just to alleviate symptoms at this advanced stage. Although, these results are not favourable as research indicates a cost exceeding $500 can be problematic for patients [37]. However, these may not be a concern to patients in countries with free or partially covered healthcare.

The study concluded that a single LDCT scan for high risk participants can be a cost-effective procedure that is "likely to be within the range of practice and policy acceptability".

1.9 Improving Screening Programmes

The independent screening programmes unanimously agree that LDCT is the preferred method to identify early stage lung cancer. Identified participants were screened annually or biennial if multiple screenings were conducted as this afforded the highest probability of capturing lung cancer in its infancy. The screening programmes identified a high risk population to target resources because a blanket screening approach was deemed impractical. All the screening trials restricted the population by age, with no-one under 55 being screened. Smoking history was also a key marker for the majority of screening trials; screening heavy smoking ever-smokers. The UKLS trial utilised the Liverpool Lung Project Model to identify high risk participants.

The screening programmes displayed a higher prevalence rate in the high risk populations than observed in the global population. This allows the screening resources to be targeted to a population that would most benefit from periodic screening. However, there were conflicting results over the cost incurred in the screening programmes with results ranging from $2,500 per life-year saved which was a promising result. Other studies reported though the screening programmes required $269,000 per QALY. The results are variable and across the estimated US $100,000 to US $160,000 per QALY that is acceptable for a screening programme [29, 30]. The variability in the results can be attributed to the volume of disease free individuals that would be screened, and considered for follow-up screening to identify one lung cancer incidence [38]. This can be evaluated by reviewing the screening guidelines in multiple sample populations. It can also be improved by reducing unnecessary screening through better testing. Improved quality assurance of CT scans may reduce the need for annual follow-ups [32] which will improve the ratio of disease free individuals screened to identify one lung cancer. Reducing unnecessary screening is crucial as there can be uncertainty

surrounding the cost of screening [39], hence improving the cost-effectiveness of screening programmes. Based on the results screening programmes are being implemented with the NLST criteria identifying participants for annual screening [27].

There is the potential to further improve screening programmes by reducing costs from unnecessary screening and increasing true positive rates. This can be achieved by identifying a more appropriate population for screening. Eligible participants could be identified by a prediction model, as seen in the UKLS programme. These could identify participants at higher risk of developing lung cancer by considering more in-depth patient information rather than merely an age and smoking history criteria.

1.10 Summary

Lung cancer is one of the leading cancers worldwide with 224,390 estimated new cases in 2016 in the United States alone [1]. Lung cancer has a causal association with a smoking history, with 85% of new incidences occurring in ever-smokers [4, 5]. Despite this smoking is still popular and there are over 1 billion smokers worldwide [7]. Additionally, lung cancer diagnosis has a bleak survival rate; with a 17.7% 5-year survival rate [1]. However, early identification increases the potential for curative treatment and improves survival rates; with a 56% 5-year survival rate for diagnoses at Stage I [1]. Early diagnosis is problematic, with no visual symptoms and common markers such as coughing, breathlessness, and chest pains common in the ever-smokers and older population. Therefore, people do not normally report the symptoms until more severe conditions, associated with later stages, develop. As a result different methods need to be incorporated to capture early stage lung cancer.

Screening programmes can improve diagnosis rates by identifying high risk participants for annual screening. The NLST and UKLS are two examples of lung cancer screening trials. While these trials have successfully identified lung cancer in individuals, the cost-effectiveness, assessed by QALYs, of these trials have been inconclusive [27] and vary either side of the acceptable screening and treatment costs per QALY. Blanket screening for lung cancer is an unrealistic solution so a more robust criteria to identify a high risk population for screening should be developed. A criteria that improves early diagnosis and true positive rates while reducing unnecessary screening could be universally applied and improve available treatment options and survival rates.

Prediction models could be used as a screening tool provided they demonstrate an ability to identify a more appropriate population for screening. In the next chapter we introduce prediction models and how a model's ability as a selective screening tool should be evaluated.

CHAPTER 2

Introduction to Prediction Modelling

2.1 Introduction

Prediction models estimate the risk of an outcome occurring within a specified time period. Prediction models can be used to predict the likelihood of lung cancer presenting in an individual. These can be used to identify the highest risk individuals who may benefit from screening, which may improve lung cancer diagnosis and survival rates. Before a prediction model can be used by individuals and physicians the model needs to be validated. This will assess the model performance and provide guidelines as to how the model can be applied to identify a target population for screening.

The chapter will introduce prediction modelling and model validations. This will introduce the most commonly used tests that assess the properties of a model, how the results of an assessment can be interpreted, and identify which performance measures should be considered when validating a model. These key tests will be encountered in previous validations of lung cancer prediction models and will be used in future testing of prediction models in this project.

2.2 Objectives

To introduce prediction models and validation methodology the chapter will;

1. Introduce prediction modelling;
    - Objectives of cancer prediction models.
    - Applications of cancer prediction models.
    - Requirements for a successful lung cancer prediction model.
2. Present the techniques to validate a prediction model;
    - Introduce the different validation tests available, presenting the methodology and how to interpret the results.
    - Identify the key tests that should be conducted in a validation study.

2.3 Introducing Prediction Models and Their Applications

Risk prediction models predict the likelihood of an outcome occurring within a time period for an individual with a particular predictor profile [44]. Prediction models are developed, in many cases, to guide healthcare professionals and individuals to make informed decisions and determine the best course of actions, based on the likelihood of the disease occurring [45, 46].

A major objective of prediction models is to identify a target population which can then be screened or tested periodically. Prior to a selective screening criteria, based upon a prediction model being implemented

as a public health tool, the selective criteria is required to demonstrate a good performance in validations and trials. A model that has the potential to be a good selective screening tool would be expected to robustly identify a target population for screening. This can be achieved by identifying a high proportion of diseased individuals while limiting unnecessary screening of healthy individuals [47]. Then the screening criteria will allow more benefits, from identifying true positives, than harms caused by false positives [44].

Indeed, in the UK, screening for a cancer will only be considered if a screening programme can demonstrate a good performance by reliably detecting cancer, not causing too many false alarms, and being cost-effective [47]. The demanding requirements has seen only three cancer screening programmes being implemented in the UK; for bowel, breast, and cervix cancer. These screening criteria choose participants based solely on age and gender; bowel cancer tests anyone aged 60-74 years, and breast and cervix cancer target females aged 50-70 years and 25-49 years respectively [47]. The participants are then invited for screening every 2-3 years [47]. A similar screening criteria has not been implemented for lung cancer; indicating a blanket screening approach for everyone within a certain age bracket is not an appropriate approach, despite most lung cancer cases occurring between 50-75 years [1]. Indeed, there is no current lung cancer screening due to a "lack of a sensitive enough test", a low proportion of cancers being identified, and the high costs incurred [47]. However, costs and potential harms, such as radiation exposure, could be reduced as new innovative tests are being designed. These include clinical trials to develop a non-invasive breath test to detect lung cancer [48]. While these tests are being developed, current lung cancer screening research is evaluating if it is possible to identify people at high risk of developing lung cancer by considering people who smoked or have contracted lung diseases such as COPD [47]. A prediction model could combine these risk factors and additional measures to identify an even more appropriate target population for screening which satisfies the UK screening guidelines.

Prediction models can also be used as an on-line risk calculator. This can then be used by the public and physicians to evaluate an individual's risk. Risk calculators have been created for many diseases and was pioneered by the Gail Model for breast cancer [49]. Clinicians can incorporate the risk assessment into routine screening [50]. An advantage of a risk calculator is the public become more aware of their personal risk, which may prompt positive lifestyle changes to reduce their risk. The greater awareness can also improve early stage disease detection rates as individuals report any relevant, concerning symptoms to a physician earlier. Additionally, patients and physicians can open up a dialogue and make informed decisions based on the risk estimate [51] which may include more frequent check-ups to monitor their risk or to undergo screening.

While a promising prediction model can be beneficial a poor model that assigned inaccurate risks could be problematic. It could cause disease free individuals to be unnecessarily screened and cause undue panic. To determine the reliability of a model, which will identify both useful and unreliable models, they need to be thoroughly evaluated. The model's predictive ability should be evaluated in many different validation studies to assess how it will perform as a risk calculator or selective screening tool in a variety of situations. Alongside thorough testing, all potential models or screening criteria (whether age, smoking history or previous lung diseases) should be compared in the same study [52]. This will allow an informed decision about the leading selective screening criteria which identifies the most appropriate target population. The different tests for model validation will now be presented, including how the results are interpreted, and the key measures that should be considered for a thorough validation will be identified.

2.4 Introduction to Prediction Model Validations

The purpose of a validation study is to evaluate the performance of a prediction model [44]. The validation evaluates a model's ability to assign accurate risks to individuals and whether the model can successfully assign a higher risk to diseased rather than disease free individuals. The results are commonly presented through the *calibration* which evaluates the accuracy of risk predictions, *discrimination* which evaluates the model's ability to consistently assign higher risks to participants with the disease than without the condition, and *prediction rules* which assess the model's performance at a specific risk threshold while

considering individuals with a predicted risk exceeding this threshold as high risk or as a target population for screening.

There are different forms of validations that can be conducted; internal (using the original data in which the model was devised), temporal (data from a new sample of the model building population), or external (a new dataset with a different population). An internal validation is commonly conducted when a new model is published. This will allow an indication of how the model will perform in practice. However, the results of an internal validation may suggest the model performs to a higher standard than observed in new populations as the model is being reviewed using the same population as the model was developed [53]. This is defined as optimism [144]. There are techniques available to attempt to eliminate optimism, in internal validations, with bootstrapping, jackknife, and cross-validation the most common methods [55]. A temporal validation considers participants who were excluded from the model building; however, these participants often have a very similar patient profile as the model building participants. This is due to the collected participants being grouped on certain conditions, whether age, exposure outcomes, or smoking history. Therefore, there is still a chance of optimism in the temporal validation and excluding these participants from model building could hinder the model as their evidence is not considered when building the model. External validations offer the most reliable and robust indication into a model performance. Here, the model is tested in a distinct population and if a model can consistently demonstrate a promising performance in multiple external validations then it could be considered to aid medical professionals.

2.5 Model Calibration

Model calibration is a commonly reported measure to assess how well a model predicts the observed incidence rate in the dataset [58]. This evaluates the extent to which a model predicts appropriate risks for individuals. The calibration can be measured differently and can be shown in a calibration plot by "plotting the observed proportions of events against the predicted probabilities for groups defined by ranges of predicted risk" [142, 60]. The participants are grouped as the outcome is dichotomous, the outcome is present or absent, so grouping participants based on their predicted risks for a set group size will allow the observed rate to be continuous between $[0, 1]$. A perfect calibration would be observed on the plot as a "$y = x$" line which indicates the observed incidence rates for each group was accurately predicted.

The most commonly reported method to assess model calibration is the Hosmer-Lemeshow Goodness-of-Fit test [61]. This generates a p-value to evaluate the null hypothesis "the model does accurately fit the observed results". If the p-value is small it gives evidence to reject the null hypothesis. The Hosmer-Lemeshow test is calculated for the number of groups, K, observed and expected incidence rates in each group, O and E, number of participants in each group, N, and degrees of freedom, $K - 2$, as follows;

$$Hosmer = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i \times (1 - \frac{E_i}{N_i})} \tag{2.1}$$

There are some concerns using the Hosmer-Lemeshow test. The calibration results can vary depending which size groups and how many groups are used to combine participants. The group size can be arbitrarily chosen however this can lead to a large change in the Hosmer-Lemeshow p-value [56]. It is important the correct group size is determined as a large volume of participants per group may not detect poor predictions generated by the model, whereas too few participants per group makes it difficult to determine whether the differences between the observed and expected risks are by chance or if the model is poorly calibrated [56]. Studies have recommended when applying the Hosmer-Lemeshow test the participants should be grouped such that there are 10 participants in each group, therefore there will be $N/10$ total groups [56]. If the null hypothesis is borderline rejected or accepted at p-value 0.05 then groups of similar sizes should be considered. This will assess the trend as the p-values changes as the group sizes changes to determine if the model has a good calibration. This approach will be taken when applying the Hosmer-Lemeshow test in any validations.

The model calibration is the key measure when assessing the ability of the model. If the model consistently reports a good calibration then there can be confidence the model would not generate misleading risks and could be considered as a public tool.

2.6 Brier Score

The Brier score is an additional measure to assess the model calibration [57]. This measures the squared differences between the estimated risk and the binary outcome whether the disease is present or absent in the participant. The score is calculated by;

$$\frac{1}{N}\sum_{i=1}^{N}(o_i - e_i)^2 \tag{2.2}$$

where N is the number of participants in the dataset with an observed outcome 'o' which is either 0 for a participant without the disease or 1 when the condition is present; and the estimated risk of developing the disease 'e'. A smaller Brier score indicates a better calibrated model with a perfect model reporting a Brier score of 0. A score of 0.25 would indicate a non-informative model in a population with 50% incidence [57].

There are some limitations to the Brier score. Unlike the Hosmer-Lemeshow test the Brier score requires personal judgement on whether the model has a good calibration. Therefore it is not the preferred test to validate model calibration and infrequently reported. However, it has an advantage over the Hosmer-Lemeshow test as there are no requirements for the participants to be grouped so there are no concerns about variable results. Additionally, it is a good comparison between multiple models evaluated in the same population.

In validations conducted in this project the Brier score will be reported as it can offer a good comparison between multiple models. This measure will be particularly useful if there are variable Hosmer-Lemeshow results for different size groups and it is unclear which model has the best calibration.

2.7 Model Discrimination

Discrimination is a measure of the model's "ability to distinguish between patients with and without the outcome" [58]. A successful model will assign a higher risk to the majority of participants with the disease rather than disease free participants.

Discrimination can be presented on a plot of the sensitivity, the proportion of individuals diagnosed with lung cancer with a higher risk than a specific risk threshold, plotted against $(1\text{–}specificity)$, where specificity is the proportion of disease free individuals with an estimated risk lower than the considered risk. This is conducted at every risk threshold. The plot is defined as the Receiver Operating Characteristic Curve. Discrimination is measured by the Area Under the Receiver Operating Characteristic Curve (AUC or AURoC). The AUC result will be a value between 0.5 – 1.0; with a higher result indicating a stronger discrimination. A perfect result of 1 would mean all individuals with the outcome were assigned a higher risk than all the disease free individuals. Whereas a result of 0.5 would indicate the model offered no better than chance at assigning a higher risk to participants with the disease than without the disease; this would be an ineffective model to discriminate between individuals.

For dichotomous outcomes, such as with or without lung cancer, the AUC is equivalent to the concordance index. The concordance index, also referred to as the c-index, is a measure of the proportion of all the possible pairs (one participant with or without the disease) where the estimated risk of disease is higher for the individual with the condition [62].

Discrimination will indicate whether the model could be a successful screening tool and is a good comparison between models. A result exceeding 0.8 indicates a model that could be a promising selective screening tool [51] although slightly lower results $0.70-0.80$ may also indicate a successful model. However,

the AUC alone is not beneficial to medical professionals as while it demonstrates the potential of a model it does not indicate how the model should be applied as a selective screening tool. A selective screening criteria using a prediction model needs a fixed risk threshold where everyone exceeding this risk would be high risk participants considered for screening. To assess this, prediction rules have to be considered.

The AUC is an important measure that should be reported, as it is a good comparison between models and an overview on how they distinguish between diseased and disease free individuals.

2.8 Prediction Rules

Prediction rules are a series of tests to assess a model's ability at a defined risk threshold. This reviews whether the model robustly identifies a high risk population that would be benefit from screening. At the specified risk threshold any participants with a risk exceeding this threshold would be considered a target population for screening.

The prediction rules can also be evaluated for additional selective screening criteria. Participants satisfying the criteria would be a person of interest for screening. Therefore, this allows the prediction rules to be compared between prediction models and alternative selective screening criteria so an informed decision can be made into which would be a better approach to identify high risk participants.

There are many different measures to report the prediction rules which are now presented.

2.8.1 Sensitivity and Specificity

A key measure of performance are the sensitivity and specificity rates. At the pre-specified threshold the sensitivity is the proportion of participants with the disease that have a risk exceeding this threshold [62], and the specificity is the proportion of participants without the disease that have a risk lower than the threshold [62]. Intuitively, as the selected risk threshold is increased, a large volume of diseased and disease free individuals would have a risk below the threshold; this would lower the sensitivity but increase the specificity. The converse also holds if a lower risk threshold is considered.

Clearly a higher performing model would have both a high sensitivity and specificity. Unfortunately, the performances of these two measures trade-off each other so a risk threshold needs to be determined that is a useful compromise between both measures. Alternatively, a sensitivity or specificity rate can be predetermined and the risk threshold selected that allows this predetermined rate. Then the leading model or criteria would have the highest performance for the other measure.

The sensitivity and specificity should be reported in a validation study to offer a review of the prediction rules.

2.8.2 Youden 'J' Index

The Youden Index or J Index is a simple but effective measure to evaluate the relationship between the sensitivity and specificity [62]. The test combines these two values into a single measure as follows;

$$J = senstivity + specificity - 1 \tag{2.3}$$

This generates a result between 0 "if the test reports the same proportion of positive tests for both control and diseased groups" [63] and 1. A higher result means a higher combined sensitivity and specificity; this indicates a better performance. The Youden Index assumes an equal weight between capturing a participant with the disease as excluding a participant without the disease. In some instances it may be more beneficial to capture participants with the disease or exclude disease free participants. There is no measure that considers an unequal weighting for the sensitivity and specificity rates as this would be down to the medical professionals own judgement.

For a model to be used as a selective screening tool an optimal or very good risk threshold should be identified as here the model would have the greatest benefit. The Youden Index can be a good indication

of where the model has this optimal performance. This simple measure will be considered in the validation of the models to identify optimal risk thresholds and compare between models.

2.8.3 Positive and Negative Predictive Values

The Positive and Negative Predictive Values (PPV/NPV) can also be presented when evaluating the prediction rules. The PPV measures the percentage of true positives (participants with the disease) amongst everyone assigned a risk above the specified threshold. The NPV measures the percentage of true negatives (disease free participants) for all participants below the threshold. These values are calculated as;

$$PPV = \frac{NumberofTruePositives}{NumberofTruePositives + NumberofFalsePositives} \tag{2.4}$$

$$NPV = \frac{NumberofTrueNegatives}{NumberofTrueNegatives + NumberofFalseNegatives} \tag{2.5}$$

These values, like sensitivity and specificity, are commonly reported together and a successful prediction rule would offer a good balance between the two measures.

The difficulty with these measures is finding an optimal performance between the PPV and NPV as they trade-off each other. A viable solution for comparison between screening criteria would be to fix one measure at a predetermined rate and assess which screening criteria would offer the highest result for the other measure. However, it is unlikely when considering a model as a selective screening tool, that medical professionals would require a predetermined PPV or NPV rate. Additionally, the PPV and NPV rates can be misleading as they are dependent on the prevalence rate in the dataset [64] and unless tested in a population cohort dataset the results may be unrealistic. This is not an appropriate measure in a case-control dataset as there is a higher prevalence rate observed. Additionally, the changing prevalence rates across the datasets makes comparisons between datasets difficult. Finally, the measures cannot be combined so they are unable to be utilised to identify an optimal risk threshold for a model or compare between models.

Overall, the PPV and NPV can be useful measures but should not be selected over sensitivity and specificity as these rates offer a clearer indication into the model performance for health decision makers considering a model for a selective screening tool. The measures would be most effective if either the PPV or NPV were restricted to a specific level; then a comparison of the other measure can be conducted between models. Unfortunately, determining an appropriate level is difficult and the results can be misleading by high prevalence rates in datasets which would not be replicated in real populations. For these reasons, the PPV and NPV rates will not be considered as key measures in any validations conducted in this project.

2.8.4 Positive and Negative Likelihood Ratios

The positive and negative likelihood ratios (PLR/NLR) combine the sensitivity and specificity rates and can be calculated by;

$$PLR = \frac{Sensitivity}{1 - Specificity} \tag{2.6}$$

$$NLR = \frac{1 - Sensitivity}{Specificity} \tag{2.7}$$

$$\tag{2.8}$$

The PLR is the ratio of the probability being considered high risk (above the threshold) when with the disease versus the probability of being considered high risk when without disease. This informs as to how

much more likely a participant considered high risk has the disease [65]. The NLR is a measure of how much more likely a participant considered low risk is disease free [65].

The ratios can be calculated in case-control datasets where there is a higher incidence than cohort studies as the sensitivity and specificity measures remain unaffected. This means the likelihood ratios can be compared between case-control and cohort studies as they are unaffected by prevalence rates.

The PLR and NLR can be important measures for clinicians as a model that has a higher probability of identifying diseased individuals for screening would be preferred. Indeed a higher PLR would indicate a better model. However, this result should not be evaluated alone and is required alongside the sensitivity and specificity results. The PLR may be misleading, if not reported with the sensitivity and specificity, as a prediction rule may have a higher value by excluding the majority of disease free participants from screening but only identifying a few diseased individuals. The NLR is not an essential measure for clinicians as this does not measure the benefit against the potential harms of the trial.

In validations conducted in this study the PLR will be reported as this informs how much more likely a participant considered to be high risk has the disease. This is a good measure of the benefits of capturing cases against the harms of screening controls if implementing a model as a screening tool.

2.8.5 Model Accuracy

The model accuracy evaluates what proportion of participants were correctly categorised; individuals with the disease are correctly assigned a risk exceeding the threshold and disease free individuals categorised below this risk threshold. The model accuracy is measured by;

$$ModelAccuracy = \frac{TruePositiveCount + TrueNegativeCount}{NumberofParticipants} \tag{2.9}$$

The result will range from $[0, 1]$; with a higher performing model achieving a result closer to 1. This can provide a good comparison between models or criteria as physicians would desire a higher model accuracy which would indicate the screening criteria is targeting a more appropriate population.

However, the model accuracy can be a misleading statistic as it is influenced by the disease prevalence rate which can lead to distorted results [66]. The more prevalent the disease the higher weighting is applied to the sensitivity, conversely a lower prevalence rate will see the model excelling at a higher specificity. In a case-control dataset the high prevalence rate will be reflected in the model performing well at a lower risk threshold where the model accuracy is skewed by the high sensitivity which counter balances a poor specificity. This may not be observed if the model is applied in real populations due to the lower prevalence rate. Additionally, the recommendation to apply the model where a high sensitivity is observed will translate into high screening rates of disease free individuals and the screening criteria may not be economically viable.

The model accuracy will not be considered when conducting validations. This is because it will add to an already large volume of statistics to analyse the same prediction rule, the results can be misleading, and it is not an essential measure for physicians selecting a screening tool.

2.8.6 Net Benefit Ratio

The Net-Benefit Ratio evaluates the relationship between the true positive and false positive rates at the risk threshold. This test considers, somewhat simplistically, the importance of a correct diagnosis over the potential harm of false positives. The net benefit ratio, at a risk threshold w, is calculated by;

$$Ratio = \frac{TruePositive - (w \times FalsePositive)}{NumberOfParticipants} \tag{2.10}$$

Here a risk threshold of 20%, resulting in a w score of 0.2, means identifying a true positive is five times more important than a false positive. A result exceeding zero is a good net benefit ratio, with a higher result indicating a more successful model.

The net benefit ratio is rarely presented as the other tests provide clearer information about the proportion of diseased participants screened and disease free participants rejected. Additionally, the importance of a true positive in comparison to a false positive seems arbitrarily chosen by the risk threshold whereas physicians may have a defined acceptable volume of false positives screened to identify a true positive. Once a specific w can be determined then this would be a very useful measure; however since this has not been determined for lung cancer then the model accuracy will not be considered in validations.

The difficulty with this measure lies in determining a good trade-off between the sensitivity and specificity. There is often not a clear rule of acceptable sensitivities for different specificity results, which would allow an appropriate w to be determined. It is often much clearer to determine an acceptable rate for the sensitivity (high enough for the model to have an impact in identifying the disease) or specificity (high enough such that the screening is economically viable). Once one of these are determined and fixed, then the other measure can be reported, which can infer whether the model would be successful and a comparison between models.

2.8.7 Net Reclassification Index

The Net-Reclassification Index (NRI) is a measure between the performances of two different selective screening criteria. One criteria offers the baseline performance and the NRI evaluates if the second criteria offers an improvement. The NRI evaluates how many participants would be correctly or incorrectly reassigned as high or low risk. The reclassification of every participant for 2 different criteria can be observed in Table 2.1 with green values representing a good reclassification and red a poor reclassification for Criteria 2 in comparison to Criteria 1.

	Criteria 1	**Criteria 2**	
		$<$ Risk Threshold	$\geq$ Risk Threshold
Disease Free	$<$ Risk Threshold	x_{111}	x_{112}
	$\geq$ Risk Threshold	x_{121}	x_{122}
With Disease	$<$ Risk Threshold	x_{211}	x_{212}
	$\geq$ Risk Threshold	x_{221}	x_{222}

Table 2.1: Reclassification Table between Two Models at Specified Risk Thresholds

The NRI can then be calculated from the reclassification table;

$$NRI = \left(\frac{x_{121} - x_{112}}{NumberDiseaseFree}\right) + \left(\frac{x_{212} - x_{221}}{NumberWithDisease}\right) \tag{2.11}$$

The NRI results are presented in terms of the improvement criteria 2 demonstrates over criteria 1, and the result will range from $[-1, 1]$. A result greater than 0 indicates criteria 2 had an improved model accuracy and would correctly reclassify more participants than incorrectly reclassify them. A result below 0 indicates the inverse. As the result tends closer to 1 or -1 the reclassification change occurs in more participants.

While this is a good measure to compare between two screening criteria there are better measures to assess a models ability as a selective screening tool. Additionally, it is more insightful for physicians to understand the model's performance in terms of true positives, false positives, and overall benefit rather than a comparison between two different criteria. Therefore, the NRI will not be considered in any subsequent validation studies.

2.9 Summary

The chapter introduced prediction models and how they can be used to predict an individual's risk or as a selective screening tool. These can be beneficial; creating awareness of ones' risk can promote lifestyle changes to reduce risk and a selective screening tool can help identify high risk participants that would benefit from lung cancer screening.

Before models can be considered for implementation they need to be validated. This will give a strong indication into the model's potential for medical decision makers who can then make an informed choice on whether to utilise the model. There are many different measures to validate a prediction model and the key measures have been identified. Calibration, measured through the Hosmer-Lemeshow test and Brier Score, is the key measure to determine if a model is to be a good individual risk calculator. This quantifies how successfully the model predicts the observed outcome rates. The AUC is an effective measure for the model discrimination. This gives an indication into how well the model assigns a higher weight to diseased individuals than disease free individuals for every pair in the dataset. This is commonly reported and is a good measure for model comparison. The prediction rules should be tested to assess how the model would perform as a selective screening tool. These measure the model's performance at a specific risk threshold; participants with a higher risk than the threshold would be considered for screening. The key measures identified were the sensitivity (percentage of cases identified), specificity (percentage of controls rejected), Youden Index, and PLR; these are commonly reported and essential measures to evaluate the prediction rules. They allow the medical professionals to determine how a model would perform as a screening tool. Evaluating prediction models with the identified validation tests would allow an informed decision into which model would be the most effective screening tool. These measures were also selected as they are not influenced by the prevalence rate; evaluating the models in case-control datasets may lead to misleading results which would not be reflected in implemented selective screening trials.

The identified measures; calibration, discrimination, sensitivity, specificity, Youden Index, and PLR; will be encountered in the systematic review of the lung cancer prediction models and used when validating prediction models in this project.

Now the measures to evaluate a prediction model have been identified a systematic review of published lung cancer prediction models and their performance in validations will be conducted. This will evaluate the models' calibration, discrimination, and prediction rules, indicating how successful they predict an individual's risk or identify a high risk population for selective screening.

CHAPTER 3

Systematic Review of Lung Cancer Prediction Models

3.1 Introduction

The review will identify all published lung cancer prediction models and their validation results. This can then assess the performance of models if they were implemented as a selective screening tool and if there is a leading selective screening criteria that could be used by medical professionals. The review will offer recommendations into how leading models should be implemented in selective screening trials based on synthesising their evidence from multiple published validations. The review will also assess the current standard of validations for lung cancer prediction models; highlighting any concerns with their reporting and identifying how the validations should be improved to allow a clearer understanding of the models' performance.

The systematic review outline was presented at the 2014 and 2015 International Lung Cancer Consortium (ILCCO) conferences and published in an article [67].

3.2 Objectives

The systematic review chapter objectives are:

1. Present the systematic review search criteria.
2. Conduct a systematic review of published lung cancer prediction models.
3. Identify and present all published lung cancer prediction models and their performance in validations.
4. Present how the prediction models have been devised, the purpose of the models and any limitations with the model design.
5. Synthesise all the evidence as presented in validation studies of prediction models to evaluate their predictive ability.
6. Assess limitations with the current reporting of lung cancer prediction models and how these limitations can be addressed.
7. Provide recommendations into how leading models could be optimally used in selective screening trials.

3.3 Search Criteria

The key search terms and search strategy used to conduct a comprehensive systematic review are presented.

3.3.1 Search Terminology

To identify all published lung cancer prediction models and validation studies in the systematic review key search terminology was defined. The relevant articles were identified through a combination of:

1. Lung cancer terminology
2. Prediction model terminology
3. Known prediction model titles

The lung cancer terminology was identified using a medical thesaurus; the Medical Subject Headings (MeSH) database [68].

Next prediction model terminology was identified. The key terms were identified through the most commonly used words and phrases in a key lung cancer prediction model validation article. The article was the "Comparison of discriminatory power and accuracy of three lung cancer risk models" by AM D'Amelio Jr et. al [69]. The most common terms in the article were identified using the TerMine database [70, 71]. More than 500 search terms were returned and those were then reduced to the highest and most relevant terms for prediction models.

The final search term was the names of known models. Incorporating the model titles into the search assisted in identifying any validation study that evaluated the named model.

The exhaustive list of search terms is presented;

Lung terminology	**Prediction model terms**	**Model titles**
Lc	Model	Bach
Cancer of the Lung	Risk Model	Liverpool Lung Project
Lung Cancer	Risk	LLP
Neoplasms, Lung	Cancer risk	Spitz
Neoplasms, Pulmonary	Incidence rate	Two-Stage Clonal Expansion
Pulmonary Cancer	Risk prediction	TSCE
Pulmonary Neoplasms	Risk-prediction model	
Lung tumours	Absolute risk	
Carcinoma, Non Small Cell	Risk estimate	
Non-Small Cell Lung Cancer		
Non-Small Cell Lung Carcinoma		
Nonsmall Cell Lung Cancer		
Carcinoma Oat Cell		
Oat Cell Carcinoma		
Small Cell Carcinoma		

Table 3.1: Search Terms for Systematic Review

Using the above terms the following combinations were searched using a whole text search in the online databases with an **OR** separating all terms under the same heading and an **AND** between different headings.

1. Lung terminology **and** Prediction model terms.
2. Model titles **and** Lung terminology.
3. Model titles **and** Prediction model terms.
4. Lung terminology **and** Prediction model terms **and** Model titles.

3.3.2 Electronic Databases

Upon identifying the key search terms, the appropriate databases to search for the articles were determined. Lung cancer prediction models are likely to be published in medical (cancer and lung cancer) journals so an exhaustive search of the follow medical databases was conducted;

1. Applied Social Sciences Index and Abstracts (ASSIA) via ProQuest
2. CINAHL via EBSCO
3. Cochrane library
4. Medline via WoS
5. Web of Science

3.3.3 Hand Searching of Key Journals

The next stage searched key medical and lung cancer journals. This aimed to identify recent publications or publications in press that would not be identified in the electronic databases. PubMed PubRefiner was used to identify the key journals with recent publications for a specific search term. "Lung Cancer Prediction Models" was searched and the journals with the highest count of relevant publications since 2013 were included;

1. Annals of Thoracic Surgery
2. BMC Cancer Journal
3. British Journal of Cancer
4. Clinical Lung Cancer Journal
5. International Journal of Radiation Oncology
6. Lung Cancer Journal
7. PLoS One

The journal's issues between 01/01/2013 and 01/07/2014 were manually searched which coincided with the time frame of the review. Early view contents were also searched to identify any relevant articles in press.

3.3.4 Grey Literature

A grey literature search was conducted in Google Scholar searching articles from 01/01/2013 till 01/07/2014 to identify new publications that may not be returned when searching the electronic databases.

3.3.5 Bibliography and Citation Search

After the inclusion criteria (Section 3.3.7) identified the relevant articles for the systematic review, a citation and bibliography search was conducted on each article. This aimed to identify additional articles that may have been missed in the original search. The bibliography search on validation studies would identify the referenced original model if previously undiscovered. Additionally, citation searching on the original model article would assist in identifying all validation studies that include the model.

3.3.6 Author Contact

Author contact was not conducted to identify articles as completed models and validations would be successfully published. Instead the objective of our systematic review were presented at international conferences, including ILCCO, to allow authors who were in the process of publishing new models or validation studies relevant to our research to contact us to be included in the systematic review. However, any articles published close to the project completion on 01/10/2106 would only be presented as a footnote of review.

3.3.7 Inclusion Criteria

Once the searches were conducted any article was included provided it satisfied all of the following criteria;

1. A model was proposed or validated that can be used to predict the risk of lung cancer incidence or absolute risk (considering competing risk of death) at an individual level.
2. The full article was obtained in English.
3. The paper was published since 1985.
4. A human population was considered to create or validate a model.
5. The model created/validated does not only incorporate patients already targeted for primary care using symptomatic predictors or a survival model for post-diagnosis.

For studies that ***only*** validate a prediction model and ***do not*** propose a unique model the following criteria will also have to be satisfied;

6. The article provided statistical results that assess at least one published model's predictive power.
7. A case-control or cohort study was used to validate the published lung cancer prediction model.

3.3.8 Outcomes Measured for Review Objectives

The objective of the systematic review was to present how the model was devised; review the model building population, the purpose of the model and a summary of the variables included in the model.

Additionally, all validation results for the lung cancer prediction model were presented. The validations evaluated the model calibration, discrimination and prediction rules and the most common measures were presented in Section 2.4 and included in summary tables.

3.3.9 Data Collection

For each stage of the search the article collection and exclusion process was detailed.

3.3.9.1 Storage

References for the articles were stored using EndNote. This allowed articles to be excluded from the systematic review in accordance with the inclusion criteria. A safety backup database was maintained while the review is being conducted between 01/07/2014 to 01/10/2015.

3.3.9.2 Study Selection

Studies were collected if they satisfied the inclusion criteria.

3.3.10 Data Analysis

The model design and all validation results will be reported through tables and discussion. The results were analysed to review how successful the model has been when validated, the model performance in comparison to other models and how successful the model would be as a selective screening decision tool. This allowed an informed decision into the leading models and the potential of published lung cancer prediction models to be used by physicians.

3.3.10.1 Study Quality

The study quality was not used to include or exclude papers from the systematic review. However, this is an integral part of reviewing how lung cancer prediction models have been developed and validated. Studies included in the systematic review were critically appraised with the population size and study design examined. This critiqued the population upon which the model was devised and the credibility of the dataset used to validate the prediction models. These were discussed for each model and the study quality was presented in tables.

3.3.10.2 Data Extraction

Data extraction on the model design and performance results was reported and discussed for each model. A series of tables throughout the review present the key information to allow comparisons of the model designs and validation results.

3.3.10.3 Synthesis – Meta Analysis

A meta-analysis was not conducted in this review. All of the results were presented separately and reviewed as there was no advantage to synthesising their evidence for the objectives of the systematic review.

3.4 Search Results

The electronic and hand searching stage of the systematic review was completed by 01/07/2014. The electronic search found 1,365 papers which were subsequently reduced to 66 articles after excluding duplicates and irrelevant articles based on the title and abstract. The remaining 66 articles were assessed at full article stage. 16 articles were not available in full because they were conference abstracts or letters to editors and another 24 did not publish or validate a lung cancer prediction model. A further 5 articles were rejected as they considered patients after diagnosis or symptomatic patients in primary care. Hand searching key journals discovered 1 additional article and the bibliography and citation search on the current 22 articles identified 4 new papers. The 66 articles reviewed at the full article stage were independently replicated by supervisor Dawn Teare in December 2014, with total agreement, in preparation for the systematic review article.

1. Epidemiological Models - basic information
2. Epidemiological and Clinical Factors Models - require some form of testing
3. TSCE Models - only consider one risk factor to measure association levels.

All models across the three categories were considered and reported in the review.

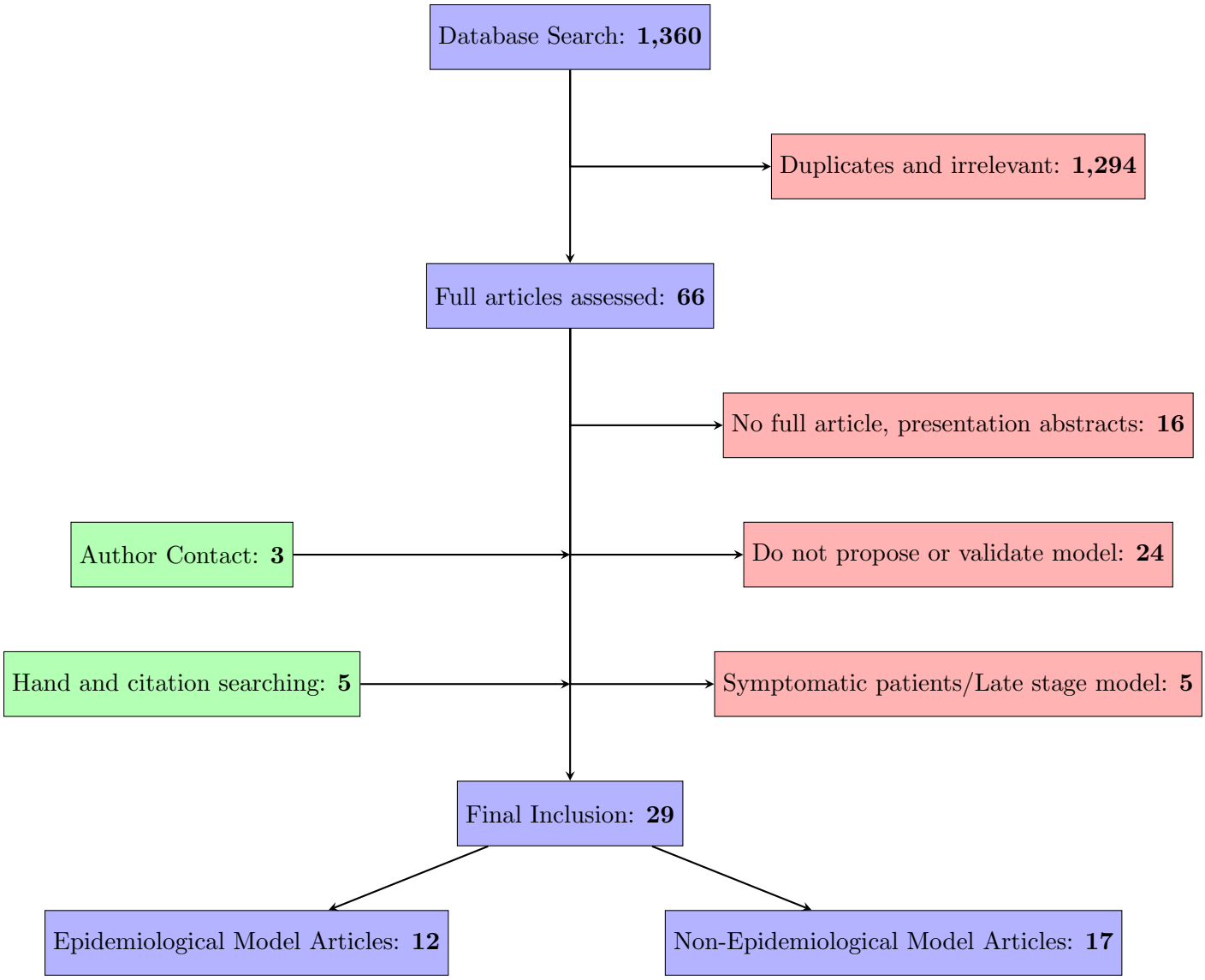

Figure 3.1: The Systematic Review Results Process

This allowed new models that were not published to be included if they were published before 01/10/2015. This identified the $PLCO_{M2014}$ Model [82] paper published in December 2014 and the Pittsburgh Predictor [85] published in March 2015. It also allowed one additional validation paper to be included in the study [83]. After completing the search strategy, 29 articles were included in the systematic review.

The systematic review was completed by 01/10/2015 and any subsequent articles were not included. If any additional papers were published after this date but before the expected submission of 01/10/2016 they were included as a footnote to the project.

The articles were subdivided into epidemiological and non-epidemiological (clinical and Two Stage Clonal Expansion) models. Epidemiological models were defined as models that used only readily available information such as age, gender and smoking history. Non-epidemiological required a clinical assessment such as CT scan or blood test. The final set of non-epidemiological models are Two Stage Clonal Expansion (TSCE) models, these commonly use one risk factor to determine incidence rates in a population but can be applied at an individual level. The models and validations are presented in these three different groups.

3.5 Epidemiological Lung Cancer Prediction Models

There were 12 epidemiological model articles identified, including 9 papers that published original models. The identified epidemiology models are the Bach, Liverpool Lung Project, Spitz, African-American, PLCO, $PLCO_{M2012}$, $PLCO_{M2014}$, Hoggart and Pittsburgh Predictor models.

These models are presented separately including their validation results. The model summaries and validation results are also present in 4 tables; these present the model variables, building dataset description, model results and a summary of model validation tests conducted.

3.5.1 Liverpool Lung Project Model

The Liverpool Lung Project (LLP) (Cassidy et al 2008) model was published in 2007 to estimate the risk of lung cancer incidence occurring in an individual within the next five years. Risk of developing lung cancer was modelled by a multivariate conditional logistic regression. The model was devised in a UK case-control population matched by age and gender. The model uses UK population specific lung cancer incidence rates stratified by age and gender in the prediction model. This could limit the model's success in different countries where the incidence rates may differ. However, considering the population specific age and gender incidence rates aims to limit concerns with matching over these variables. The model also considered diagnosis of a prior malignant tumour, smoking duration, pneumonia, asbestos exposure and family history of lung cancer. The model is only applicable to people aged 40-80 years as the incidence rates stratified by age and gender were only provided for this age range. The model is applicable to all people in this age range; whether they are never- or ever-smokers [72].

The model was internally validated in the article using a 10-fold cross validation in an attempt to eliminate optimism [72]. The internal validation reported an AUC of 0.71 which was a reasonable discrimination. The internal validation also evaluated the prediction rules. These were reviewed at 2.5% and 6% risk thresholds and the sensitivity and specificity were reported. At 2.5% threshold the sensitivity and specificity results were 0.62 and 0.70 respectively. These results are promising, which was expected as the model reported a good overall discriminative ability, highlighted by the AUC result. At the 6% threshold the sensitivity was 34% while the specificity was 90%. This would allow the prediction model, if implemented as a selective screening tool, to remove a high proportion of lung cancer free individuals from unnecessary screening; a benefit if the screening programme objective was to reduce unnecessary screening costs. The ability to still capture 34% of individuals with lung cancer while limiting screening of disease free individuals is an advantage of the model. However, the results should be verified in an external validation to allow a stronger understanding. Additionally, validating the model in a case-control population often leads to enhanced validation results as the cases are commonly at an advanced stage and have an enhanced risk.

The LLP model was validated in five distinct datasets and the results showed a similar level of performance to the internal validation. One external validation [69] evaluated the model performance using 4,900 case-control study participants in a US sample population. This provided an opportunity to evaluate how the model performed in a distinct population while considering UK age and gender specific lung cancer incidence rates. The AUC had a slightly poorer performance at 0.69 in comparison to the internal validation; however, the model still performed well when considering the prediction rules. The sensitivity and specificity rates were reported at the 2.5%, 5% and 7.5% thresholds for 5-year risk. The model reported similar results at the 2.5% risk threshold as the internal validation, with a sensitivity and specificity both of 0.67. This represents a good trade-off between the sensitivity and specificity highlighted by a Youden's (J index) score of 0.33 in comparison to 0.32 in the internal validation. At 5% risk, the sensitivity dramatically reduced to 45.5% meaning over half of all individuals with lung cancer would not be screened and the J index reduced to 0.304. This trend continued at 7.5%, where a very low sensitivity of 31.2% was reported, although the specificity exceeded 90%, which resulted in a J Index of 0.235. The external validation demonstrates that at the low 2.5% threshold the model would be able to capture a high proportion of individuals who develop lung cancer, which may improve diagnosis and survival rates for lung

cancer. In contrast, if a selective screening programme had a limited budget, then a high risk threshold of 6% (internal validation) or 7.5% would allow the model to avoid screening over 90% of controls.

The LLP model was validated in another article [73]; where the model was validated in three large datasets. The studies collected were case-control and prospective population cohorts that were based across Europe and North America. The results varied between studies but suggested that the LLP model has strong potential as a selective screening criteria. The three AUC results were 0.67, 0.76 and an impressive 0.82 in the large population-based prospective study cohort of 7,652 patients. However, this was a Liverpool population similar to which the model was devised. In this cohort the sensitivity and specificity rates at the 2.5% threshold were 74.3% and 67.4% respectively, which combine to give the highest recorded J-Index for the LLP Model at 0.42; a very strong result. The model also reported a model accuracy of 67.8% in the cohort. The results were very promising and the model's ability to perform to a high standard in a population cohort demonstrates its potential for a selective screening trial. The model reported good results in the other studies although they were not as high, which are presented in Table 3.3. The difference between the three datasets indicate the model cannot perform to the same standard in populations outside of Liverpool or the UK.

One final external validation was conducted [74]. This used a case-control study with 1,275 participants and reported the sensitivity and specificity rates. The results were slightly poorer and were validated at the 5.12% risk threshold. Here, the model reported a sensitivity of 49.9% but a specificity of 79.8%. The calibration was not formally reported but the model recorded a "moderate overall calibration and improved accuracy at higher values of predicted risks" [74].

Based on the success of the LLP Model in validation studies the model has been evaluated as a selective screening tool in a screening trial. The trial is reported in Section 1.8.2. Participants were considered to be at high risk and recommended for screening, if their risk exceeded 5%. At this risk threshold the model would maintain a specificity of approximately 45-50% for a specificity around 75-80% based on reported validated results. This level of performance would result in the LLP Model reporting a Youden's Index 0.25, which is a good result.

Across studies, the AUC was reasonable with internal and external results ranging from 0.67 to 0.76 indicating a good discrimination. This was surpassed by an AUC of 0.82 in one validation, although the validation and model building data were both samples from Liverpool populations where the UK age and gender specific incidence rates are relevant. A slightly lower performance was observed in validations conducted outside of the UK. Based upon these findings the model should be tested in non UK cohorts to assess how successfully these can be used in different populations. Additionally, the calibration has not been adequately tested and needs to be more comprehensively assessed. At this stage it is clear that the model has potential but there is room for improvement and more thorough validations to assess the calibration and differences between UK and non-UK populations.

3.5.2 Spitz Model

The Spitz Model (Spitz et al 2007) predicts one-year absolute risk of lung cancer. The logistic regression model calculates an individual's likelihood of developing lung cancer within one year and the competing risk of mortality without a lung cancer diagnosis. The model is applicable to anyone; never- and ever-smokers. While individuals are required to be at least 20 years of age, lung cancer rarely presents at younger ages so this is not a limitation of the model. Therefore, versatility of the model to be applicable to everyone older than 20 years of age is a key advantage.

The model was devised with 3,852 case-control participants matched by age, sex and smoking status collected from the USA [75]. To avoid concerns with matching; age and gender lung cancer incidence and death rates were incorporated into the model by using the SEER rates. The SEER rates are the observed incidence rates across the USA. Smoking status is incorporated into the model using gender-smoking incidence groups.

The dataset was split into a model building set (75%) and an external validation set (25%) and three distinct models were devised for never, former and current smokers. These consider different variables in

different logistic regression models. Ideally, the model building dataset would include all the participants to allow as much evidence as possible to be included when devising the model. Then the Spitz Model could be validated internally and subsequently assessed in external validations.

The calibration and discrimination were measured in the external validation in the original article [75]. The validation results were presented separately for the never-, former- and current-smoking models. For the never-smokers model the model reported a good calibration with a Hosmer-Lemeshow p-value of 0.777. However, the never-smoker model reported a poor AUC result of 0.57 (95% CI [0.47, 0.66]) which includes 0.5; suggesting the model has no better than chance in assigning a higher risk to an individual with lung cancer than an individual who is disease free. The results slightly improved in the former-smokers model with a good calibration (p-value 0.712) and an improved AUC of 0.63 (95% CI [0.58, 0.69]). However, the AUC result does not suggest the model is a robust tool to distinguish between individuals with or without lung cancer. Finally, the current-smokers model was also well calibrated (p-value 0.688) but reported another poor AUC of 0.58 (95% CI [0.52, 0.64]). Across the three distinct models for different smoking statuses there was a reoccurring pattern. The models were well calibrated indicating that the model could be used to provide accurate risks for individuals. However, the model reported a very poor discriminative ability suggesting the model would have a limited ability to identify high risk participants for screening. However, this could be more extensively validated by evaluating the prediction rules, which was not conducted in the article in which the Spitz Model was presented.

The Spitz Model was externally validated [69] but the model was extended to predict absolute risk over 5-years by "combining the risk of cancer from the relative risk model with age and gender incidence and mortality models recursively five times" [69]. In the 4,900 case-control population the model reported a much improved discrimination with an AUC of 0.69 (95% CI [0.66, 0.71]). This would suggest extending the Spitz Model duration may improve the model's discriminative ability as it can more robustly distinguish between individuals with and without lung cancer. This could be due to the original 1-year time period being too small as the majority of individuals, both diseased and disease free, have low risks of developing cancer over such a small time period.

Despite the improved discriminative ability, the prediction rules results were surprising. The results at the 2.5% risk threshold suggest this was too high with the majority of the participants being eliminated as shown by a low sensitivity of 26.6% and a specificity rate of 94.4%. Decision makers would need to decide whether the low sensitivity is beneficial as a trade-off of eliminating a high proportion of unnecessary screening with the high specificity. To allow an informed decision the prediction rules should be evaluated at a lower risk threshold. The PPV at the 2.5% risk threshold was 0.882 and the NPV was 0.45, this is a very good PPV, although a case-control population favourably reflects these results due to the high incidence rates in the validation set that would not be observed if implementing the model as a selective screening tool.

The final external validation of the Spitz model used 1,340 patients in another case-control study [76]. This study reverted back to applying the Spitz Model to predict the absolute one-year risk. The discrimination was measured and the model reported AUC values of 0.67 and 0.68 for former- and current-smokers respectively, which indicates a reasonable discriminative ability. Unfortunately, the prediction rules were not assessed in the paper.

Current reporting for the Spitz Model has been subpar with only the AUC results being consistently reported. Indeed, the optimal duration for the prediction model has not been determined at this stage with validations alternating between applying the model to predict 1 or 5-year absolute risk. The AUC results for the Spitz Model varied from very poor to reasonable but further testing needs to be conducted. This includes reviewing the prediction rules, to offer a more comprehensive review of the model's potential, which currently have only been evaluated in one article. The wide applicability is a clear strength of the model, which would suggest if a robust optimal risk threshold can be identified then the Spitz Model may be considered to identify a target population for screening. Additionally, the model was well calibrated, when validated and based upon these results the model could be made available to the public to calculate their risk and increase awareness.

Variables	Bach	LLP	Spitz	Afr.-Amer.	PLCO	$PLCO_{12/14}$	Hoggart	Pittsburgh
Personal Information								
Age	X	X	X	X	X	X	X	X
Gender	X	X	X	X				
Ethnicity						X		
Body Mass Index					X	X		
Education					X	X		
Prior Malignant Tumour		X				X		
X-Rays					X			
Smoking History								
Smoking Status			X	X	X	X	X	X
Start Age							X	
Cessation Age			X	X				
Smoking Duration	X	X			X	X	X	X
CPD	X					X	X	X
Pack Years			X	X	X			
Quit Duration	X			X	X	X		
ETS			X					
Family History of Cancer								
Cases of Smoking Related Cancer			X					
Cases of Lung Cancer		X	X		X	X		
Age of Onset of Lung Cancer		X						
Exposures and Conditions								
Asbestos Exposure	X	X	X					
Dust			X	X				
Hay fever			X	X				
Emphysema			X					
COPD				X	X	X		
Pneumonia		X		X				
The following models have restrictions and designs;								
Applicable to Never Smokers		X	X	X	X	M2014		
Applicable to Ever Smokers	X	X	X	X	X	X	X	X
Age	50 - 75	40 - 80	20+	20+			35+	
Smoking History	30 PY							
Predicts Incidence	X	X			X	X	X	X
Predicts Survival	X		X	X			X	
Risk Length (Years)	1+	5	1+	5	9	6	1+	6

Table 3.2: Epidemiological Model Variables and Restrictions

3.5.3 Bach Model

The Bach Model (Bach et al 2003) was devised using a large 18,172 cohort collected in the USA [77]. The model estimates an individual's absolute risk by combining two separate one-year logistic regression models that calculate "risk of developing lung cancer" and the competing "risk of dying without a lung cancer diagnosis". The model duration can be extended by running the models recursively and the model was designed to predict absolute risk for 10 years [77].

The Bach Model is quite restrictive and can only be applied to a target population aged 45–69 years who are ever- smokers with a minimum 30 pack year smoking history and former-smokers are required to have quit within the last 15 years [77]. The restrictive population lowers the model utility as the 15% of lung cancer incidences that occur in never-smokers [4, 5] and further incidences that occur in lower smoking ever-smokers will not be identified. This lowers the impact of the model as an early detection tool as a sizeable proportion of lung cancers would not be identified in a selective screening trial.

The Bach Model was validated in a 300 participant cohort from the Mayo Clinic, USA in the original article. The Mayo Clinic entries satisfied the original criteria as "subjects must be aged 55–74 years, have smoked a minimum of 30 pack-years and be current smokers or former smokers who quit within the last 15 years" [77]. The validation compared the estimated one-year absolute risk and observed incidences over the duration. The study found the participants with the lowest 25% of risk contained only 8% of incidences whilst the top risk quartile contained 50%. While this demonstrates the model's ability to assign higher risk to individuals who develop lung cancer the risk thresholds at these quartiles were not reported which may indicate how the model could be applied as a selective screening tool. The AUC result (0.72) supported the suggestion the model had a reasonable discriminative ability. Unfortunately, no tests were conducted to evaluate the prediction rules and the model's clinical utility.

The Bach Model was externally validated in "Validation of a Model of Lung Cancer Risk Prediction among Smokers" [78] with 6,239 smokers recruited by the Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC) Study. The study recruited some participants outside the range specified in the original article; men aged "50-69 who had smoked 5 or more cigarettes a day" [78]. This tested the model's adaptability in different environments, in this case for lower ever-smokers. The expected incidences in the cohort over 10 years was 297.07 in comparison to the observed 333. This equates to expecting 89% of the observed incidences. The under estimation could indicate the model struggled to estimate risk in the lower risk participants with a reduced smoking history. The validation study only estimates total group deaths rather than formally assessing the calibration, discrimination and prediction rules. The study could be improved by splitting the validation to assess performance for patients inside and outside the original entry criteria to determine if the model's target population could be expanded.

Finally, the model was externally validated in an article that compared the Bach, LLP and Spitz models [69]. The Bach Model predicted absolute risk over 5 years in the study. The study included a large range of participants but non-smokers were excluded since their risk is not calculable by the model. The AUC was 0.66, indicating a reasonable discriminative ability and the study reported no significant difference in the AUC when stratified between age groups. However, this was the lowest AUC of the models considered in the same dataset (Spitz and LLP). This was reflected in the sensitivity and specificity rates with a highest J-Index of only 0.19 at the 2.5% risk. The poor results could indicate that the model has difficulty in predicting risk in less intense smoking populations and when applied for a shorter time frame than the 10 years original proposed.

In summary, the Bach Model is limited by only considering ever-smokers. The current validation testing has been poor; the discrimination, calibration and prediction rules have been infrequently recorded so it is difficult to judge the model's performance. The preliminary testing indicates that the Bach Model may not be as robust a selective screening tool as the LLP and Spitz models. Further testing is required to thoroughly evaluate the model; this should assess if the model can demonstrate an improved discriminative ability and the prediction rules should be more extensively evaluated to provide insight into the Bach Model's clinical utility.

Model	Article	Validation Type	Sample Size	H-L P-Value	AUC	Threshold (%)	Sensitivity	Specificity
Bach Model	Bach (2003) [12]	External	300	NR	0.72	NR	NR	NR
	D'Amelio Jr (2010) [14]	External	4,900	NR	0.66	2.5 5 7.5	0.302 0.155 0.064	0.888 0.976 0.988
	Wilson (2015)	External	26732 26722	0.06 0.81	0.695 0.687	NR NR	NR NR	NR NR
Liverpool Lung Project	Cassidy (2008) [16]	Internal	1,736	NR	0.71	2.5 6	0.62 0.34	0.7 0.9
	Raji (2012) [18]	External (x3)	1,868 -7,652	NR	0.67 - 0.82	2.5 5 10	0.552 - 0.743 0.357 - 0.574 0.142 - 0.279	0.674 - 0.731 0.811 - 0.871 0.925 - 0.957
	Raji (2010) [17]	External	1,275	NR	NR	0.91 2.5 5.12	0.861 0.678 0.499	0.395 0.642 0.798
	D'Amelio Jr (2010) [14]	External	4,900	NR	0.69	2.5 5 7.5	0.667 0.455 0.312	0.666 0.849 0.923
Spitz Model	Spitz (2007) [19]	External (x3)	963	0.688 - 0.777	0.57-0.63	NR	NR	NR
	Spitz (2008) [26]	External (x2)	1,340	NR	0.67-0.68	NR	NR	NR
	D'Amelio Jr (2010) [14]	External	4,900	NR	0.69	2.5 5 7.5	0.266 0.068 0.022	0.944 0.993 1
African-American Model	Etzel (2008)	Internal External	156 325	NR NR	0.75 0.63	NR NR	NR NR	NR NR
PLCO Model	Tammemagi (2011)	Internal	70,962 38,258	0.274 0.416	0.859 0.809	NR NR	NR NR	NR NR
PLCOM2012 Model	Tammemagi (2013)	Internal External	36,286 37,332	NR NR	0.803 0.797	NR 1.35%	NR 0.83	NR 0.629
	Wilson (2015)	External	26732 26722	0.04 0.01	0.702 0.69	NR NR	NR NR	NR NR
PLCOM2014ALL Model	Tammemagi (2014)	Internal	53,454	NR	0.848	NR	NR	NR
Hoggart Model	Hoggart (2012) [23]	External 1Y External 5Y	16,906 14,032	NR NR	0.843 0.787	NR NR	NR NR	NR NR
Pittsburgh Model	Wilson (2015)	Internal Internal	26732 26722	0.19 0.08	0.688 0.687	NR NR	NR NR	NR NR

Table 3.3: Epidemiological Models All Validation Results

3.5.4 Hoggart Model

The Hoggart Model (Hoggart et al 2012) is a Weibull regression model, devised in a prospective cohort of 169,035 ever-smokers [79]. From this, 90% of the population was used to devise the model and the remaining 10% formed an external validation set. The model was applicable to ever-smokers aged 35+ years [79]. The model only considers age and smoking history to calculate one-year absolute risk (incidence and competing risk of death). The reduce volume of variables considered in comparison to other models could allow the Hoggart Model to be successful in new environments by not over-fitting the model. Additionally, the model has the potential to be improved at a later stage by extending the model to consider additional variables. Conversely, the small number of variables could hinder the model performance and the model could be limited when attempting to assign a higher risk to individuals with lung cancer when only considering smoking history and age.

The formula and incidence rates presented in the article were incorrect. A correct version was obtained through author contact and are presented in the Appendix.

The model can be run recursively and in the external validation in the original article, was evaluated over 1 and 5 years. The model reported some very promising results and the AUCs were 0.843 and 0.787 for 1 and 5-years respectively. These are very promising results and suggest the model has a good discriminative ability. The results suggest the model performs stronger over a smaller time frame and there were no significant changes reported when stratified for former- and current-smokers. While the initial results are encouraging it is important that the model is tested in distinct environments to see if a similar or better performance can be obtained.

It is unfortunate that the model has only been validated in the original article and there has been no evaluation of the prediction rules. Therefore, despite the primary success there are no guidelines on how the model could be applied as a selective screening tool. Future testing should evaluate the prediction rules to identify at which risk threshold the model performed optimally. Further testing will also provide an opportunity to assess whether the Hoggart Model can replicate the initial results while only considering a few variables.

3.5.5 PLCO Model

The Prostate, Lung, Clonal and Ovarian Cancer (PLCO) screening trial created two logistic regression models to predict risk of developing lung cancer within 9 years, one was applicable to everyone and one for ever-smokers. There were no further model restrictions. The models (Tammemagi et al 2011) use a cubic spline for the predictors which include age, pack years, smoking duration, BMI, family history, COPD and chest x-rays [80]. The two models were created in a cohort; the model applicable to everyone used 70,962 participants and the ever-smokers model used a sub cohort of 38,258 ever-smokers [80].

The model was internally validated in the original article and bootstrapping was conducted in an attempt to eliminate optimism and give a fairer reflection of the model potential in distinct environments. The calibration and discrimination were assessed. The model version that was applicable to everyone reported a very high AUC of 0.859 and good calibration (p-value 0.274). The strong performance was replicated in the ever-smokers model with an AUC of 0.809 and a Hosmer-Lemeshow p-value of 0.416.

Currently the models have not been externally validated. Future testing needs to be conducted on the PLCO models which recorded the highest AUC results of any model currently, although this was reported in an internal validation. The assessment should evaluate if the good model calibration and AUC results can be replicated in new populations and the prediction rules need to be evaluated. The initial results would indicate the model could be a useful tool for the public to calculate their risk or utilised to identify a high risk target group for screening. However, failure to report the prediction rules means currently there is no indication how the model could be optimally utilised as a selective screening tool.

Since publishing these models the author has developed two new versions of the model discussed in Section 3.5.6 & 3.5.7.

3.5.6 $PLCO_{M2012}$ Model

Following the promising results of the PLCO models, these were revised to create the $PLCO_{M2012}$ logistic regression model to calculate risk over 6 years. This was applicable to ever-smokers with no further restrictions [81]. However, the revised model did not consider pack years or x-rays but did include ethnicity.

The $PLCO_{M2012}$ Model was devised in a cohort of 77,456 participants; these were heavy smokers as participants used to devise the model were all eligible using the NLST criteria to offer a direct comparison. Therefore, the model was devised in ever-smokers aged 55-74, with a minimum 30 pack year smoking history and all former-smokers had quit within the last 15 years [81].

An external validation was conducted in the original article using a cohort of 53,454. This was also a heavy smoking cohort with participants required to have a minimum 30 pack year smoking history. The model was compared to the original PLCO Model and the study found "the AUCs in the validation data suggest that predictive discrimination with the $PLCO_{M2012}$ was slightly improved" with an impressive result of 0.803 [81]. The strong discriminative ability was further supported by the prediction rules results which were optimal at a risk threshold of 1.3455%. Here the model recorded a Youden's Index of 0.459, which is very high, as the model demonstrated a good trade-off between the sensitivity and specificity. Indeed, the model reported a sensitivity of 83% while maintaining a high specificity of 63%; this very impressive result indicates the model's potential as a selective screening tool. However, it is important that the model is validated in multiple distinct populations at the reported threshold to assess if a consistent level of performance is recorded.

Tammemagi et al. (2014) externally validated the model [82]. The study found 0.0151 (1.51%) was the optimal 6-year risk threshold, which is not too dissimilar to the 1.3455% reported in the original article. The model was compared to the NLST screening population in a high risk 53,455 cohort that satisfied the NLST screening criteria [83]. At the 1.51% risk threshold the model performed strongly and demonstrated an improvement upon the NLST criteria. The $PLCO_{M2012}$ Model reported a sensitivity of 80.9%, specificity of 65.9% and a PPV of 4.1%. This PPV result is an improvement on the NLST criteria which scored 3.4%. The model and NLST criteria were validated in a cohort, with a reasonable lung cancer incidence rate, so the PPV results are a fair reflection of what would be expected in a genuine screening trial. However, it is important to note this is only in ever-smokers with a 30+ pack year history who were aged 55-74 rather than all ever-smokers as the model is applicable to. This result and the sensitivity and specificity rates may be weaker in a complete ever-smoker population. Future testing could evaluate the model in a wider population to assess if the results can be replicated and if the model still improves upon the NLST screening programme. The Hosmer-Lemeshow test and AUC were not recorded in the study but the reported results demonstrated how prediction models may offer an improved criterion for selecting high risk individuals. An 11-year risk version of the model was also examined. However, the article concluded this is an excessive time frame and there was no improvement to the model [83].

Overall this is an extremely promising model offering an improvement over the original PLCO model and NLST criteria. The articles also indicated the model was optimal around the 1.5% risk threshold for 6-year risk of incidence. The model needs further external validations, particularly in populations that are not heavy smokers, to assess the model's robustness in distinct populations. If the model continually reports the high results such as sensitivities of 80% and specificities of 63% then the model may be considered in screening trials. In contrast the reported calibration results have been poor when reported and the model may not be preferred to be made available to the public to estimate their risks.

3.5.7 $PLCO_{M2014}$ Model

Another PLCO logistic regression model was created (Tammemagi 2014) to include never-smokers when predicting 6-year incidence [82]. The $PLCO_{M2014}$ Model included a new variable for smoking status in the logit model while the other variables included in the $PLCO_{M2012}$ Model remained with a minor recalibration to adjust for the never-smokers.

Model	Lead Author	Test/Validation Set	Size	Study Type	Study Eligibility Criteria
Bach	Bach	Test and internal validation	18,172	Cohort	Aged 45-69, Ever smoker, 20+ PY, quit within 15 years
	Crunin	External validation	6,239	Cohort	Aged 50-69, Ever smoker, Male, 5+ CPD
	Maisonneuve	External validation	5,203	Cohort	Aged 50+, Ever smoker, 20+ PY, quit within 10 years
	Wilson (2015)	Test and internal validation	53,454	Cohort	Aged 55-74, Ever smoker, 30+ PY, quit within 15 years
LLP	Cassidy	Test and internal validation	1,736	Case-Control	Aged 20-80
	Raji	External validation	388	Case-Control	Aged 20-80
			1,823	Case-Control	None
	Raji	External validation	2,922	Case-Control	None
			7,652	Cohort	Aged 40-79
Spitz	Spitz	75% test and 25% external validation	3,852	Case-Control	None
Bach, LLP, Spitz	D'Amelio	External validation	4,900	Case-Control	Caucasion
African-American	Etzel	Test and internal validation	988	Case-Control	African-American
PLCO	Tammemagi	Test	70,962	Cohort	Aged 55-74
		External validation	44,223	Cohort	Aged 55-74
$PLCO_{M2012}$	Tammemagi	Test and internal validation	77,456	Cohort	Ever smoker, 30+ PY
	Wilson (2015)	Test and internal validation	53,454	Cohort	Aged 55-74, Ever smoker, 30+ PY, quit within 15 years
$PLCO_{M2014}$	Tammemagi (2014)	Test and internal validation	154,910	Cohort	Aged 55-74
Hoggart	Hoggart	90% test and 10% external validation	169,035	Cohort	Aged 40-65, Ever smoker
Pittsburgh	Wilson (2015)	Test and internal validation	53,454	Cohort	Aged 55-74, Ever smoker, 30+ PY, quit within 15 years
			3,654	Cohort	Aged 50-79, Ever smoker, Quit within 10 year, 10+ CPD for 25+ years

Table 3.4: Study Design of Building and Validating Datasets for Epidemiological Models

The model was devised and internally validated in the PLCO cohort. The study included 154,910 participants who were aged 55-74 with no further restrictions [82].

Despite creating a model applicable for never-smokers, the article argued "it has not been demonstrated that never-smokers can be at high enough risk to warrant screening" [82]. As a result, the study did not comprehensively review the performance of the $PLCO_{M2014}$ Model despite the model reported an impressive AUC of 0.848 in an internal validation. This is a very promising result although the model needs to be externally validated to give a more accurate review of the model's discriminative ability. The article did not formally record the calibration using the Hosmer-Lemeshow test but reported "the $PLCO_{M2014}$ Model calibration is good for risks below 0.10" but overestimates risk in individuals with a risk above 0.15 for 6-year risk of incidence [82]. However, only 0.2% of the participants had a risk above 0.15 in the validation.

The study assessed the model at the optimal risk threshold identified for the $PLCO_{M2012}$ Model (1.51%). The prediction rules were not reported since none of the 65,711 never-smokers in the cohort were estimated a risk exceeding 1.47% and as a consequence would not be considered for screening [82]. This is despite the fact never-smokers can be assigned a risk up to 3.5% using the prediction model. The study supported their previous argument that never-smokers do not warrant a high enough risk for screening and they recommended the $PLCO_{M2012}$ Model in preference.

However, the model would benefit from an independent validation that identified an optimal risk threshold specific for this model to evaluate the prediction rules. Additionally, the model recalibration may have improved the model's ability as a selective screening tool; even if never-smokers are not considered for screening the model may improve how ever-smokers are classified as high or low risk. Clearly, the model demonstrates a good discriminative ability in the internal validation. Future testing should evaluate the model in an external population and identify a risk threshold specific for the $PLCO_{M2014}$ Model to evaluate the prediction rules. The calibration should also be evaluated using the Hosmer-Lemeshow test. Currently there has been no independent validations of the model as a direct consequence of the model being published very recently in December 2014 with the systematic review deadline 01/10/2015.

3.5.8 African-American Model

The African-American Model [84] is closely related to the Spitz Model and devised by the same lead authors. The logistic regression model was devised to predict accurate risks in an African-American population whereas previous models commonly considered a Caucasian population. The model was devised to predict 5-year absolute risk by considering lung cancer incidence rates and survival rates. However, the model was devised using a very small 988 case-control population [84]. This could be a limitation of the model as there is relatively little evidence to develop the prediction model. Although some concerns are alleviated by considering the external SEER lung cancer incidence and survival rates for age and gender specifically for African-Americans [84]. This may assist in developing a model that estimates reasonable risks of developing lung cancer in African-Americans. By considering these age and gender specific incidence rates the model cannot be applied to participants under 20, although lung cancer rarely occurs under 40 [1] so this is not a concern. The model includes many of the same variables as the Spitz Model, although does not consider asbestos exposure and emphysema in the prediction model [84].

The model was internally and externally validated in the original article [84]; unfortunately, these were conducted in very small populations. The internal validation only included 156 participants where the model reported a good AUC of 0.75 [84]. Unfortunately, this was not replicated in an external validation of 388 participants with a poorer AUC of 0.63 [84]. It could be argued that the small external validation dataset may have hindered the model performance as it is insufficient to determine if the model has a poor overall discriminative ability or only performs poorly in the dataset. However, the result does raise the concern that only using 988 participants to build the model could result in the model performing poorly in new environments.

There has been no additional testing other than the original article and the calibration and prediction rules have not been evaluated. While the original results are poor this model may report the leading performance in African-American populations, as this is currently the only model developed for this target group.

Certainly, there is the potential for specific models for different populations whether based on ethnicity or location provided they offer a leading performance when assessing the model calibration, discrimination and prediction rules. Future testing should validate the African-American Model in an African-American population in a direct comparison to universal models to assess which report the leading performance. Additionally, testing the model in a much larger validation should also allow a better understanding into the model's performance and answer concerns on whether the model is limited in new populations.

	Overall		**Prediction Rules**						**Same Dataset Testing**
Model	Calibration	AUC	Risk Threshold	Sens and Spec	PPV	NPV	Model Accuracy	NRI	Compared With?
Bach	X	X	X	X	X	X	X		LLP, Spitz, Pittsburgh
LLP		X	X	X	X	X	X		Bach, Spitz
Spitz	X	X	X	X	X	X	X		Bach, LLP, African-American
African-American		X							Spitz
PLCO	X	X							
PLCOM2012	X	X	X	X	X				PLCOM2014, US Screening Trial
PLCOM2014		X	X						PLCOM2012, US Screening Trail
Hoggart		X							
Pittsburgh	X	X							Recalibrated Bach, Recalibrated $PLCO_{M2012}$

Table 3.5: Testing and Model Comparison Reported for each Epidemiological Model

3.5.9 The Pittsburgh Predictor

The Pittsburgh Predictor (Wilson 2015) [85] is a recent logistic regression model that was identified for the systematic review through author contact. It is a simple model that considers only four variables; age, smoking duration, CPD and smoking status in order to predict lung cancer risk of incidence within 6-years [85]. The simplicity of the model may be an advantage when applied to new environments as the model only considers age and smoking history, both linked to lung cancer risk worldwide. Indeed, the model was designed to consider fewer predictors to make it "less complicated than currently available models, perhaps more amenable to widespread application, yet devised to place individuals on a lung-cancer continuum." [85]. Conversely it could be argued, the limited number of predictors in the model may disadvantage the model's discriminative ability. The model is only applicable to ever-smokers, although there are no additional restrictions [85].

Two versions of the model were created in slightly different cohorts. The first cohort recruited 53,454 high risk ever-smokers cohort aged 55-74 with a minimum 30 pack year smoking history. A second cohort of ever-smokers recruited 3654 50–79 year-old ever-smokers who had smoked a minimum 10 CPD for at least 25 years and former-smokers had quit within the previous 10 years [85]. Despite being applicable to all ever-smokers the two cohorts collected only heavy smokers which may limit the model's performance in lower smoking populations.

The two model versions considered the same parameters but were slightly recalibrated based on the evidence in the cohort. At this stage the models have only been internally validated in the original study where they were compared to a recalibrated version of the Bach Model and the $PLCO_{M2012}$ Model. The models were recalibrated using an odds ratio scaling factor to reflect the observed incidence rates and are presented in the article [85].

The calibration (Hosmer-Lemeshow) and discrimination (AUC) results were presented. The two versions of the Pittsburgh Predictor demonstrated a good calibration, exceeding 0.05 (p-values 0.19 and 0.08). In comparison the recalibrated Bach Model also reported a good calibration but the $PLCO_{M2012}$ (p-values 0.04 and 0.01) underperformed in both cohorts. In the internal validation the Pittsburgh Predictor showed a reasonable discriminative ability with AUC results of 0.688 and 0.678. However, this was surpassed by the other models. This adds weight to the argument that the small number of variables considered in the Pittsburgh Predictor may limit the model's ability to assign a higher risk to individuals with lung cancer.

Based on the AUC evidence in the internal validations the leading Pittsburgh Predictor published in the article is as follows;

$$\frac{1}{1+\exp-(-4.2195+(0.1\times S))} \tag{3.1}$$

Here S is the additive model using the 4 variables as presented in the article [85].

The AUC results indicate that the Pittsburgh Predictor may not be the favoured model for a selective screening tool. At this stage the model should be validated in external populations which currently have not been conducted due to the recent publication of the model. Future testing should also assess the prediction rules which were not presented in the original publication. Any future validations should also compare the performance to other models in the same datasets. This can then evaluate whether only considering four variables has allowed the model to be successful in distinct environments in comparison to more sophisticated models that might consider variables not relevant in new populations. Conversely, the validation may highlight, by restricting the volume of variables this limited the model's discriminative ability.

3.6 Clinical Lung Cancer Models

The next group of models presented in the systematic are defined as clinical lung cancer models. Models are categorised as such if they include variables that require a clinical assessment (CT/PET Scan, blood test, sputum test) to estimate an individual's risk. These models have the potential to incorporate crucial markers that increase or decrease ones' susceptibility to developing lung cancer. However, at this stage there is ongoing work to identify the markers that best explain risk [86] indicating the possibility of more clinical models being developed in parallel to further research into SNPs. Additionally, at the time the review was conducted the key markers may not have been identified so a leading model may not be developed. There is an expectancy that these markers will improve prediction model performance, however, these models all require a costly procedure, such as a scan or blood test. Therefore, they will in all likelihood need to show a significant improvement over leading epidemiological models to justify the costs involved to generate risk predictions. Unfortunately, there is no exact measure to assess level of improvement for a clinical model while considering increased costs incurred, this would most likely be reviewed in a cost-effectiveness analysis during a screening trial which have not been conducted for the new clinical models.

The review identified and presented 7 non-epidemiological prediction models. These models are a combination of new models and updated epidemiological models to include clinical variables. The results are presented and discussed separately for each model with the results combined in one table.

3.6.1 Pulmonary Function and Sputum DNA Image Cytometry Model

In 2011 an article published two models that predicted lung cancer incidence risk over eight years in ever-smokers [86]. The models were designed in a prospective cohort of 2,596 high risk participants aged at least 40 years with a minimum 20 pack year smoking history [86]. The models were an extension of the original PLCO models and presented by the same lead author. As a result, the epidemiological variables considered in the model originate from the PLCO models [86]. The original model was then extended to consider forced expiratory volume (FEV) and sputum DNA imaging.

To apply the models, participants would be required to be tested by a Spirometer and provide a sputum sample which required "subjects were instructed to cough intermittently during the induction procedure and for at least 2 hours afterward to produce sputum samples" [86].

The combined model including forced expiratory volume and sputum DNA imaging recorded the best results in the internal validation with a high AUC (0.773) and good calibration (p-value 0.313) [86]. However, the AUC did not exceed the 0.859 or 0.809 observed from the original model when tested amongst everyone and ever-smokers respectively in the internal validations. The initial results indicate that considering FEV and DNA have not improved the model's discriminative ability, it would in all probability lead to a poorer selective screening tool although this has not been thoroughly evaluated.

The model has only been tested in the internal validation. Therefore, the next stage should evaluate the model in an external population ideally in comparison to the original PLCO models. Currently,

the AUC result of the extended model were slightly inferior to the epidemiological PLCO model, which would not require 2 hours from the participant to estimate their risk. This would suggest the clinical model would not be a better screening tool in terms of convenience, costs incurred and discriminative ability. However, FEV and DNA may be universal factors that allow the model to perform robustly in new environments and outperform the original PLCO model. The current results would argue the new model will have to demonstrate a significant improvement in AUC (and most likely prediction rules which have not be assessed) for clinicians to favour this model as a selective screening tool over leading epidemiological models, however, at the current stage of research this should be assessed.

3.6.2 Extended LLP Model

The LPP Model was recalibrated and extended to include the SEZ6L gene as an additional variable [74]. The model predicted 5-year absolute risk of developing lung cancer similar to the original LLP Model. The model was devised using a subset of the case-control LLP study which collected information for this genotype. As a result, the model was only based on 388 participants [74]. It is possible a larger population may have indicated that the SEZ6L gene was not a key marker to explain lung cancer susceptibility. Additionally, this may also lead to over-fitting as a result of the small case-control study upon which the model extension was based; this variable may explain the difference between predicted and observed risks in individuals in the original model rather than predicting an individual's absolute risk of developing lung cancer. However, there is only a small concern about over-fitting as the model is only extended to include one additional variable [87].

The study participants were aged $20-80$ but there were no further restrictions. Controls were matched on age and gender and were also from the Liverpool, UK area [74]. However, the model used age and gender specific lung cancer incidence and death rates, determined from the overall Liverpool area, rather than the case-control study to reduce concerns with matching for these variables.

The SEZ6L gene was identified in participants by "extract(ing) the genomic DNA from blood peripheral leukocytes" [74] using a blood kit which can be a costly procedure on a large scale.

The extended model was assessed in an internal validation and reported a good AUC (0.75), although the original LLP Model had only a slightly reduced discrimination (0.72) in the same dataset. Unfortunately, the prediction rules for both models were not presented as this would indicate by what extent the extended model was an improved selective screening tool, which is expected from the improved AUC results. Both the original and expanded model reported a good calibration.

There has been no subsequent testing and while there was an improvement in AUC, the inclusion of the SEZ6L SNP does not appear to be more beneficial in comparison to the original LLP model. This marginal improvement may not justify this model being preferred as a selective screening tool when considering the extra cost incurred from conducting blood tests. Future testing of the model should compare the extended and original models in the same external population with particular attention to comparing the prediction rules.

3.6.3 Extended Spitz Model

The original Spitz Model was extended to consider two markers for DNA repair capacity [76]. The model calculates 1-year absolute risk of developing lung cancer in ever-smokers [76], whereas the original model was applicable to everyone, due to a lack of reporting for the new variables in never-smokers. To be applicable participants were required to be at least 20 years of age, as this is the youngest age provided for the age and gender lung cancer incidence rates using the SEER rates. However, this is not a limitation as lung cancer rarely occurs in individuals younger than 40 [1]. Additionally, incorporating these SEER rates into the model limits concerns about matching as the case-control study recruited 1,340 participants with controls matched by age, gender, ethnicity and smoking status.

The extended model considers the same variables as the original model but included DNA repair capacity and Bieomycin sensitivity as additional predictors [76]. These required blood samples to be

measured. There is a concern in including these variables as the case-control participants were matched by smoking status which may be confounded with DNA repair capacity and Bieomycin sensitivity. Therefore, these new variables may not assist the model performance any more than considering smoking status.

The model was internally validated in the original article. The model demonstrated an improved AUC in comparison to the original model with results of 0.70 and 0.73 for 746 former and 594 current-smokers respectively. In comparison, the Spitz Model reported an AUC of 0.67 and 0.68 for former and current-smokers in the same dataset. The clinical model demonstrated a good calibration with Hosmer-Lemeshow p-values of 0.61 and 0.433. Unfortunately, the prediction rules were not evaluated in the article.

Further testing should be conducted in an external validation and compare between the original and extended models. The prediction rules need to be evaluated as there is no current indication as to how the model could be best utilised as a screening tool. Additionally, an external validation will demonstrate if a more robust model has been created; considering the additional markers to improve the performance of the original model may only explain the difference between the predicted risks of the original model and the observed incidences in the dataset rather than predicting lung cancer risk. This would limit the model's ability in new environments, which can be evaluated by testing the calibration and discrimination.

3.6.4 Model for Korean Men

A new model was developed using a large cohort of 1,324,804 Korean men [88]. This is one of only two models that were devised using a population outside of Western Europe and North America. The study collected information on males aged between 30-80 with no further restrictions [88]. It is unfortunate that the model was restricted to males, as lung cancer has the second highest incidence rate and highest mortality rate of all cancers amongst females [1].

The model considered age, smoking history, BMI, physical activity and fasting glucose levels to predict 8-year risk of incidence. Fasting glucose levels requires blood and urine testing which classifies the model as a non-epidemiological model. There were some concerns with the ambiguity of some of the variables considered in the model and individuals may be unsure how to classify themselves. For example, physical activity was defined as light, moderate, or heavy [88], which is subjective to the individual's interpretation as what qualifies as exercise, which could affect model accuracy.

The model was externally validated in the original article in a substantial 507,046 cohort of males aged $30 - 80$ years [88]. The Korean Model recorded an extremely impressive AUC of 0.875; this was one of the leading results of all models reviewed. In contrast, the calibration was poor (p-value < 0.0001) as the model failed to accurately predict risks in individuals. The study did not report the prediction rules, which is disappointing as the high AUC results could be reflected in impressive prediction rules results.

It would be of interest to assess how the model performed in other Asian and worldwide populations. These should evaluate if the model can replicate its impressive performance in distinct environments. It would be insightful to compare the model in the same populations (in Asia, Europe and North America) to other leading models such as the PLCO and Hoggart models. This can evaluate if there is a leading universal model or leading models for different populations. While the model showed potential based on the AUC results the prediction rules require evaluation. This will allow medical decision makers to make an informed decision as to whether the prediction model should be used as a screening tool and provide guidance on how to apply the model to identify a high risk group who would benefit from screening. If the model still maintains a leading performance, then the costs of taking blood samples for the fasting glucose levels may be justified, after an assessment in a cost-effectiveness analysis, in preference to epidemiological models.

3.6.5 COSMOS (Extended Bach) Model

The COSMOS Model is a recalibrated and extended version of the Bach Model [89]. The model was devised using a cohort of 5,203 participants who were aged 50 years or over and heavy smokers with a minimum

Model	Article	Validation	Study Size	H-L P-Value	AUC	Risk (%)	Sens	Spec
P & S DNA Model	Tammemagi (2011)	Internal	2,596	0.313	0.773	NR	NR	NR
Extended LLP	Raji (2010)	Internal	388	NR	0.75	NR	NR	NR
Extended Spitz Model	Spitz (2008)	Internal	746 594	NR NR	0.7 0.73	NR NR	NR NR	NR NR
Korean Men Model	Nam (2013)	External	507,046	< 0.0001	0.875	NR	NR	NR
COSMOS Model	Maisonneuve (2011)	Internal	5,203	0.63	NR	NR	NR	NR
Gene Based Risk Score	Young (2009)	Internal	439	NR	0.79	NR	NR	NR
Chinese Genetic Model	Li (2012)	External	1267	NR	0.639	NR	NR	NR

Table 3.6: Systematic Review Results of Epidemiological and Clinical Models

20 pack years smoking history [89]. This is similar to the Bach Model criteria, although it also included ever-smokers with a lower smoking history compared to the Bach Model, which used ever-smokers with a minimum 30 pack year smoking history. The COSMOS Model was developed as the Bach Model recorded a poor calibration in this population. The COSMOS Model included nodule description and size from CT scans as additional variables. It could be argued that if CT scans are required for each participant then a blanket screening approach of everyone applicable to the model can be conducted as LDCT trials have been successful in previous lung cancer screening programmes [27]. The model has a limited potential of identifying general participants for screening as it considers nodules; these can develop into lung cancer, therefore, it could be argued participants with nodules are already high risk participants.

The COSMOS Model was internally validated and the Hosmer-Lemeshow was measured (p-value 0.63) [89]. The model successfully predicted lung cancer risk, although this is in asymptomatic participants.

An external validation was conducted by Veronesi et al. [90]. The model was tested in the same target population as the model building population; aged at least 50 years and a minimum 20 pack year smoking history. The study collected 1,035 patients who were receiving annual LDCT screening [90] and evaluated the model's estimated lung cancer incidence rate that would accrue during annual LDCT screening over the next 10 years. The COSMOS Model accurately predicted lung cancer incidence rates for the first two years of screening with expected incidence rates of 12.5 and 13.4 in comparison to the observed 12 and 11 incidences [90]. However, the COSMOS Model over predicted incidence rates for the next 8 rounds of annual screening. Over the 8 screening rounds the COSMOS Model predicted an accumulated 156.2 incidences in comparison to the much lower 60 incidences observed [90]. This heavily over-predicted, predicting a 260.3% increase on the observed rate. In comparison, the Bach model predicted 58.3 incidences over the same 8-year period equating to 96.7% of the observed incidences [90]. The results suggest that the COSMOS Model had been recalibrated too highly based on initial high capture rate at the baseline round of screening. The expected versus observed calibration was the only measure reported. The poor result suggests that the COSMOS Model failed to adapt to new environments.

Based on the calibration results provided, the extended COSMOS Model does not seem to offer an improvement over the Bach Model despite requiring costly procedures to apply the prediction model. Additionally, the model may not be appropriate as the objective of the prediction model is to identify participants for screening. Requiring screening to make the predictions would negate the model's practicality. The COSMOS Model offers an indication into the observed incidence rate in this high risk population, although the external validation results suggest the original Bach Model would be a better tool for this purpose.

3.6.6 Gene Based Risk Score

A model was created to incorporate SNPs that could influence lung cancer susceptibility [91]. The model was devised in a population based case-control dataset. Ever-smokers with a minimum 15 pack year smoking history who were at least 40 years of age were considered [91]. The model considers for inclusion 20 SNPs associated with lung cancer; research determined 12 of the SNPs increased susceptibility of developing lung cancer and the remaining 8 SNPs reduced risk of lung cancer. Information for these SNPs were identified through blood tests. These SNPs are considered alongside family history, gender, age and COPD in an additive model.

In the original article a small internal validation was conducted using 439 participants. The model reported a strong AUC of 0.79 [91], which demonstrates the potential to incorporate SNPs into prediction models. However, additional testing in larger external populations would allow a more extensive understanding into the potential of SNPs to create robust prediction models.

The research into key SNP markers is still ongoing and newly identified SNPs could be considered in future lung cancer prediction models. Similar to all reported clinical model validations there has been no focus on prediction rules which need to be evaluated before the model would be considered to identify participants for selective screening trials. Additionally, research could assess whether the identified SNPs can be incorporated into a robust prediction model that also considers never-smokers.

3.6.7 Chinese Genetic Model

The Chinese Genetic Model (Li et al 2012) [92] was created using 75% of a 5,068 participant case-control dataset with the remaining 25% utilised as a validation set. There were no restrictions in collecting the 2,283 individuals with lung cancer and disease free participants were matched by age, gender and residency [92]. This was only the second prediction model that did not consider a European or North American subpopulation.

The model considers 4 SNPs (rs2736100, rs402710, rs4488809 and rs4083914) alongside age, gender and smoking status in a log additive model [92]. Several versions of the models that considered different combinations of these risk factors were published in the article. The first version of the model only considered SNPs and reported a poor performance despite "estimating these four SNPs accounted for 4.02% of genetic variance in lung cancer" [92]. The external validation in the original article reported a weak AUC of 0.551 (95% CI [0.532, 0.564]) [92]. This suggests that solely considering non-epidemiological factors cannot create a good prediction model and a combination of epidemiological and non-epidemiological factors will generate better predictions. Indeed, a second version of the model considering gender and smoking status alongside the SNPs improved the AUC to 0.639 (95% CI [0.621, 0.652]) [92], which is a more reasonable discriminative ability.

In a final model version that considered the SNPs, age, gender and smoking status, the article evaluated the prediction rules at the optimal risk threshold. This was identified by the highest Youden Index across all risk thresholds. However, the article did not report the risk threshold that had the optimal performance which means that this cannot be evaluated in future validations to assess if a similar performance can be obtained. The model reported an optimal performance with a sensitivity of 53%, specificity of 68% and a PPV of 57% [92]. This was not an outstanding performance with a Youden Index of 0.21 in comparison to the performance of other models, which have reported a value exceeding 0.4. The model would still screen a high proportion of individuals who are disease free while only identifying 57% of individuals who developed lung cancer. The calibration was not evaluated in the article.

Overall the model had a poor performance with lower AUC and prediction rule results. The model's poor performance may suggest the most significant genetic markers for lung cancer susceptibility have still not been identified. Unfortunately, genetic prediction models are continually being published without identifying the key SNPs. This increases the volume of prediction models published, including some poor performing models, leading to confusion as to which model(s) would be beneficial as a clinical utility.

3.7 Two-Stage Clonal Expansion Models

The final collection of models identified in the systematic review are the Two-Stage Clonal Expansion (TSCE) models. These models differ from previously considered models as they consider one risk factor and are commonly developed and applied to a subpopulation that has a high exposure to the risk factor. The TSCE models offer a comparison of "the strength and goodness of fit of the association between exposure and cancer risk" [93] and are normally devised to evaluate if a risk factor is associated with the outcome. By only considering one factor they are most likely to be limited in accurately predicting an individual's risk of developing lung cancer as including additional variables can improve the models' discriminative ability rather than an increased exposure equates to an increased risk. Despite their potential limitations TSCE models can be applied to calculate an individual's risk.

In the majority of studies, the model is validated by comparing observed and estimated lung cancer incidences/mortalities to assess the model calibration. Since the models assess the level of association of a variable with lung cancer risk they are often not applied in an external population as they are not considered as a universal selective screening tool.

The review will offer an overview of the TSCE process followed by the different models and their results.

3.7.1 TSCE Model Process

TSCE models are aptly named as they are split into two distinct stages. These stages evaluate the intensity that normal cells become pre-malignant and the rate of expansion, which considers cell birth and death rates plus the rate of transformation from pre-malignant to malignant.

3.7.1.1 Stage 1

The intensity of the transformation from normal cells to pre-malignant cells relates to the quantity of cells that are susceptible at a time, t. This will be a fixed number of cells in the lung defined as $X_{(t)}$. Cells are 'initiated' into premalignant cells which is "a (nonhomogeneous) Poisson process" [94]. This rate is defined as;

$$\alpha_{1(t)} = v_t \times X_t \tag{3.2}$$

Where $\alpha_{1(t)}$ is the initiation rate for a fixed number of cells based on the intensity of initiation (v_t) at time t.

3.7.1.2 Stage 2

The second stage considers the rate of change of cells from pre-malignant to malignant. This mutation rate is the asymmetrical division of a pre-malignant cell into one pre-malignant cell and one malignant cell [94]. The rate at time t is denoted as $\alpha_{2(t)}$.

3.7.1.3 Intermediate Section / Clonal Expansion Rate

The two final rates in the model are specific to the pre-malignant cells. These are the birth rate (cell division) b_t and death rate (differentiate) d_t at time t [94].

The methodology of how to form the TCSE model, including notation at stages comes courtesy of detailed description in Heidenreich et. Al [94] and Zeka et. Al [93]. The key terms of the TSCE models are summarised as:

X_t	Number of cells - fixed
v_t	Intensity of cell initiation, from normal to premalignant
$\alpha_{1(t)}$	Initiation rate
$\alpha_{1(t)}$	Mutation rate, from premalignant to malignant
b_t	Premalignant cell birth rate
d_t	Premalignant cell death rate

Figure 3.2: List of TSCE notation specific at time (t)

The overall process is best described by Zeka et. al. [93] and can be summarised as follows;

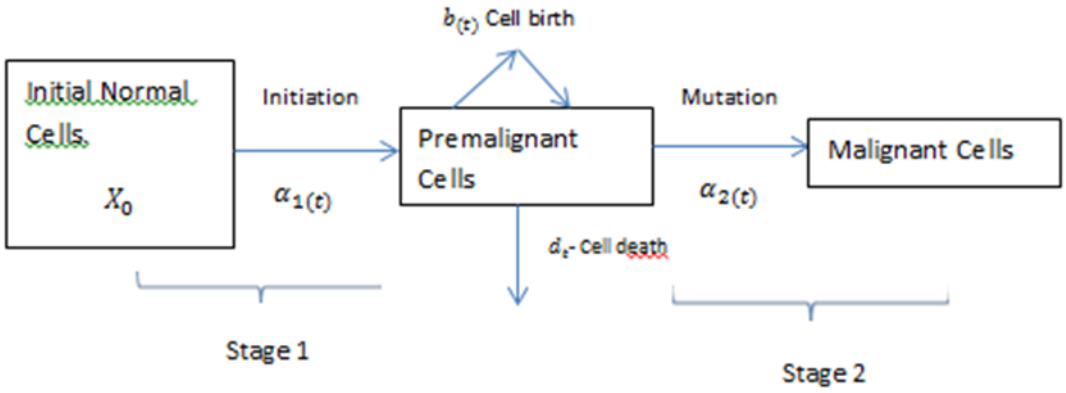

Figure 3.3: Flow of TSCE Model Process

3.7.2 Review of TSCE Prediction Models

Eight models were identified and were each presented in a paper ([93], [95]-[101]). Across these models asbestos or silica exposure was commonly considered as the sole variable. In all the TSCE models the volume of lung cells remained fixed at 10^7, but the mutation rates varied based upon the variable considered in the model and the study population. The working formulas with mutation rates are listed in the published articles ([93], [95]-[101]).

The aim of all the identified models was to predict lung cancer mortality rates across a population and were then evaluated by comparing the estimated and observed mortality rates and the results showed potential [96]. None of the models have ever been validated or used in new environments outside the original study. The lack of interest in conducting external validations for these models is partly due to the model design; by only considering one, often uncommon variable, they would be limited as a selective screening utility. Further, while they can be easily applied to high risk or high exposed populations their simplicity can be a limitation; a single factor can struggle to distinguish between individuals who would benefit or not benefit from screening. Finally, the models are applied in a highly exposed population and are not very transferable to new environments.

The TSCE models are unlikely to ever be considered outside of the original study nor would these models be employed to assist physicians. There is no further work or validations required for the TSCE models as their primary purpose is to show the level of association of a risk factor with lung cancer in the original article.

3.8 Discussion

The published literature on lung cancer prediction models indicates that the focus has been in publishing new models rather than validating existing prediction model. As a result, there was a varied standard of validations for models with some being extensively validated while others have only been considered in the original article. The epidemiological models have been the most extensively reviewed, which is expected since they were published earlier. Since then there has been a shift to consider incorporating genetic markers in prediction models. These published clinical models are relatively new and therefore have rarely been evaluated outside the original article. Finally, the TSCE models were seldom used and have never been considered beyond the original study. These targeted high risk, niche populations so were not practical in new environments.

When validating the epidemiological models the results are often presented for different models using distinct datasets so direct comparisons between models has been difficult (Table 3.5). The calibration was commonly reported and in most instances the model demonstrated a good calibration based on the

Hosmer-Lemeshow p-values. This suggests a model could be used to accurately provide predictions for an individual's risk. The AUC was also commonly reported for the epidemiological models. The AUC results ranged from a weak 0.57 (Spitz Model) to a very strong 0.86 (PLCO Model). The leading reported AUC results were for the PLCO, $\text{PLCO}_{\text{M2012}}$, $\text{PLCO}_{\text{M2014}}$ and Hoggart models. The results would suggest these models have the highest potential to be a successful selective screening tool. To assess the ability of a model as a selective screening tool the prediction rules need to be validated, unfortunately, only for the $\text{PLCO}_{\text{M2012}}$ Model have these been extensively evaluated. This model at the optimal risk threshold (1.34%) where it improved upon the previously implemented NLST criteria to identify a target group for screening.

At this stage, the optimal risk threshold for the remaining epidemiological models needs to be identified and the prediction rules externally validated. Janssen et al. (2008) recommends that all models should be externally validated to quantify their predictive performance through calibration, discrimination and classification predictors [45]. Additionally, comparing all the models in the same dataset is recommended. This should address limitations with the current reporting, where direct comparisons between the models is difficult. This may also identify a different leading model for distinct populations, based on ethnicity or location, which may allow different optimal screening programmes to be conducted worldwide.

The review found that the models which consider clinical factors (such as blood testing and scans) do not offer a universal improvement over epidemiological models. The discrimination was commonly reported with an AUC between $0.639 - 0.875$ and the majority of results between $0.7 - 0.79$. The lack of improvement for clinical models was highlighted by comparisons between the COSMOS (Bach), Extended Spitz and Extended LLP Models and their original models where any improvement (AUC/calibration) was minimal. The slightly disappointing results may be a result of not identifying the correct genetic factors associated with lung cancer risk. Further work is being conducted to identify the key markers which could lead to new prediction models that improve upon the existing literature. However, it is important that the new clinical markers assist in developing a robust prediction model rather than publishing a poor prediction model, as has been observed with some published clinical models. The performance of a clinical model is critical; and before a clinical model would be considered as a selective screening tool the costs to apply the model will need to be justified, which can be demonstrated by a high performing model to identify a high risk target population.

Clinical models are still at a very early stage of development and have usually been published since 2010. As a direct consequence they have not been properly tested at this stage and commonly only validated in the original article in an internal validation. A conclusion of the systematic review was that it would be beneficial if every clinical model was external validated to develop stronger conclusions about the models' performances.

In conclusion the review found some flaws in the current validations of the models that will be addressed in our study. Firstly, models have inconsistently been compared with each other in the same dataset making it difficult to draw direct comparisons between models. This has been observed for models developed for many different diseases, and the current reporting of validations across different studies has been poor [46]. Secondly, prediction rules have been poorly reported in validation studies and for some models these have never been reported (Table 3.5). It could be argued that the lack of clear guidelines, through assessing the prediction rules, on how to optimally apply models has prevented these being incorporated in selective screening programmes. Therefore, this study will address the poor validation reporting for lung cancer prediction models. The models will also be compared to the current screening programmes to evaluate if there is conclusive evidence that a leading prediction model or criteria has been identified. This will be conducted for the epidemiological lung cancer prediction models using a series of datasets received from the International Lung Cancer Consortium (ILCCO). These models will be considered because the variables will be commonly collected in the ILCCO datasets. Additionally, these models demonstrated the potential to improve upon current screening guidelines whether the UKLS or NLST criteria. Therefore, a leading model could be identified and implemented.

3.9 Summary

The systematic review identified 29 different models that predict lung cancer risk in individuals. These were classified as epidemiological, clinical assessment and TSCE models. There were ten epidemiological lung cancer prediction models identified; the Bach, LLP, Spitz, African-American, Hoggart, Pittsburgh, two PLCO versions, $PLCO_{M2012}$ and $PLCO_{M2014}$ models. The epidemiological models showed potential to be utilised as a clinical utility with some good performances for calibration, discrimination and prediction rules in validations. However, the models had often been validated independently in distinct datasets so direct comparisons between the models was difficult. The current reporting of validations was inconsistent with some validation studies not considering prediction rules. Additionally, when a model had been evaluated in multiple studies the prediction rules were often reviewed at different risk thresholds. This resulted in no clear indication into a leading model or being able to provide recommendations into the risk threshold that will allow the model to perform optimally. The standard of validations did not allow a confident recommendation into a leading model and how this may be utilised as a selective screening tool.

Based on the systematic review the next stage of research should be to address the lack of consistent reporting for models in the validations. An external validation, comparing all the models in the same dataset, is required which should provide better understanding into the different model performances.

To perform an external validation individual patient level datasets (IPD) will be collected and prepared, which is detailed in the next chapter. The collected datasets will then be analysed in the subsequent chapter to identify any limitations with the datasets that could influence the model results in the validation. Then the validation will be conducted and the results presented and analysed.

CHAPTER 4

Dataset Collection, Preparation and Imputation

4.1 Introduction

To conduct external validations on the identified lung cancer prediction models datasets were made available through the International Lung Cancer Consortium (ILCCO). Prior to performing the validation, there needs to be confidence that the datasets would allow the models to be accurately evaluated. Therefore, reported information on the variables in the datasets were reviewed and modified when required. This included harmonising the reporting of the variables so they were applicable to the models' specifications, removing participants with unreliable information and imputing missing information. The modifications made to any of the variables in the datasets were recorded and reported in this chapter.

4.2 Objectives

This chapter aims to detail how the datasets were obtained and prepared to perform an external validation of lung cancer prediction models. The chapter will:

1. Detail the ILCCO application process to acquire datasets for the project.
2. Introduce the datasets that were made available by ILCCO.
3. Present which models were applicable to which dataset based upon having complete information for each variable required by the model.
4. Present how the datasets required modifications to be compatible for the models.
    - Report any modifications to the variable information to allow it to be in the correct form required by the models.
    - Detail any participants removed because of unreliable information that could negatively affect the model validation.
5. Impute missing information in the datasets where possible.
    - Introduce the imputation objectives.
    - Present different imputation methods and identify an appropriate method.
    - Conduct and report the imputation.

4.3 Dataset Collection

To conduct an external validation on the identified epidemiological risk models, datasets with individual patient level (IPD) data were required. Applications were sent, in December 2014, to the study holders of datasets that were released to the ILCCO repository, asking for permission to use their datasets in our study.

The studies shared to ILCCO's repository were originally devised and collected for different lung cancer research objectives. Therefore, the studies had differing participant recruitment strategies and collected information for different variables which satisfied their research objectives. As a consequence, the studies that were released to the ILCCO repository vastly differed and there was no minimum requirement of essential participant information that should be provided in the datasets other than the case or control status for each participant. For the objectives of our research datasets were requested provided they collected enough variable information so that they could be applied to a minimum of two models. This minimum standard was set as this would allow the external validation to compare between models in the datasets and avoid evaluating models separately in distinct datasets.

When releasing studies to the ILCCO repository, the datasets principle investigators (PIs) were also invited to provide information on the variables collected in the study and share the questionnaire used to collect information. Approximately half of the studies provided this information. This allowed datasets which would not be applicable for our research objectives, because they did not collect important information required by the models such as a detailed smoking history, to be automatically excluded. The remaining eligible studies and all studies that did not provide a questionnaire were invited to participate in our study. An email was sent to all the study holder principle investigators informing them of our intention to use the datasets to perform an external validation and potentially use them to update prediction models. Approximately 60 datasets requests were sent with 20 responses. Unfortunately, only 10 had enough information for at least 2 of the prediction models and these were collected.

4.4 Dataset Introduction

The 10 studies obtained are listed by the names they will commonly be referred to throughout the project:

1. ReSoLuCENT
2. UCLA
3. CARET
4. New York (NY) Wynder
5. Singapore
6. New Zealand
7. CREST
8. Israel
9. ESTHER
10. MSH-PMH

These will be presented in more detail during the dataset descriptive analysis (Chapter 6) where the participant recruitment policy and population demographic will be presented and discussed.

In the obtained datasets, only the variables required by a model were provided, except in the ReSoLuCENT dataset where the complete IPD was received.

Unfortunately, none of the datasets were applicable to all the models and the Spitz and African-American model could only be applied to the CREST dataset (Table 4.1). However, the remaining models could be applied in multiple datasets, this should allow a comprehensive review of these models and allow comparisons between the models and implemented screening trials' criteria.

	Bach	LLP	Spitz	Af.-Am.	PLCO (×2)	Pitts.	Hogg.	NLST	UKLS
ReS.	X	X			X	X	X	X	X
UCLA					X	X	X	X	
CARET	X	X			X	X	X	X	X
NY Wyn.	X				X	X	X	X	
Singapore						X	X	X	
New Zea.	X					X	X	X	
CREST	X		X	X		X	X	X	
Israel						X	X	X	
ESTHER						X	X	X	
MSH-PMH	X	X			X	X	X	X	X

Table 4.1: Models and Screening Trials which could be Evaluated in each Dataset

4.5 Summary of Dataset Modifications

Once the datasets were received, the reported participant's responses for each variable was reviewed. Some of the information provided needed to be harmonised to ensure it was classified correctly for a prediction model. Additionally, all erroneous, unreliable, or incomplete information, that could negatively affect the proposed validation, was reviewed and rectified. The dataset preparation allowed every participant in the IPDs to have complete, reliable information.

Some information and participants had to removed or modified. A complete summary of dataset modifications for every variable in each dataset with justifications is presented in Appendix 12.9. A summary of the major modifications are discussed and presented in Tables 4.2 and 4.3. Imputation of missing information was considered, if possible for the variable under consideration if there was over 10% of missing information across all the variables in the dataset. This aimed to avoid removing too many participants. In datasets where there was under 10% missing information, participants with missing information were removed and the complete case analysis was preferred. Imputation in the datasets is presented in more detail later in the chapter.

Age and BMI were checked for unreasonable entries. There were no concerns with the ages provided in the datasets. However, BMI entries not in the range [8, 60] were removed as these results were unrealistic and most likely due to reporting error. Across the studies a few participants were removed for erroneous information.

Education and ethnicity the entries had to be harmonised into the form required by the PLCO models. This was problematic in the ReSoLuCENT dataset where participants could provide a free text answer for their highest level of education. Details of the way in which participants were reclassified is provided in the appendix (Section 12.10.1). For the remaining ILCCO studies participants were already grouped into their highest education level; these were then successfully reclassified into PLCO group criteria, as detailed in the Appendix.

Smoking history required some major modifications. If the smoking status was missing and could not be inferred from other information, such as a cessation age to indicate a former-smoker, then these were removed from the studies. Unrealistic information was also removed for start age and cigarettes per day (CPD). It was unrealistic to have a CPD exceeding 100 and while there were some young start ages, any age below 8 was deemed unrealistic and removed. Further checks confirmed if $StartAge + SmokingDuration =$

$CessationAge$ for all former smokers and $StartAge + SmokingDuration = CurrentAge$ for all current smokers. Finally, pack years and CPD could be inferred from each other using;

$$PackYears = \frac{CPD}{20} \times SmokingDuration \tag{4.1}$$

Checks confirmed if these values agreed when both were provided. Any incorrect information during these checks resulted in the participant being removed.

The ReSoLuCENT study collected information differently and reported CPD entries at 20, 30, 40, 50 years and currently. Participants reported their CPD for the entries that were applicable. From this the average CPD and pack years were inferred assuming participants had a linear increase or decrease between two adjacent groups. It was also assumed smoking before 20 years of age remained constant at the 20-year rate, additionally, former-smokers who quit some years after their last CPD entry were assumed to smoke consistently at the last provided rate until they quit. This assumption allowed us to determine the average CPD using the most relevant information available however, there is a concern this could be incorrect. Participants may fluctuate from the last known value; it could be expected individuals may increase smoking once they have started (to the CPD rate observed at 20) and may gradually decline before eventually stopping (since the last known CPD provided). However, with no information on which to further ascertain a more in-depth smoking history, assuming smoking remained constant at the last observable rate, provided the most reliable estimate for CPD. Additionally, some participants had missing information for one of these entries despite occurring during their smoking range. Rather than removing participants with missing information at this stage, attempts were made to impute the missing information which is presented in more detail during the dataset imputation (Section 4.10).

Family history of cancer required data preparation. This needed to be classified for the models into the following;

1. Any cancer
2. Any cancer (excluding skin melanoma)
3. Smoking related cancers
4. Lung cancer

Every cancer in the later groups are also included in the earlier groups. The cancers were classified using the Cancer Research UK website using their information on the causes of each type of cancer [102]. Each reclassified cancer is presented in the Appendix (Section 12.11). Checks were made to ensure that the cancers were only recorded for first degree relatives; i.e. parent, sibling, or child. Additionally, the age of onset was reported for each relative and, in the case of multiple relatives with a positive family history, the youngest age was recorded for each cancer subgroup.

The final variables to review were the exposures and lung conditions. These were binary responses with the majority of participants having complete information. To avoid removing participants, we explored methods to impute missing information using additional information for the participants in the dataset where possible.

At this stage the datasets were modified and the variables harmonised to be applicable to the models. The final stage was to attempt to impute missing information and then the datasets will be prepared to conduct the external validation.

Variable	ReSoLuCENT	UCLA	CARET	NY Wynder	Singapore
Raw Participants	1,363	1,651	2,381	10,072	1,087
Eligibility	29 removed as stated as ineligible				
Age					7 removed for missing information
Gender	1 missing information Considered for imputation				
Ethnicity	2 missing information Considered for imputation	46 removed for missing information		2 removed for missing information	
BMI	205 removed for missing or unreasonable entries	4 removed for missing information	15 removed for missing or unreasonable entries	534 removed for missing or unreasonable entries	
Education	7 missing information. Considered for imputation. Participants reclassified	1 removed for missing information	197 removed for missing information	14 removed for missing information	
PMT[1]			1 removed for missing information		
Smoking Status	183 removed for missing information			1 removed for missing information	1 removed for missing information
Start Age	4 removed for missing information	23 removed for missing or unreasonable entries		255 removed for missing or unreasonable entries	6 removed for missing information
Cessation Age	6 removed for missing or unreasonable entries				
Smoking Duration					
CPD/Pack Years	372 missing information. Considered for imputation.	23 removed for missing information			3 removed for missing information
Quit Duration					
ETS					16 removed for missing information
FH[2] of Cancer	169 removed for missing or unreliable age of onset	Determined by ICD9 Code	Determined by ICD9 Code	Determined by ICD9 Code	
Asbestos Exposure	46 missing will be attempted to be imputed				
Dust					
Hay fever					
Emphysema	1 missing information will be attempted to be imputed				
COPD	1 missing information will be attempted to be imputed				
Pneumonia			6 removed for missing information		

[1]PMT = Prior Malignant Tumour; [2]FH = Family History

Table 4.2: Summary of Participants Removed from the ILCCO Datasets (1/2)

Variable	New Zealand	CREST	Israel	ESTHER	MSH-PMH
Raw Participants	403	948	665	413	3,176
Eligibility					
Age			1 removed for missing information		1 missing information Considered for imputation
Gender					
Ethnicity	28 removed for missing information				108 missing information Considered for imputation
BMI					125 missing information Considered for imputation
Education					116 missing information Considered for imputation
PMT[1]					
Smoking Status		2 removed for missing information	8 removed for missing information	6 removed for missing information	373 removed for missing or unreliable information
Start Age	7 removed for missing information	10 removed for missing information		7 removed for missing information	1 removed for unreliable information
Cessation Age			1 removed for unreliable information		
Smoking Duration					
CPD/Pack Years	21 removed for missing or unreliable information	2 removed for missing or unreliable information	5 removed for missing or unreliable information	15 removed for missing or unreliable information	
Quit Duration					
ETS					
FH[2] of Cancer	Determined by ICD9 Code	Determined by ICD9 Code			Determined by ICD9 Code
Asbestos Exposure	2 removed for missing information				214 missing information Considered for imputation
Dust					
Hay fever		Identified from ICD9 Code			
Emphysema		Identified from ICD9 Code			240 missing information Considered for imputation
COPD		Identified from ICD9 Code			249 missing information Considered for imputation
Pneumonia		Identified from ICD9 Code			215 missing information Considered for imputation

[1]PMT = Prior Malignant Tumour; [2]FH = Family History

Table 4.3: Summary of Participants Removed from the ILCCO Datasets (2/2)

4.6 Dataset Imputation Objectives

The within-study imputation aimed to avoid removing a large volume of participants in the datasets (exceeding 10%) by estimating the missing information. A larger dataset may allow a fairer validation of the models as removing a large volume of participants may make the dataset less representative of the original population.

It is imperative if conducting an imputation that there is confidence the imputed values are accurate estimates. Incorrect imputations could negatively affect the validation results by generating biased predictions for the participant's risk of developing lung cancer using the prediction models. It would be better to remove the participants and have smaller datasets than keep participants with incorrect, misleading variable estimates.

Imputations will only be conducted if a suitable method can be identified and there is evidence in the dataset the imputation can be conducted correctly. If an imputation is conducted the imputed values will also be assessed to confirm that the new values are realistic, to allow confidence in the estimated values, before using them in a validation.

4.7 Dataset Imputation: The "Missingness" of Data and Rubin's Rules

Prior to imputing the missing information, it is important to demonstrate that the variables which require imputing are "Missing at Random" (MAR), rather than "Missing Completely at Random" (MCAR) or "Information Missing/Missing Not at Random" (IM/MNAR). This is defined as the missingness of the data.

Data is MAR if the likelihood of the data being missing is independent of what the missing value is, but can be explained by the observed data. If the data is MAR then the missing values can be imputed [108, 109]. When the data is MCAR, where the missingness of data does not depend on the observed or unobserved data [108, 109] and MNAR, where the missingness of data depends on the unobserved data [108, 109], it should not be imputed.

It can be difficult to determine, based on only the observable data, if the data points are MAR or MNAR. In some instances, to impute the data, it has to be assumed the information is MAR rather than MNAR, although for low levels of missingness the imputed estimates should be robust. Additionally, very extreme scenarios of MNAR would have to occur for the imputation to be unreliable.

A review of the data will be conducted to assess if the data is MAR. The first test will analyse the observed values for the variable with the missing information. This will review whether the information for the variable is likely to be withheld as it is sensitivity information. This may occur for information such as CPD, where participants may be reluctant to reveal how much they smoke if the number is high. A review of the observed data points aims to assess if it is likely the information has been concealed for extreme values. This test will review, as best as possible using only observed data point, if the data is MAR or MNAR.

A second test aims to assess if the data is MAR or MCAR. It evaluates whether the likelihood of the data point missing can be explained by the observed entries for other variables. This will be evaluated through t-tests across the variables, X_i, grouped on whether variable, X_j, is missing or observed. This can be tested in *Stata 12* by the following code;

```
ttest X_i, by (miss_X_j)
```

This test evaluates whether the observed data in both groups are not significantly different. A $|T| > 1.96$ means the null hypothesis is rejected and there is a significant difference between the two groups. Then it can be determined X_j is MAR dependant on X_i rather than MCAR.

Once the testing to evaluate whether the missingness is MAR in comparison MCAR and, as best attempted to justify using only the observable data, that the missingness is MAR against MNAR, then the imputation can be conducted.

Upon identifying an appropriate method to impute the missing information (Section 4.8) the imputation will be undertaken. The imputation will be conducted 11 times independently. This will produce 11 distinct datasets where the imputed values can differ between the datasets. 11 imputations were selected to allow confidence there is an adequate level of reproducibility in the validation results. This is achieved by considering approximately the same volume of imputations and the percentage of missing data [110]. At this stage, a review of the imputed values will be conducted to ensure the estimates are reasonable. Once this has been conducted, the models will then be applied to the 11 distinct datasets and the model performance in each of the datasets evaluated.

When presenting the model performance, the aggregated results across the 11 imputed datasets will be presented where possible. This is calculated using Rubin's Rules which combines the estimates and also allows for variability in the model performance across the datasets by incorporating this uncertainty into the 95% confidence intervals. The results for the Brier Score, sensitivity and specificity can be combined using Rubin's Rules as follows;

$$\hat{\theta}_{MI} = \frac{1}{J}\sum_{n=1}^{J}\hat{\theta}_n \tag{4.2}$$

Where $\hat{\theta}_n$ is the Brier Score, sensitivity, or specificity result in the n^{th} dataset. The PLR results will be calculated from the sensitivity and specificity results obtained from Equation 4.2.

The AUC will also be calculated using Equation 4.2 but the standard error is calculated to determine the 95% confidence interval for the AUC. This is calculated to include the within-imputation variance, $\hat{\sigma}_w^2$ and between-imputation variance, $\hat{\sigma}_b^2$;

$$\hat{\sigma}_w^2 = \frac{1}{J}\sum_{n=1}^{J}\hat{\sigma}_n^2 \tag{4.3}$$

$$\hat{\sigma}_b^2 = \frac{1}{J-1}\sum_{n=1}^{J}\left(\hat{\theta}_n - \hat{\theta}_{MI}\right)^2 \tag{4.4}$$

Where $\hat{\theta}_{MI}$ is determined from Equation 4.2. Then the standard error for the combined estimate is as follows;

$$\hat{\sigma}_{MI} = \left(1 + \frac{1}{J}\right)\hat{\sigma}_b^2 + \hat{\sigma}_w^2 \tag{4.5}$$

Now the standard error has been calculated the 95% CI can be determined by;

$$95\%\text{ Confidence Interval} = \left[\hat{\theta}_{MI} - \hat{\sigma}_{MI}t_{(v,0.975)}, \hat{\theta}_{MI} + \hat{\sigma}_{MI}t_{(v,0.975)}\right] \tag{4.6}$$

$$\equiv \left[\hat{\theta}_{MI} - 1.96\hat{\sigma}_{MI}, \hat{\theta}_{MI} + 1.96\hat{\sigma}_{MI}\right] \tag{4.7}$$

The Hosmer-Lemeshow result will not be aggregated using Rubin's Rules. The result from the aggregated dataset will be presented. However, if the result in any of the 11 imputed datasets changes for the critical p-value of 0.05 all results will be disclosed.

The tests to evaluate the missingness of the data and to combine the validation results using Rubin's Rules will be employed to ensure there can be confidence using imputed values to evaluate the models' performance. The next stage is to identify a practical method to impute the missing information.

4.8 Review of the Imputation Methods

A review of the available methods to impute missing information is presented to assess if there is a method that is suitable for the purpose of our study.

4.8.1 Complete Case Analysis

The first approach is to remove any participants with missing information. This is referred to as complete case analysis. There are advantages to this method as it avoids concerns with inappropriate imputations. Poor imputations for a participant will see the models estimate inaccurate risks for the likelihood of lung cancer developing. This will influence the performance of the prediction models when validated. Complete case analysis can be preferred as there are no concerns with subsequent results generated using imputed information.

However, removing all participants with incomplete information could drastically reduce the IPD size. A severely reduced dataset may compromise the validation results. There is also potential the complete case analysis may cause bias [111]. By removing participants with incomplete information this can remove evidence that should be considered if the missingness of data is MAR [111]. The final population may become biased, when removing participants who have missing information for a variable, which can create unreliable validation results, especially if a large proportion of participants would be removed. This is a concern with complete case analysis and why imputation should be considered, particularly if a large proportion of the study would be removed.

4.8.2 Mean Imputation

Mean imputation assigns the mean value, for all the missing values for a variable, using the complete information for this variable [107]. This is not a preferential method as it fails to assign accurate estimates for the missing variable. The method has the potential to create bias and understates the variability for the missing information [112].

The purpose of the imputation is to create a complete dataset in which the prediction models can be applied and reviewed. Therefore, assigning the same values for participants is unreasonable and unrealistic. As a consequence, the performance of the models in the dataset will be reduced, the lack of variability for the variable, will restrict the models' ability to discriminate between individuals with or without lung cancer, or generate accurate risks for an individual.

4.8.3 Last Value Carried Forward

The last value carried forward method imputes the information using previous information for a participant. This method is used when data is collected 'longitudinally' over time and there are follow-ups for the participants. The last plausible value for the participant is then carried forward [108]. This method is most likely to be considered for clinical trials that conduct follow-ups, however the datasets we received did not record the majority of the data 'longitudinally'. There is the potential for this method to be used for the smoking quantities in the ReSoLuCENT dataset which are recorded at 20, 30, 40 and 50 years. It could be reasonable to assume there would not be a large variance between smoking quantities at different ages for an individual.

There are limitations to this method. Basing the imputation on as much information as possible, rather than the last plausible value, can generate more accurate estimates for the variable. Additionally, this can underestimate the variability that can occur in a variable across time between follow-ups and create bias [112]. Overall, this method is not the most appropriate for the participant information received across the datasets.

4.8.4 Simple Random Imputation

This method imputes a random value for the missing information of a variable based on the observed results for this variable [108]. The purpose of this method is to create random noise for the variable and account for the variability that is observed [112]. To achieve this the imputed values are based on the observed information for the variable recorded for different participants, rather than using additional information for

the participant with the missing information. A random value is then assigned for the missing information based on the evidence for this variable.

These are concerns in using simple random imputation to impute the missing information. The estimated values are not based upon available evidence for the participants with missing information. Therefore, the imputed values may be unreliable for the participant. This concern is magnified when imputing missing information for multiple variables, all of which have the risk of being inappropriate for the observed participant profile. Consequently, this will negatively affect the model's ability to accurately predict the likelihood of an individual developing lung cancer based upon these imputed values.

4.8.5 Regression Predictions to Perform Deterministic Imputation

This method uses observable information specific to the participant in a regression model to predict the missing values [108]. By using the observed data for each participant, this method aims to accurately estimate the missing variable. This can allow more confidence in the imputed variable as the method utilises the most relevant explanatory variables in the estimate. This method can be successful when there is a large range of additional information available for each participant to build a more accurate participant profile.

This method is most appropriate for single imputation, when there is missing information for only one variable [108]. This method has some limitations for imputation of multiple variables. If participants have missing information for some of explanatory variables then these cannot be used to impute the missing information. Additionally, if imputed values are then used in further imputations, there is no consideration of the uncertainty of the accuracy of these imputed values. Any error in imputing the first variables will distort the estimates in the subsequent rounds of imputation. It would be preferential if there was a penalty to account for the uncertainty in these values. This concern could be avoided by only considering variables with complete information to base the imputation. However, in a dataset with missing information, it is unlikely the remaining variables will have complete information. It would also be unwise to exclude an explanatory variable because it reported some missing information. Then the imputation would be based on weaker information and lower the accuracy of the imputations.

4.8.6 Random Regression Imputation

Random regression imputation considers relevant participant explanatory variables in a regression equation and incorporates randomness for the variable by including an error value in the estimate [108].

Similar to Regression Prediction to Perform Deterministic Imputation (Section 4.8.5) this method is most appropriate for single imputation rather than missing information for a range of variables. If there are imputed values used in subsequent imputations then there is a greater potential for poor predictions. This could be counteracted by assigning a penalty to variables considered in the imputation that have previously been imputed. Additionally, only variables with complete information could be considered in the imputation but this reduces the volume of variables to accurately estimate the missing information.

Random regression imputation has some additional limitations; by incorporating randomness the estimates differ from the most likely estimate for the participant. This aims to incorporate the variability that observed in the variable [112]. However, this is an inappropriate method if the purpose of the imputation is to create accurate participant profiles. The method could drastically deviate the estimated risk from the most likely estimate.

4.8.7 Matching

This method matches a participant with missing information to another participant with a similar profile that has complete information based on the observed information. The missing value is replaced by the value the other participant has for that variable [108]. Matching considers the relationship between participants using the observed data to pair participants with incomplete information with participants

Imputation Method	**Method Application**	**Method Advantages**	**Method Disadvantages**
No Imputation	No imputation. Participants removed. Keep complete case info.	Accurate information. Confident validations.	Reduced population. Increased uncertainty in validation results.
Mean Imputation	Mean for variable applied to all missing values.		Inaccurate imputed values. Not based on participant info. Assigned same values. Model cannot discriminate between participants.
Last Value Carried	Value from last point taken. Consider the last point of follow-up.	Consider relevant participant information.	Datasets do not have follow up information. Different variables will need to be considered that are not relevant. Snap shot of history can be misleading.
Simple Random Imputation	Random value assigned. Based on the existing variable values.		Random value assigned. Not based on any participant info. Negative influence of validation results.

Table 4.4: Summary of Imputation Methods *(Methods 1-4)*

Imputation Method	Method Application	Method Advantages	Method Disadvantages
Regression Predictions	Additional variables considered. Regression models estimates new value.	Considers relevant predictors. Accurate predictions for participants.	Imputed values could be reused in an additional imputation without allowing for uncertainty of imputations. Need complete information for explanatory variables or imputations for every missing variable separately.
Random Regression	Regression model used to predict values and error is accounted for with a random error value assigned.	Uses reliable participant information. Error allows for unexpectedness observed in real world scenarios.	Desire accurate predictions. Error could distort true value affecting model validations. Values could be reused in further imputations.
Matching	Participants are matched over similar predictors. The same value is assigned as the closest match.	Uses similar participant information who were captured in the study design. No unrealistic values generated.	Difficulty choosing variables to match over. Could subgroup predictors too much and create distorted imputations. Participants still differ.
Multiple Imputation	All variables that want to be considered are imputed simultaneously. Logistic or logit regression. Results are run until the imputed values have stabilised.	Relevant participant information considered to create accurate predictions. Imputed values used in imputations are considered by stabilising the results. Can consider continuous, categorical and dichotomous variables. Range of values possible as not matched or restricted to observed values.	Averaging reduces extreme values and gathers towards the mean.

Table 4.5: Summary of Imputation Methods *(Methods 5-8)*

with complete information. This allows the method to generate realistic estimates for the missing values as they are observed in the dataset.

It is imperative with this method that the participants are matched over key variables that most explain the missing value. It is important that multiple variables to match participants are identified and weighted, otherwise the participants could have similar profiles to multiple other participants but they could have drastically different values for the variable that will be imputed. This method works best when matching over continuous variables [108]; for dichotomous outcomes there will be many participants in the dataset that have a matching profile.

This method can be limited in IPDs where there is missing information for multiple variables. This makes it more difficult to successfully match participant profiles to generate accurate estimates for the missing information.

4.8.8 Multiple Iterative Regression Imputation

Multiple imputation using chained equations (MICE) imputes missing values across multiple variables simultaneously [113]. This is the only method that imputes information where the explanatory variables can have missing information. As a result this allows a large volume of evidence to be considered in the imputation regardless of whether there is complete or missing information for the variable. MICE can consider categorical and dichotomous data through ordinal and logit functions allowing it to be used for all forms of missing data.

To apply the method for the variables considered, any missing values for x_1 are imputed using $x_2, ..., x_n$. When imputing this value if any of the variables $[x_2, x_n]$ have missing information they are not considered in the imputation of x_1. Next, x_2 is imputed considering the imputed value or already compete value for x_1 and $x_3, ..., x_n$ [113]. After a complete cycle this is re-run until the imputed results converge and are stable between each cycle. This reduces the likelihood of inappropriate imputations in an earlier cycle when there may be limited information.

While this method can deal with multiple missing values it is important that the participants have some complete information. Otherwise, the imputation will may be based on insufficient evidence and may generate poor estimates. Participants with too much missing information across all the variables considered in the imputation should be considered for removal from the IPD before the imputation is conducted.

An advantage of this method is whether the variables originally have complete or missing information they can still be considered during the imputation. Then by stabilising all variables with missing information it can create confidence the estimated values are reasonable. This also allows the key explanatory variables to be used without concern even when these values contain missing information for some participants.

4.8.9 Imputation Methods Summary

In the ILCCO datasets received there is missing information across multiple variables. Therefore, MICE is the most appropriate method to generate accurate estimates for the missing information. The other identified methods are more appropriate for missing information for a single variable. However, when there are participants with missing information across multiple variables several independent imputations would have to be considered. There is a concern that any erroneous imputed values could be considered in subsequent imputations. Additionally, these methods need a set of core explanatory variables with complete information to conduct the original imputation. This can reduce the evidence on which the original imputation is based. In contrast MICE will consider all the variables in one single imputation whether they have complete or missing information. This method also considers whether the variable in the imputation has been imputed and then considered in the imputation for another variable. MICE accounts for this by recursively imputing the imputation until all the imputed values have stabilised.

4.9 Conducting the Imputation

MICE will be conducted in the ReSoLuCENT and MSH-PMH dataset as there was more than 10% of participants with some level of missing information across the variables.

The first stage to conducting the imputation will review if the missingness of data is MAR or MNAR. This will be achieved by evaluating the observed data and assessing if there are any indications the data has been purposely withheld. However, since there are low levels of missing data the imputed values should be robust whether the missingness of data is MAR or MNAR. Next the association between variables will be reviewed through t-tests. This will evaluate if the data is MAR, as desired, or MCAR. This is explained in more detail in Section 4.7.

One round of MICE will then be conducted separately in each of these two datasets which will impute missing information across all the variables. This round of imputation will consist of 11 independent iterations. This aims to eliminate the possibility of the imputation for specific participants being unrealistic. 11 imputations were selected as this seemed reasonable to allow the imputation results to not be significantly skewed by an unusual result for one imputation. An odd number also allowed a clear result for imputed variables with a dichotomous outcome.

When conducting the imputation multiple variables including both complete and missing information will be considered in the imputation. However, case-control status will not be considered in the imputation as this may bias the imputed values, such as assigning a heavier smoking history or prior lung conditions to individuals with lung cancer. This will bias the models' performance and improve their discriminative ability [115].

The imputation will be reported to review whether the imputation has been successful and the imputed values are reasonable. This will assess whether there are any erroneous values imputed. However, to avoid incorrect imputations, such as a negative CPD, the imputed values will be matched to their nearest neighbour based on the observed results when performing the imputation.

The imputed dataset will then be used in the model validations and the uncertainty in the imputed values will be considered using Rubin's Rules, across the 11 independent imputations, as presented in Section 4.7.

The imputation in the ReSoLuCENT and MSH-PMH datasets will now be conducted, presented and reviewed.

4.10 ReSoLuCENT Dataset Imputation

There were 18% of participants in the ReSoLuCENT dataset who had missing information for at least one variable where imputation was attempted.

A summary of the missing information is presented in Figure 4.6. The figure shows there was a large volume of missing information for the smoking variables (S31_1ya, S31_20, S31_30, S31_40 and S31_50). The ReSoLuCENT dataset reported information on how much participants smoked last year and at 20, 30, 40, 50 years.

Table 4.6 also displayed a high level of missing information for the age of onset of cancers. However, this value was only included if there was a positive family history. All missing values are accounted for, as the participants did not have a positive family history of cancer. As a result this was not considered in the multiple imputation.

There were 46 participants who did not report asbestos exposure and 17 who did not report their ETS exposure. There was missing information for gender, ethnicity, education and other lung conditions which will be imputed, although this was only missing for a small number of participants (1-7).

Variable	Missing	Observed
Gender	1	1,077
Ethnicity	2	1,076
Education	7	1,071
Smoking Last Year	342	736
Smoking 20	32	1,046
Smoking 30	64	1,014
Smoking 40	146	932
Smoking 50	324	754
Age of Any Cancer	431	647
Age of Any Cancer (Excluding Melonoma)	433	645
Age of Smoking Cancer	534	544
Age of Lung Cancer	718	360
Asthma	1	1,077
Emphysema	1	1,077
COPD	1	1,077
Dust	4	1,074
Asbestos	46	1,032

Table 4.6: ReSoLuCENT Dataset: Identifying the Missing Information

Now that the variables that required imputation had been identified, the next stage was to evaluate if the data was MAR. This missingness of information was unlikely to be MNAR as they agreed to participate in a hospital study for lung cancer. Therefore, it is unlikely the information would then be purposely withheld.

The next stage reviewed if the data was MAR in comparison to MCAR. A t-test was conducted using the observed information for all the variables which had at least 10 participants with missing information and additional variables. The results are displayed in Table 4.7.

Variable One	Variable Two	T-Test
Smoked at 50	Smoked Last Year	-6.14
Smoked at 50	Smoked at 20	-2.70
Smoked at 20	Smoked at 30	-2.96
Smoked at 20	Smoked at 40	-3.83
Smoked at 20	Smoked at 50	-7.18
Asbestos Exposure	ETS Exposure	-2.85
Gender	Asbestos Exposure	2.36

Table 4.7: T-Test Results by Missing Information in the ReSoLuCENT Dataset for Variables considered in Imputation

The data was shown to be MAR in comparison to MCAR with $|T| > 1.96$ for all the variables considered in the imputation. The smoking information was shown to be MAR in comparison to MCAR dependant on entries for different smoking quantities.

ETS was MAR dependant on asbestos exposure; an increased asbestos exposure rate was shown in individuals who did not report their ETS exposure. Finally, asbestos exposure was MAR dependant on gender. Males were shown more likely not to report whether they had been exposed to asbestos. Now the variables considered in the imputation have been shown to be MAR the imputation can be conducted.

The imputation was successfully conducted in *Stata 12* using the following code;

```
\textit{*Set the imputation}
mi set wide
\textit{*Register the variables for imputation}
mi register imputed AgeRegistered BMI ETSNumeric PMT Pneumonia ///
SmokingStatusNumeric StartAge CessationAge Asbestos COPD Emphysema Asthma ///
LastYear Smoked20 Smoked30 Smoked40 Smoked50 PLCOEducationScale ///
EthnicityScale Gender
\textit{*Impute the missing information}
mi imp chained (pmm) AgeRegistered BMI StartAge CessationAge LastYear Smoked20 ///
Smoked30 Smoked40 Smoked50 (logit, augment) ETSNumeric PMT Pneumonia Asbestos ///
COPD Emphysema Asthma (ologit, augment) Gender SmokingStatusNumeric ///
EthnicityScale PLCOEducationScale, add(11) rseed(19) dots
```

All missing values were all imputed.

The next stage was to evaluate the missing information to assess if the results were reliable. A review of the imputed results indicated the missing information was successful imputed. There were no unexpected values as the imputed values were matched to observed values after imputation.

In conclusion the imputation in the ReSoLuCENT dataset was successfully conducted to avoid removing 18% of participants. The missingness of data was shown to be MAR in comparison to MCAR. The missing values were imputed and there can be confidence using these imputed values in an external validation of models.

4.11 MSH-PMH Dataset Imputation

There were 26% of participants who had missing information for at least one variable in the dataset. Attempts were made to impute this missing information rather than remove these participants.

The level of missing information for each variable is presented in Figure 4.8. There were high levels of information missing for lung conditions; pneumonia, COPD and emphysema as well as asbestos exposure, ethnicity, education and BMI. Finally, there was one missing entry for age and 37 participants who did not provide ETS information.

Variable	Missing	Observed
Age Registered	1	2,743
Ethnicity	108	2,636
Education	116	2,628
BMI	123	2,619
ETS	37	2,707
Asbestos	214	2,530
Pneumonia	215	2,529
COPD	249	2,495
Emphysema	240	2,504
Eczema	95	2,649

Table 4.8: MSH-PMH Dataset: Identifying the Missing Information

As the MSH-PMH study recruited participants in a hospital setting, it is unlikely the information is purposely withheld. Additionally, for the small level of missing information the imputation should be robust whether MAR or MNAR. Then for all the variables that required imputation, except for age with only 1 missing entry, t-tests were performed to evaluate if the missingness of data was dependant on the observable information for another variable. The t-test results are presented in Table 4.9. All the variables

were shown to be MAR in comparison to MCAR as the missingness of the data was dependant on the results for different variables.

Variable One	Variable Two	T-Test
Age	Ethnicity	3,41
Age	Education	-5.01
Age	BMI	-4.31
CPD	ETS	-3.38
Age	Asbestos	-5.99
COPD	Pneumonia	-9.36
Emphysema	COPD	-9.44
COPD	Emphysema	-17.01

Table 4.9: T-Test Results by Missing Information in the MSH-PMH Dataset for Variables considered in Imputation

The next stage was to conduct the imputation and assess if all missing information was successfully imputed. The following code was used in the imputation;

```
\textit{*Set the imputation}
mi set wide
\textit{*Register the variables for imputations}
mi register imputed AgeRegistered EthnicityScale PLCOEducationScale BMI ETS ///
Asbestos Pneumonia COPD Emphysema Eczema CPD Gender PMT LCC
\textit{*Impute the missing information}
mi imp chained (pmm) AgeRegistered BMI CPD(logit, augment) ETS Asbestos ///
Pneumonia COPD Emphysema Eczema Gender PMT LCC (ologit, augment) ///
EthnicityScale PLCOEducationScale, add(11) rseed(19) dots
```

The imputation was successfully conducted and all missing information imputed. The estimated values were reasonable as a review of the new values did not identify any unexpected values. These values will be used to apply the prediction models in the dataset.

4.12 Summary

Ten datasets were collected from ILCCO. These could be applied to multiple lung cancer prediction models although not all models were applicable to each dataset. The datasets were prepared so that the variable information was complete for each model that was applicable.

MI used to impute missing information in instances of large levels of missing information (exceeding 10%). This was conducted if the data was shown to be MAR in comparison to MCAR as the complete case analysis may lead to biased results. Attempts were made to demonstrate the missingness of data was MAR in comparison to MNAR, however, for small levels of missing information the imputations should be robust. Two imputations were conducted, using MICE, in the ReSoLuCENT and MSH-PMH datasets.

The datasets were now fully prepared for the models to be applied and validated. The next stage is to review the datasets, their population demographic and their participant recruitment strategy.

CHAPTER 5

Dataset Descriptive Analysis

5.1 Introduction

This chapter will present the characteristics of the ten studies which will be used in the external validation. The datasets were uniquely collected for different study objectives; this may result in some unusual study populations. The chapter will provide a descriptive analysis which will highlight any similarities, differences, strengths and limitations of the datasets available by reviewing how participants were recruited for the study and population demographic. This will give an indication into any models that could have a sub-par performance in a dataset due to distinct dataset characteristics. This follows the guidelines promoted by TRIPOD in how to report any influences and limitations in the datasets used in a validation study to allow confidence in the validation results [46].

5.2 Objectives

To perform a dataset descriptive analysis, the chapter will:

1. Present how participants were recruited for each study.
2. Present the population demographic for every variable required by the models for each dataset.
3. Discuss the dataset collection and population demographic to assess if this could influence any model performance.

5.3 ReSoLuCENT Dataset

The Resource for the Study of Lung Cancer Epidemiology in North Trent (ReSoLuCENT) is a study in the National Institute for Health Research Clinical Research Network (NIHR CRN) Portfolio. It is an ongoing study (at the time the dataset was collected) and started data collection on 06.04.2006 with a proposed finish on 31.08.2015. The dataset was made available in March 2015, as the study drew to a conclusion. The participants were collected from various locations around the United Kingdom [123];

1. Airedale
2. Chesterfield
3. Doncaster
4. Edinburgh
5. Manchester

6. Rotherham
7. Sheffield
8. Southampton

The study entry conditions were as follows [123];

"You can enter this trial if you have lung cancer or you need to have an operation for suspected lung cancer. As well as this you can enter this trial if you

- *Are 60 years old or less,* ***OR***
- *Have a first degree relative aged 60 or less who has lung cancer,* ***OR***
- *Have two or more first or second degree relatives of any age who have lung cancer*

You may be able to be in the 'control' group if you

- *Are the partner of someone taking part in the study,* ***OR***
- *Are a first degree relative of someone taking part in the study and are at least 18 years old*

A first degree relative is a parent, child, brother or sister. A second degree relative is an aunt, uncle, grandchild or grandparent, for example."

This study was designed to investigate lung cancer, considering both small cell and non-small cell type cancers. The protocol stated "The aim of this study is to collect good quality epidemiological and biological data from lung cancer patients with a family history of the disease or with early onset lung cancer." [124]

The ReSoLuCENT study was designed to examine early onset lung cancer and recruited participants accordingly. This results in a young collection of lung cancer patients in whom lung cancer is a rare occurrence [1]. This is presented in the population demographic in Table 5.1 where the diseased and disease free individuals are both concentrated around 50 years of age.

The ReSoLuCENT dataset is a case-control study and controls were recruited for the study if they were a first degree relative of an identified case [124]. By initially collecting controls who are related to a lung cancer case, a high proportion of controls in the study will have a positive family history of lung cancer. In an attempt to counteract this concern, the ReSoLuCENT study proposed collecting 1,000 cases and 1,500 controls with each case "to be matched with at least one control (ideally one related and one non-related)" [124]. In a general population only a small proportion of participants would have a positive family history and despite the additional controls collected the dataset still reported a positive lung cancer family history in approximately 50% of controls (Table 5.1).

5.4 University of California, Los Angeles Study (UCLA)

The University of California, Los Angeles (UCLA) study is an American based case-control study. The study was collected in LA using the following methodology [125];

"All subjects in this study were:

(a) residents of Los Angeles County at the time of diagnosis for cases or at the time of recruitment for controls

(b) were 18 to 65 years of age during the study period, 1999 to 2004;

(c) spoke either English or Spanish, or had translators available at home."

The study did not focus solely on lung cancer incidents, but the data received from ILCCO repository contained 611 lung cancer cases and 1,040 control participants. The original study also identified 601 individuals who had oral, pharynx, larynx and oesophageal cancers, but these were not provided in the dataset recieved. Individuals with lung cancer were collected "by the rapid ascertainment system of the Cancer Surveillance Program for Los Angeles County, which is administered by the Keck School of Medicine and the Norris Comprehensive Cancer Center at the University of Southern California" [125]. Controls "were individually matched to cases on age by decade, gender and residential neighbourhood" [125].

Another concern with the participant recruitment is limiting the maximum age of inclusion to 65 years for cases and controls. The original study purpose was to evaluate lung cancer onset in unusual young cases. This creates a final dataset where the participants are considered at low risk of developing lung cancer. Therefore, the age range for this study is younger than the age range that any lung cancer screening programme would likely consider. The young population is highlighted in Table 5.1 with the cases and controls reporting an average age of 50 years old.

One final concern with the UCLA dataset was that participant's education status was pre-categorised into five levels. However, the PLCO models categorised participants' education status into six levels. The participants in the UCLA could not be reclassified into the six PLCO levels (Section 12.4) and so no participants were classified as "More than high school" as the other UCLA categories coincided with a PLCO Model category.

5.5 CARET Dataset

The CARET dataset obtained is a subset of the study used to create the Bach Model. Therefore it is likely the results from the validation will be favourable for the Bach Model. The dataset will provide a high risk population to evaluate the prediction models.

In contrast to the ReSoLuCENT and UCLA studies, this study recruited high risk participants who were "heavy smokers and asbestos-exposed workers" [126]. To recruit high risk participants, never-smokers were excluded from the study.

The participants were collected at various centres across the United States, with collections taking place in "Seattle, Portland, San Francisco, Baltimore, Connecticut and Irvine" [126]. The study protocol set out to recruit participants for their trial if they satisfied [126];

"CARET is a randomized, double-masked, placebo-controlled chemoprevention trial in two groups of adults at high risk for lung cancer:

(a) men and women age 50-69 years with a history of at least 20 pack-years of cigarette smoking

(b) men age 45-69 years with evidence of extensive occupational exposure to asbestos and a history of cigarette smoking."

"Through April 30, 1993, CARET had randomized 4,000 asbestos-exposed workers, exceeding accrual goals at all five CARET asbestos centres and 11,105 heavy smokers" [126].

The study design means that the participants were mainly males (Table 5.1) however; lung cancer prevalence is evenly distributed among men and women so this is not a concern. Excluding never-smokers might be a problem when coupled with the high volume of asbestos exposed workers. These two selection criteria means that participants are at increased risk of developing lung cancer. These participants may all be considered high enough risk, due to their smoking and asbestos history, to be considered for screening.

The age restriction during recruitment should not be a major concern as most target screening populations are included in this 45-70 years age group. Indeed, this is the main age bracket of individuals that would benefit from screening.

Variable	Dataset	ReSoLuCENT Case	ReSoLuCENT Control	ReSoLuCENT P-Value	UCLA Case	UCLA Control	UCLA P-Value	CARET Case	CARET Control	CARET P-Value	NY Wynder Case	NY Wynder Control	NY Wynder P-Value	Singapore Case	Singapore Control	Singapore P-Value
	Total (%)	656 (60.85)	422 (39.15)		597 (37.86)	980 (62.14)		695 (33.49)	1,380 (66.51)		4,789 (51.60)	4,492 (48.4)		314 (29.35)	756 (70.65)	
	Age (s.e.)	55.49 (0.266)	49.81 (0.622)	< 0.0001	52.27 (0.222)	49.86 (0.234)	< 0.0001	60.37 (0.204)	59.45 (0.150)	0.0004	60.56 (0.142)	60.28 (0.151)	0.1698	65.18 (0.70)	63.13 (0.459)	0.0150
Gender	Male (%)	318 (48.48)	157 (37.20)	0.0003	297 (49.75)	582 (59.39)	0.0002	473 (68.06)	938 (67.97)	0.9682	2,917 (60.91)	2,568 (57.17)	0.0002	0 (0)	0 (0)	.
	Female (%)	338 (51.52)	265 (62.80)		300 (50.25)	398 (40.61)		222 (31.94)	442 (32.03)		1,872 (42.83)	1,924 (100)		314 (100)	756 (70.65)	
	BMI (s.e.)	26.51 (0.214)	27.81 (0.293)	0.0003	26.30 (0.235)	27.48 (0.178)	0.0001	26.89 (0.171)	27.32 (0.132)	0.0560	25.28 (0.062)	26.35 (0.073)	< 0.0001	NA	NA	
Education	< High School (%)	238 (36.28)	89 (21.09)	< 0.0001	3 (0.50)	2 (0.20)	< 0.0001	0 (0)	1 (0.07)	0.1320	5 (0.10)	1 (0.02)	< 0.0001	184 (58.60)	353 (46.69)	0.0018
	High School (%)	182 (27.74)	135 (31.99)		40 (6.70)	66 (6.73)		17 (2.45)	25 (1.81)		679 (14.18)	433 (9.64)		90 (28.66)	269 (35.58)	
	> High School (%)	173 (26.37)	138 (32.70)		0 (0)	0 (0)		84 (12.09)	129 (9.35)		3,050 (63.69)	2,635 (58.66)		0 (0)	0 (0)	
	Some College (%)	15 (2.29)	12 (2.84)		215 (36.01)	220 (22.45)		171 (24.60)	373 (27.03)		0 (0)	0 (0)		35 (11.15)	109 (14.42)	
	College Grad (%)	14 (2.13)	10 (2.37)		272 (45.56)	454 (46.33)		277 (39.86)	530 (38.41)		622 (12.99)	759 (16.90)		5 (1.59)	25 (3.31)	
	Postgrad./Degree (%)	34 (5.18)	38 (9.00)		67 (11.22)	238 (24.29)		146 (21.01)	322 (23.33)		433 (9.04)	664 (14.78)		0 (0)	0 (0)	
Ethnicity	White (%)	651 (99.24)	422 (100)	0.1009	357 (59.80)	614 (62.65)	0.2244	655 (92.24)	1,300 (94.20)	0.9968	4,242 (88.58)	4,053 (90.23)	0.0099	0 (0)	0 (0)	.
	Black (%)	2 (0.30)	0 (0)		96 (16.08)	101 (10.31)		20 (2.88)	41 (2.97)		471 (9.84)	377 (8.39)		0 (0)	0 (0)	
	Hispanic (%)	0 (0)	0 (0)		69 (11.56)	202 (20.61)		5 (0.72)	9 (0.65)		57 (1.19)	54 (1.20)		0 (0)	0 (0)	
	Asian (%)	3 (0.46)	0 (0)		70 (11.73)	60 (6.12)		11 (1.58)	22 (1.59)		19 (0.40)	8 (0.09)		314 (100)	756 (100)	
	Amer Indian/Alaskan (%)	0 (0)	0 (0)		5 (0.84)	3 (0.31)		4 (0.58)	8 (0.18)		0 (0)	0 (0)		0 (0)	0 (0)	
	Hawaiian/Pacific Isl. (%)	0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)	
Never Smoker	Number (%)	62 (9.45)	159 (37.68)		106 (17.76)	455 (46.43)		0 (0)	0 (0)		257 (5.36)	1,782 (39.67)		195 (62.1)	661 (87.43)	
Former Smoker	Number (%)	388 (59.15)	135 (31.99)		136 (22.78)	293 (29.90)		139 (20.00)	268 (19.42)		1,619 (33,81)	1,621 (36.09)		39 (12.42)	37 (4.89)	
	CPD (s.e.)	18.26 (0.395)	16.46 (0.691)	0.0221	20.38 (1.063)	13.57 (0.678)	< 0.0001	29.25 (0.867)	26.56 (0.700)	0.0204	27.18 (0.345)	22.16 (0.349)	< 0.0001	11.21 (2.157)	8.54 (1.542)	0.3227
	Pack Years (s.e.)	31.73 (0.872)	22.37 (1.379)	< 0.0001	26.49 (1.767)	12.19 (0.867)	< 0.0001	57.45 (1.958)	47.85 (1.372)	0.0001	46.64 (0.743)	28.89 (0.632)	< 0.0001	20.90 (3.641)	14.08 (3.212)	0.1661
Current Smoker	Number (%)	206 (31.40)	128 (30.33)		355 (59.46)	232 (22.65)		556 (80.00)	1,112 (80.58)		2,913 (60.83)	1,089 (24.24)		80 (25.48)	58 (7.67)	
	CPD (s.e.)	22.49 (0.612)	18.51 (0.718)	< 0.0001	22.84 (0.522)	15.63 (0.676)	< 0.0001	26.03 (0.379)	24.58 (0.263)	0.0016	26.76 (0.215)	21.83 (0.340)	< 0.0001	12.41 (0.937)	10.71 (1.319)	0.2797
	Pack Years (s.e.)	44.27 (1.310)	30.87 (1.710)	< 0.0001	41.95 (1.115)	23.88 (1.268)	< 0.0001	54.99 (0.890)	49.65 (0.601)	< 0.0001	53.72 (0.512)	40.58 (0.793)	< 0.0001	31.05 (2.450)	21.47 (3.025)	0.0144
Any Cancer	0 (%)	250 (38.11)	112 (26.54)	0.0407	279 (46.73)	506 (51.63)	0.0244	269 (38.71)	556 (40.29)	0.1271	4,060 (84.78)	3,800 (84.59)	0.6374	NA	NA	
	1 (%)	238 (36.28)	192 (45.50)		203 (34.00)	325 (33.16)		228 (32.81)	484 (35.07)		469 (9.79)	473 (10.53)		NA	NA	
	2+ (%)	168 (25.61)	118 (27.96)		115 (19.26)	149 (15.20)		198 (28.49)	340 (24.64)		260 (5.43)	219 (4.88)		NA	NA	
Smoking Cancer	0 (%)	334 (50.91)	154 (36.49)	0.0074	365 (61.14)	695 (70.92)	0.0001	415 (59.71)	849 (61.52)	0.2085	4,275 (89.27)	4,020 (89.49)	0.2990	NA	NA	
	1 (%)	206 (31.40)	195 (46.21)		181 (30.32)	230 (23.47)		197 (28.35)	385 (27.90)		373 (7.79)	375 (8.35)		NA	NA	
	2+ (%)	116 (17.68)	73 (17.30)		51 (8.54)	55 (5.61)		83 (11.94)	146 (10.58)		141 (2.94)	97 (2.16)		NA	NA	
Lung Cancer	0 (%)	467 (71.19)	221 (52.37)	< 0.0001	494 (82.75)	889 (90.71)	< 0.0001	593 (85.32)	1,220 (88.41)	0.0793	4,591 (95.87)	4,361 (97.08)	0.0010	NA	NA	
	1 (%)	148 (22.56)	167 (39.57)		95 (15.91)	87 (8.88)		90 (12.95)	138 (10.00)		177 (2.44)	120 (2.67)		NA	NA	
	2+ (%)	41 (6.25)	34 (8.06)		8 (1.34)	4 (0.41)		12 (1.73)	22 (1.59)		21 (0.44)	11 (0.24)		NA	NA	
Prior Tumour	No (%)	592 (90.24)	400 (94.79)	0.0072	517 (68.60)	872 (88.98)	0.1574	601 (86.47)	1,246 (90.29)	0.0111	4,641 (96.91)	4,301 (95.75)	0.0029	NA	NA	
	Yes (%)	64 (9.76)	22 (5.21)		80 (13.40)	108 (11.02)		93 (13.38)	134 (9.71)		148 (3.09)	191 (4.25)		NA	NA	
Dust	No (%)	309 (47.10)	270 (63.98)	< 0.0001	NA	NA		NA	NA		4,690 (97.93)	4,398 (97.91)	0.9318	NA	NA	
	Yes (%)	344 (52.90)	151 (35.78)		NA	NA		NA	NA		99 (2.07)	94 (2.09)		NA	NA	
Asbestos	No (%)	540 (82.32)	397 (94.08)	< 0.0001	NA	NA		581 (83.60)	1,149 (83.26)	0.8461	4,779 (99.79)	4,487 (99.89)	0.2426	NA	NA	
	Yes (%)	116 (17.68)	25 (5.92)		NA	NA		114 (16.40)	231 (16.74)		10 (0.21)	5 (0.11)		NA	NA	
COPD	No (%)	540 (82.32)	370 (87.68)	0.0178	527 (88.27)	926 (94.49)	< 0.0001	578 (83.17)	1,236 (89.57)	< 0.0001	4,302 (89.83)	4,304 (95.81)	< 0.0001	NA	NA	
	Yes (%)	116 (17.68)	52 (12.32)		70 (11.73)	54 (5.51)		117 (16.83)	144 (10.43)		487 (10.17)	188 (4.19)		NA	NA	
Emphysema	No (%)	620 (94.51)	416 (98.58)	0.0007	526 (88.11)	973 (99.29)	< 0.0001	578 (83.17)	1,236 (89.57)	< 0.0001	4,494 (93.84)	4,412 (98.22)	< 0.0001	NA	NA	
	Yes (%)	36 (5.49)	6 (1.42)		71 (11.89)	7 (0.71)		117 (16.83)	144 (10.43)		295 (6.16)	80 (1.78)		NA	NA	
Pneumonia	No (%)	554 (84.45)	391 (92.65)	0.0001	389 (65.16)	799 (81.53)	< 0.0001	500 (71.94)	1,065 (77.17)	0.0090	NA	NA		NA	NA	
	Yes (%)	102 (15.55)	31 (7.35)		208 (34.84)	181 (18.47)		195 (28.06)	315 (22.83)		NA	NA		NA	NA	
Hay Fever	No (%)	NA	NA		530 (88.78)	812 (82.86)	0.0014	564 (81.17)	1,113 (80.65)	0.7854	NA	NA		NA	NA	
	Yes (%)	NA	NA		67 (11.22)	168 (17.14)		131 (45.04)	267 (19.35)		NA	NA		NA	NA	

Table 5.1: Population Demographic of Five Participating Studies [1-5]

Finally, the CARET dataset is a cohort study, this means the cases and controls were not matched and were recruited provided they satisfied the entry criteria. This should allow a fair validation of the prediction models.

5.6 New York Wynder Dataset

The NY Wynder study is a large multicenter case-control study collected across the United States. Cases were recruited as "men and women who were diagnosed with lung cancer were interviewed by trained personnel in nine hospitals in New York, Michigan, Illinois and Pennsylvania" [127]. Controls were matched by location and age where "controls were interviewed in the same hospital as the case and within 2 months of the case interview. Controls matched on age to the cases (within 2-5 years)" [127]. There were no further restrictions in recruiting the participants to the study.

Asbestos exposure and smoking status was categorised based on the questionnaire responses [127];

"Subjects were asked separate questions on whether they had been exposed to asbestos dust on the job for at least 8 hr a week for 1 or more years.

Lifetime exposure to tobacco smoke was assessed according to current smoking status, number of cigarettes smoked per day and number of years since quitting (for ex-smokers). Never smokers included subjects who smoked fewer than 100 cigarettes; current smoking was defined as smoking one or more cigarettes per day. Ex-smokers were defined as having quit smoking for at least 1 year."

This is consistent with how smokers and exposures were defined by the prediction models.

Limited additional information was provided in the study protocol, although Table 5.1 shows that education was collected using 5 levels rather than the 6 levels used in the PLCO models. This could not be reclassified into the 6 levels required by the PLCO models; participants were not classified as "some college".

Overall, the NY Wynder study provides a good population to validate the models. This is supported by the population demographic (Table 5.1); as the NY Wynder study collected a population that would be considered for selective screening.

5.7 Singapore Dataset

The Singapore Chinese Health Study is a large prospective cohort study [128]. The cohort design prevented the cases and controls being matched over key variables to provide a representative sample of participants with which to evaluate the models. The study recruited participants between 1993 and 1998 accumulating 63,257 Asian participants. The trial restricted inclusion by age and location recruiting participants aged between 45-74 years who were Singapore residents. The location will allow an opportunity to evaluate the models outside of a European or North American population. Additionally, the age restrictions in the recruitment will not be a concern, this incorporates the eligible population for previous screening trials and it is most likely any selective screening programme would consider a similar age range.

Unfortunately, only a small proportion of the large cohort was received from the ILCCO repository. The received datasets contained 314 lung cancer cases and 1,070 controls. Lung cancer was identified by records "with the Singapore Cancer Registry and the Registry of Births and Death" [128]. All lung cancer cases identified in the study were provided for our analysis. This included 16 cancers identified at baseline testing and in the most recent dataset after final follow ups by 31st December 2005 a further "298 women cohort participants who were free of cancer at baseline had developed lung cancer" [128]. However, the received dataset restricted the volume of controls that were provided. By restricting the lung cancer free individuals received, this subset of the cohort reports a higher lung cancer prevalence rate than would be observed in a selective screening population.

Study	P.I.	Location	Study Period	Study Design	Eligibilty Restrictions	Cases (n)	Controls (n)
ReSoLuCENT	P. Woll	UK, Various	2006-2015	Case-control; Controls family to a case	Younger than 60 without a family history; Older than 60 permitted with a positive family history	630	419
UCLA	Z. Zhang	Los Angeles, US	1999-2004	Matched case-control on location, age	18-65	597	980
CARET	C. Chen	US, Various	1996-2993	Cohort; Heavy smokers and asbestos exposure	50-69, 20+ pack years; ever smoker	695	1,380
NY Wynder	J. Muscat	New York State, US	1969-1999	Matched case-control; hospital population	None	4,789	4,492
Singapore	A. Seow	Singapore	1993-1998	Cohort; hospital population	Female; 45-74	314	756
New Zealand	B. Cox	New Zealand	2001-2005	Matched case-control; global population	Younger than 55	62	286
CREST	M. Neri	Genoa, Italy	1996-2008	Matched case-control; hospital population	None	382	526
Israel	G. Rennert	Northern Israel	2006	Matched case-control on gender, age, ethnicity; global population	None	301	316
ESTHER	H. Brenner	Saarland, Germany	2000-2003	Matched case-control; hospital population	50-74	185	184
MSH-PMH	G. Liu; R. Hung	Toronto, Canada	2006-2012	Matched case-control on gender, age, location	18+	1,086	1,315

Table 5.2: Summary of Dataset Design

The Singapore dataset collected little information other than smoking history, therefore, only allowed the Hoggart and Pittsburgh models to be applied.

Finally, the participants in the dataset were all female. While a mixed population would be preferred this will not influence the validation results obtained as lung cancer has an equally high prevalence rate in males and females [1] and there is no significant difference in lung cancer susceptibility between the two genders [1]. Additionally, the Hoggart and Pittsburgh models do not incorporate gender into their risk prediction model.

5.8 New Zealand Dataset

The New Zealand study is a small case-control dataset which prioritised collecting lung cancer cases observed in younger participants. The study recruited participants between January 2001 to July 2005 from "hospital databases and the New Zealand Cancer Registry" [130] and by collecting subjects from "eight district boards which represents just under half of the New Zealand population" and allowed a different, nationwide sample of participants to be included.

The study targeted rarer occurrences because "cases were patients with confirmed lung cancer aged 55 years and under" [130] and controls were matched by "5-year age groups and district health boards" [130]. This created a young population demographic (Table 5.3). In additional to participants being young, they reported a low pack year smoking history because of a reduced smoking period. As a result the majority of participants could be considered at low risk of developing lung cancer. Therefore, this population would most likely be extended to include older participants, who are at higher risk of developing lung cancer, in any selective screening programme.

The study participants who smoked cigarettes occasionally, but not daily, were classified as never smokers. There is no concern with this as their CPD and pack year values were nominal and in terms of how the models originally defined ever-smokers they would also be classified as a never-smoker. Table 5.3 highlights that there is no significant difference in CPD rates between individuals with or without lung cancer, as a result models may be limited in distinguishing between individuals with or without lung cancer.

Finally, for a family history of cancer only information for siblings and parents was collected and children excluded. However, since the cases and controls are of a young age their offspring are highly unlikely to be of an age where lung cancer can realistically develop [130].

5.9 CREST Dataset

The CREST study provided a case-control population with participants collected in Genoa, Italy since 1996. The primary focus of the data collection was to create a biorepository at the National Cancer Institute for all cancers.

The participants were collected for the study if they satisfied [131];

- *"Patients with respiratory tract cancer. Patients who have been diagnosed with lung cancer or pleural malignant mesothelioma and that are sampled before any treatment, including surgery. Samples are collected within the framework of ad hoc molecular epidemiology studies or are obtained through a collaborative regional network of pneumologic departments.*

- ***Controls.*** *Patients hospitalized for nonneoplastic and nonrespiratory conditions in orthopedic or ophthalmology departments. Other reference subjects are recruited from blood donors, from social and recreational clubs and homes for the elderly located in the same geographic area where cancer patients are hospitalized. Community outreach programs, including information provided through local TV and newspapers, are occasionally carried out to increase compliance in this group."*

Variable	Dataset	New Zealand			CREST			Israel			ESTHER			MSH-PMH		
		Case	Control	P-value	Case	Control	P-value	Case	Control	P-value	Case	Control	P-value	Case	Control	P-value
	Total (%)	62 (17.82)	286 (82.18)		382 (42.07)	526 (57.93)		301 (48.78)	316 (51.22)		185 (50.14)	184 (49.86)		1,447 (52.73)	1,297 (47.27)	
	Age (s.e.)	48.72 (0.692)	49.32 (0.254)	0.3420	65.87 (0.459)	57.19 (0.728)	< 0.0001	67.75 (0.591)	70.25 (0.586)	0.0033	63.81 (0.490)	63.47 (0.493)	0.6215	65.62 (0.297)	58.79 (0.337)	< 0.0001
Gender	Male (%)	32 (51.61)	132 (46.15)	0.4365	328 (46.26)	381 (53.74)	< 0.0001	192 (48.85)	201 (51.15)	0.9609	140 (50.72)	136 (49.28)	0.6975	672 (51.81)	663 (45.82)	0.0017
	Female (%)	30 (48.39)	154 (53.85)		54 (27.14)	145 (72.86)		109 (48.66)	115 (51.34)		45 (48.38)	48 (51.62)		625 (48.19)	784 (54.18)	
	BMI (s.e.)	NA	NA		NA	NA		26.16 (0.247)	27.58 (0.281)	0.0003	24.96 (0.316)	27.33 (0.300)	< 0.0001	25.28 (0.148)	25.80 (0.130)	0.0076
Education	< High School (%)	0 (0)	0 (0)	0.0046	7 (1.83)	9 (1.71)	< 0.0001	8 (2.66)	9 (2.85)	0.0065	8 (4.32)	4 (2.17)	0.0127	0 (0)	0 (0)	< 0.0001
	High School (%)	1 (1.61)	0 (0)		286 (74.87)	234 (44.49)		33 (10.96)	24 (7.59)		145 (78.38)	127 (69.02)		162 (13.32)	11 (0.78)	
	> High School (%)	40 (64.52)	124 (43.36)		0 (0)	0 (0)		81 (26.91)	46 (14.56)		0 (0)	3 (1.63)		186 (15.30)	51 (3.61)	
	Some College (%)	3 (4.84)	19 (6.64)		72 (18.85)	212 (40.30)		102 (33.89)	136 (43.04)		15 (8.11)	22 (11.96)		251 (20.64)	120 (8.50)	
	College Grad (%)	5 (8.06)	67 (23.43)		0 (0)	0 (0)		0 (0)	0 (0)		11 (5.95)	18 (9.78)		92 (7.57)	52 (3.68)	
	Postgrad./Degree (%)	13 (20.97)	76 (26.57)		17 (4.45)	71 (13.50)		77 (25.58)	101 (31.96)		10 (5.43)	6 (3.26)		525 (43.17)	1,178 (83.43)	
Ethnicity	White (%)	48 (77.42)	253 (88.46)	0.0187	382 (100)	526 (100)	.	300 (99.67)	316 (100)	.	185 (100)	184 (100)	.	1,025 (81.87)	1,239 (89.52)	< 0.0001
	Black (%)	0 (0)	0 (0)		0 (0)	0 (0)		1 (0.33)	0 (0)		0 (0)	0 (0)		26 (2.08)	23 (1.73)	
	Hispanic (%)	0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		12 (0.92)	15 (1.08)	
	Asian (%)	1 (1.61)	5 (1.75)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		180 (14.38)	104 (7.51)	
	Amer Indian/Alaskan (%)	12 (19.35)	25 (8.74)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		9 (0.72)	2 (0.14)	
	Hawaiian/Pacific Isl. (%)	1 (1.61)	3 (1.05)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)		0 (0)	0 (0)	
Never Smoker	Number (%)	9 (14.52)	156 (54.5)		37 (9.69)	231 (43.92)		86 (28.57)	179 (56.65)		19 (10.27)	67 (36.41)		262 (20.20)	829 (57.29)	
Former Smoker	Number (%)	26 (41.94)	79 (27.62)		158 (41.36)	176 (33.46)		98 (32.56)	94 (29.75)		121 (65.41)	84 (45.65)		494 (38.09)	507 (35.04)	
	CPD (s.e.)	17.18 (2.582)	12.32 (1.277)	0.0722	27.89 (1.142)	21.97 (1.255)	0.0006	25.74 (1.556)	23.43 (1.720)	0.3189	20.96 (0.899)	17 (1.096)	0.0056	21.41 (0.525)	15.26 (0.518)	< 0.0001
	Pack Years (s.e.)	22.88 (3.662)	9.48 (1.155)	< 0.0001	53.43 (2.750)	29.95 (2.093)	0.0006	45.65 (3.525)	36.34 (3.345)	0.0574	41.18 (2.222)	21.41 (1.665)	< 0.0001	36.66 (1.200)	16.15 (0.817)	< 0.0001
Current Smoker	Number (%)	27 (43.55)	51 (17.83)		187 (48.95)	119 (22.62)		117 (38.87)	43 (13.61)		45 (24.32)	33 (17.93)		541 (41.71)	111 (7.67)	
	CPD (s.e.)	18.69 (1.272)	15.50 (1.036)	0.0695	28.44 (0.993)	17.68 (0.949)	0.0006	26.57 (1.322)	21.24 (1.983)	0.0338	20.18 (1.400)	16.91 (1.430)	0.1138	20.71 (0.423)	13.20 (1.008)	< 0.0001
	Pack Years (s.e.)	33.12 (2.653)	25.11 (2.108)	0.0265	66.74 (2.500)	32.65 (2.522)	0.0006	60.90 (2.898)	50.18 (5.107)	0.0617	44.53 (3.551)	34.42 (3.264)	0.0467	46.98 (1.126)	24.68 (2.405)	< 0.0001
Any Cancer	0 (%)	24 (38.71)	145 (50.7)	0.0201	221 (57.85)	334 (63.50)	0.0041	151 (50.17)	151 (47.78)	0.9771	NA	NA		NA	NA	
	1 (%)	24 (38.71)	101 (35.31)		105 (27.49)	142 (27.00)		92 (30.56)	104 (32.91)		NA	NA		NA	NA	
	2+ (%)	14 (0.226)	40 (13.99)		56 (14.66)	50 (9.51)		58 (19.27)	61 (19.30)		NA	NA		NA	NA	
Smoking Cancer	0 (%)	33 (53.23)	206 (72.03)	0.0022	267 (69.89)	378 (71.86)	0.0220	193 (64.12)	219 (69.30)	0.1310	NA	NA		NA	NA	
	1 (%)	20 (32.26)	63 (22.03)		83 (21.73)	120 (22.81)		78 (25.91)	73 (23.10)		NA	NA		NA	NA	
	2+ (%)	9 (14.52)	17 (5.94)		32 (8.38)	28 (5.32)		30 (9.97)	24 (7.59)		NA	NA		NA	NA	
Lung Cancer	0 (%)	46 (74.19)	266 (93.01)	< 0.0001	327 (85.60)	476 (90.49)	0.0929	263 (87.38)	292 (92.41)	0.0545	NA	NA		1,230 (94.83)	1,298 (98.70)	0.0001
	1 (%)	15 (24.19)	19 (6.64)		51 (13.35)	47 (8.94)		34 (11.30)	21 (6.65)		NA	NA		58 (4.47)	143 (9.88)	
	2+ (%)	1 (1.61)	1 (0.35)		4 (1.05)	3 (0.57)		4 (1.33)	3 (0.95)		NA	NA		9 (0.69)	6 (0.42)	
Prior Tumour	No (%)	53 (85.48)	250 (87.41)	0.6826	340 (89.01)	526 (100)	< 0.0001	NA	NA		170 (91.89)	171 (92.93)	0.7062	1,297 (100)	1,138 (78.65)	0.0001
	Yes (%)	9 (14.52)	36 (12.59)		42 (10.99)	0 (0)		NA	NA		15 (8.11)	13 (7.07)		0 (0)	309 (21.35)	
Dust	No (%)	43 (69.35)	210 (73.43)	0.5155	374 (97.91)	525 (99.81)	0.0042	NA	NA		NA	NA		NA	NA	
	Yes (%)	19 (30.65)	76 (26.57)		8 (2.09)	1 (0.19)		NA	NA		NA	NA		NA	NA	
Asbestos	No (%)	50 (80.65)	252 (88.11)	0.1162	374 (97.91)	525 (99.81)	0.0042	NA	NA		NA	NA		1,105 (98.93)	1,413 (100)	0.0001
	Yes (%)	12 (19.35)	34 (11.89)		8 (2.09)	1 (0.19)		NA	NA		NA	NA		12 (1.07)	0 (0)	
COPD	No (%)	NA	NA		317 (82.98)	520 (98.86)	< 0.0001	270 (89.70)	295 (93.67)	0.0981	NA	NA		908 (82.55)	1,315 (94.27)	< 0.0001
	Yes (%)	NA	NA		65 (17.02)	6 (1.14)		30 (10.30)	20 (6.33)		NA	NA		191 (17.45)	80 (5.73)	
Emphysema	No (%)	NA	NA		373 (97.64)	523 (99.43)	0.0200	NA	NA		NA	NA		951 (86.45)	1,391 (99.07)	< 0.0001
	Yes (%)	NA	NA		9 (2.36)	3 (0.57)		NA	NA		NA	NA		149 (13.55)	13 (0.93)	
Pneumonia	No (%)	NA	NA		362 (94.76)	512 (97.34)	0.0438	NA	NA		NA	NA		762 (67.55)	1,081 (77.16)	< 0.0001
	Yes (%)	NA	NA		20 (5.24)	14 (2.66)		NA	NA		NA	NA		366 (32.45)	320 (22.84)	
Hay Fever	No (%)	NA	NA		382 (100)	526 (100)	.	282 (93.69)	296 (93.67)	0.9936	NA	NA		NA	NA	
	Yes (%)	NA	NA		0 (0)	0 (0)		19 (6.31)	20 (6.33)		NA	NA		NA	NA	

Table 5.3: Population Demographic of Five Participating Studies [6-10]

All lung cancer cases were eligible for the study with no age or smoking history restrictions enforced during the recruitment. The controls were identified using a hospital population. However, participants hospitalised with lung conditions were excluded, this is unfortunate because lung conditions are considered by models to assign higher risk to individuals. The controls were also matched on age.

When the data collection had concluded by January 31st 2008, 446 lung cancer cases were identified and 935 controls from the ILCCO repository. All cases were provided, although only 542 controls were received. This is due to the ILCCO repository only receiving information for a proportion of the controls. This may create bias as we have a restricted subgroup of the sample population.

5.10 Israel Dataset

The Israel study is a case-control dataset that collected participants who were located in northern Israel. The study was conducted in 2006 and the cases were identified after diagnosis in local hospitals. The controls were also collected from residents in northern Israel and were matched 1:1 between cases and controls based on location, gender, age and ethnicity [132]. These three variables are required by models which could limit the models' discriminative ability. However, the PLCO models, the only models that require ethnicity were not applicable to this dataset as other variable information was missing.

The final dataset provided is quite small with 301 cases and 316 controls. An analysis of this small dataset, presented in Table 5.3, indicates an elderly population has been recruited. It could be argued the participants in the Israel dataset are elderly with the mean age for cases approximately 67 years and controls of 70 years. This is certainly higher than the average population in previously implemented screening trials for lung cancer. As an elevated lung cancer risk is associated with an increased age then it could be argued a marginally high risk population has been collected in the Israel dataset.

Despite the elderly population, the study does not appear to have targeted a specific sample population, whether high risk ever-smokers or a population exposed to a material such as asbestos. Unfortunately, this could not be categorically verified because the study protocol was not provided to enable a review of the exclusion/inclusion criteria.

5.11 ESTHER Dataset

The ESTHER study collected participants from "Saarland in Southwest Germany" [133] and is a combination of two sub-studies; ESTHER 1 and ESTHER 2. Both studies were matched case-control datasets collected at hospitals and general practices between June 2000 and December 2002. The final dataset obtained included 185 cases and 184 controls.

ESTHER 1 collected participants who had a general check-up with the primary study focus on chronic diseases. ESTHER 2 focused on common forms of cancers (colon/rectum, lung, prostate, breast) and was composed of participants diagnosed between January 2001 to December 2003. Age was restricted to 50-74 years in both studies. However, this is a reasonable age range that is appropriate for lung cancer selective screening programmes.

ESTHER 2 identified 197 lung cancer cases during the study, "confirmed using a serum sample" [133] and these were included in the dataset received from the ILCCO repository. They were matched 1-1 with participants from ESTHER 1 who served as controls in the datasets. Unfortunately, the specifics for the matching between cases and controls could not be obtained. There were concerns that the controls were identified from check-ups for chronic conditions and that this would lead to an unexpectedly high volume of controls with lung conditions, particularly COPD. However, these conditions were not reported in the ESTHER dataset received. Indeed, the lack of reporting for variables other than smoking history restricted the dataset to only the Hoggart and Pittsburgh model. While all the participants are classified as white this is not a concern as this variable is not considered by any models applicable to the dataset.

5.12 MSH-PMH Dataset

The MSH-PMH study was a case-control dataset using a clinic based sample population [135]. The primary information for the study was obtained through contact with the Principle Investigator (R.Hung [134]), providing information on the key points of the study and design.

The key information obtained from the P.I. was as follows;

- *The cases are recruited from Princess Margaret Hospital, a designated cancer hospital in Greater Toronto Area.*
- *The controls are recruited from the family medicine registration database in Mount Sinai Hospital, which is next to Princess Margaret Hospital and hosted one of the largest family medicine practices in Greater Toronto Area.*
- *The cases and controls completed standardized study questionnaires and clarifications were followed up by study personnel."*

The MSH-PMH study was collected in Ontario, Canada between 2006 and 2013. Controls were matched to cases on a 1-1 ratio based on age, gender and location [136]. There were no other demographic restrictions in the model recruitment as both males and females aged 18 years and up were eligible for the study [136]. The study only recruited NSCLC cases [136]. However, this does not affect the prediction models as the cancer type is not important when predicting risk.

When obtaining the dataset participants were already classified into 5 education levels whereas the PLCO models require 6 levels. As a result, no participant was classified in the lowest education level (Table 5.3). Table 5.3 also shows that a high proportion of cases and controls with pneumonia and COPD exposure was recruited.

Overall, the large dataset should allow a thorough evaluation of the models' ability as the study did not restrict eligibility for inclusion. This created a sample population to be collected that would be considered in a lung cancer selective screening programme. One concern is the slightly high positive lung conditions observed for both cases and controls; in particular for pneumonia and COPD.

5.13 Dataset Summary

Upon reviewing the study recruitment process, the key strengths and limitations of the datasets, in the context of using the sample populations to validate prediction models, are summarised.

The majority of the datasets obtained were case-control studies. These were the ReSoLuCENT, UCLA, Ny Wynder, New Zealand, CREST, Israel, ESTHER and MSH-PMH datasets. Cohort studies are preferential as the study will have collected a disease free population and observed whether lung cancer developed in individuals. This is the precise purpose of the lung cancer prediction models, to be applied to a disease free population and predict the likelihood of lung cancer developing. Therefore, a cohort would give a comprehensive review of how the models perform. In contrast case-control studies may bias the results. The cases are commonly at an advanced stage of the disease when recruited for the study. This could assist the prediction models, as the cases may report a high risk of developing the disease, which will boost the discriminative ability of the models. As a result, an optimistic model performance may be reported which would not be observed in a selective screening programme. This is a limitation of using case-control studies to validate prediction models and provide accurate summaries of how to apply the model as a selective screening tool. While results for the discrimination, sensitivity and specificity, may be slightly higher than would be observed in a selective screening tool, the case-control datasets will still allow a comprehensive comparison between models and screening criteria to identify a leading model.

Secondly, the obtained datasets have a high proportion of individuals with lung cancer, commonly around 50% of the dataset participants (Tables 5.1 and 5.3). This is a considerably higher prevalence rate

than observed worldwide ([1]). This will affect the calibration results. The models were designed in cohort studies with a prevalence rate similar to that observed globally. Therefore, the models are not designed to predict such a high incidence rate observed in the datasets obtained. This may lead to models reporting a poor calibration result through the Hosmer-Lemeshow test and Brier score. This could make it difficult to identify which models may have any calibration deficiencies as we expect all the models may under perform. Unfortunately, this is unavoidable but will be considered when reviewing the calibration results. It is likely further validations will have to be conducted in cohorts with a realistic prevalence rate, to gain a more accurate review of the models' ability to accurately predict lung cancer incidence.

In recruiting the participants for the study, controls were often matched to cases. Commonly, lung cancer free individuals were matched on age (within 5 years), gender and location or hospital to an individual with lung cancer. Age and gender are two key variables considered by prediction models. The UCLA, NY Wynder, New Zealand, Israel and MSH-PMH datasets were matched on either age, gender, or both. The Israel study also matched on ethnicity, however the PLCO models which utilise this variable was not applicable to the dataset. Matching over these variables can impede the models' performance as these are key variables to discriminate between individuals with or without lung cancer [121]. This will be considered when reviewing the models in these datasets as this may slightly lower the model performance. However, further matching on location and hospital should not influence the models' performance as these are not variables considered by the models. In addition the hospital matched datasets did not specifically collect over diseases that are key factors in models. If hospital controls had been recruited using key lung cancer predictive markers then this would have hindered the discriminative ability of the models as the controls would have an elevated risk predicted. Additionally, none of the recruited studies matched over smoking history, this is an advantage as this is the crucial variable, considered by all the models, to discriminate between individuals with or without or lung cancer.

The studies recruited individuals for different research purposes. This included some studies reviewing lung cancer occurring at young ages (ReSoLuCENT, UCLA and New Zealand), or recruited heavy smoking participants (CARET). These are unorthodox sample populations in comparison to what would be considered in a selective screening trial. The youthful sample populations recruited include a large proportion of participants that are younger than have previously been considered for selective screening. This can be addressed by performing two validations, one assessing the models in the complete dataset and a secondary validation that reviews the models in the subpopulation that would be considered in screening trials. It is important to consider when validating the models that they may be limited in assigning a higher risk to individuals with lung cancer than without lung cancer, as the recruited sample population would be considered low risk of developing lung cancer. Similarly, the CARET population which recruited heavy smoking, asbestos exposed workers may prove challenging for the prediction models. These participants would all be considered at high risk of developing lung cancer, so the models may be unsuccessful in assigning a lower risk to the disease free individuals. The different recruitment strategies will be considered when reviewing the models in these datasets as this may lower their discriminative ability.

The Israel and ESTHER datasets only provided a small sample population to evaluate the models. There may be uncertainty in the results reported in these datasets. However, this uncertainty will be reflected in the 95% confidence intervals when reported for the AUC. While larger studies will allow for a more comprehensive review of the prediction models, the small datasets will still provide a good comparison between the models and an indication into their expected performance as a selective screening tool.

Another identified limitation of the datasets was how the variables were collected. This was highlighted with how education was reported in the UCLA, NY Wynder and MSH-PMH datasets. The datasets received had stratified education into five levels rather than the six levels the PLCO models required. In some instances, this was due to collecting studies in distinct education systems to the USA education system in which the PLCO models were developed. Unfortunately, the predefined education levels in the studies could not be split into the six levels. Therefore one level had zero participants when applying the model in the datasets. This has the potential to limit the ability of the PLCO models as this is an important factor to distinguish between individuals with or without lung cancer. The missing education

level cannot be rectified, but will be considered when reviewing the PLCO models in these datasets.

Next the recruitment strategy for some of the controls may limit the prediction models' performance. The ReSoLuCENT datasets collected disease free individuals who were a first degree relative of a case. Consequently, the dataset has a high positive history of lung cancer amongst controls. This heightens the risk generated by the LLP, Spitz and PLCO models for the disease free individuals and the models may report a poor discriminative ability. Conversely, the CREST dataset excluded some disease free participants with lung conditions. This may also bias the discriminative ability of the prediction models, as the disease free participants with a heightened risk because they have a prior lung condition were excluded. This will allow the models to easier discriminate between individuals with or without lung cancer.

The Bach Model was developed in a high risk study. The received CARET dataset was a subset of the study used to develop the Bach Model. This could see the Bach Model have a heightened performance in this dataset than may be observed in different populations. However, the CARET dataset will still provide a good opportunity to validate the remaining models, especially in a sample population of participants who would be considered of higher risk of developing lung cancer.

While there are concerns with the ILCCO datasets, which may lead to more variable results than observed in selective screening populations, this will still provide a good opportunity to compare between different models and selective screening criteria. Validations across the datasets may identify a leading model which could then be reviewed in selective screening trials to comprehensively assess the models' performance. Additionally, the ILCCO datasets provide an opportunity to evaluate the model in different sample populations. This includes some sample populations that recruited participants who would be considered at low or high risk of developing lung cancer. Also, the studies recruited participants from a large range of countries, which includes Singapore and New Zealand. This provides a good opportunity to assess the models in sample populations outside of Europe and North America, where the systematic review identified models are commonly developed and validated. This can determine whether there is a universally leading model or selective screening criteria, or whether there are leading models for different locations due to the differing populations observed around the world.

5.14 Summary

The chapter discussed the datasets provided for this project and reviewed their distinct participant recruitment strategies. Some studies recruited disease free individuals in an unorthodox manner; whether with a positive history of lung cancer or rejecting individuals with prior lung conditions. Another limitation of the studies was matching. Some studies matched individuals with or without lung cancer over variables required by the models, such as age or gender. This may negate the variable in the models and the models may report a sub-par performance. Finally, the high volume of individuals with lung cancer in the datasets is unrealistic in comparison to the worldwide prevalence rates. These may all influence the models' performance in these datasets and should be considered when reporting the results.

However, the datasets provided do offer a range of distinct sample populations to review the models. This includes some sample populations of participants that would be considered of low or high risk of developing lung cancer. Additionally, the sample populations were recruited in different locations and will allow an opportunity to review the models outside of European and North American subpopulations, where most models have been developed and validated. The differing sample populations will allow a good comparison between the models and selective screening criteria to identify a leading performer and can also review if there is a global leading model or different models with a leading performance in differing locations.

CHAPTER 6

External Validation of Lung Cancer Models: Part One

6.1 Introduction

The systematic review identified that previous validations of lung cancer prediction models have been inconsistent. Based upon the results in the systematic review a leading model with clear guidelines on how it could be utilised as a selective screening programme could not be recommended. Therefore an external validation of the lung cancer prediction models will be presented and reviewed. This will provide an opportunity to review multiple models in the same datasets in a direct comparison. The datasets will also allow the models to be reviewed in distinct sample populations to determine if there is a universal screening strategy or different optimal screening strategies in different populations.

The discrimination, calibration and prediction rules will be evaluated and compared. The prediction rules will be considered at six chosen risk thresholds. Here any participant with a higher risk generated by the model than the specified risk in the threshold would be considered of high enough risk of developing lung cancer to be considered for screening. This will assess if the models can consistently perform strongly at one risk threshold that could then be considered in a selective screening programme. Finally, the models will be compared to previously implemented screening criteria to assess if the models can select a more appropriate high risk population with a high volume of individuals with lung cancer selected and disease free individuals rejected for screening.

6.2 Objectives

To conduct an external validation, which could identify a leading model and how to apply the model to perform optimally, the following objectives were devised:

1. Record the calibration and discrimination for each applicable model in each dataset.
2. Assess the prediction rules at several risk thresholds.
    - Identify the risk threshold at which each model reports the strongest performance.
    - Identify a model that would be optimal as a selective screening tool through comparing the prediction rules.
    - Compare the model to previously implemented screening criteria.

6.3 Methodology

The models were applied to predict risk for the duration specified in the original model; this meant that the Bach, Spitz and Hoggart models that could be run recursively predicted risk for 10, 5 and 1 years respectively. Models were applied to predict absolute risk when they could be applied to predict either risk

of incidence or absolute risk; this was the case for the Bach and Hoggart models. An external validation was then conducted which is divided into sections across the next two chapters. This chapter reviews the models considering the objective of a selective screening programme is to identify a high proportion of individuals with lung cancer for screening. The next chapter reduces the dataset population to that which has previously been considered in selective screening trials. This then reviews how the models perform while aiming to reduce a high proportion of unnecessary costs from screening disease free individuals.

The model calibration, using the Hosmer-Lemeshow test and Brier Score, was evaluated. The AUC was calculated to measure the model discrimination potential and the prediction rules were evaluated through the sensitivity, specificity, Youden index and PLR measures. The prediction rules were measured at the 0.1%, 0.25%, 0.5%, 1%, 1.5% and 2.5% risk thresholds. These thresholds were selected as they would provide a range of different thresholds to evaluate the models. Fixed thresholds were selected rather than evaluating the model at their optimal risk threshold, as this could vary between the datasets and therefore not allow a recommendation as to where a model may perform robustly. The prediction rules will also be compared to the NLST screening criteria to offer a baseline performance of previously implemented screening trials. The objective of the external validation was to identify where the models consistently reported a good performance. If a model is consistently optimal at a risk threshold then this information could be used to inform future validation analysis using cohort studies to ascertain if the model maintains the impressive performance.

The results are presented per dataset and reviewed. The chapter then concludes with a summary analysis of each model's performance.

6.4 ReSoLuCENT Dataset

The ReSoLuCENT dataset was collected to evaluate lung cancer risk factors in low risk populations. The study recruited individuals with lung cancer who were under 60 years old. Additionally 50% of disease free individuals reported a positive family history of lung cancer due to the participant recruitment strategy. Finally, the dataset has a significantly higher prevalence rate than observed for lung cancer worldwide.

The Bach, $\text{PLCO}_{\text{M2012}}$, $\text{PLCO}_{\text{M2014}}$, Hoggart, LLP and Pittsburgh models were applied to the dataset alongside the NLST criteria.

6.4.1 Model Calibration

The calibration was very poor in the ReSoLuCENT dataset for all models except the Bach Model. The Bach Model had a good calibration with a Hosmer-Lemeshow p-value exceeding 0.05 (0.38). It is surprising that the Bach Model, designed for heavy smoking ever-smokers, performed well in a dataset that recruited participants considered as low risk of developing lung cancer. However, this could be a result of the high volume of individuals with lung cancer in the ReSoLuCENT dataset. The other models were calibrated to predict low incidence rates in this sample population. In contrast the Bach Model, which predicted risk over 10 years and was only applicable to the older, heavy smoking participants in the dataset, predicted a high incidence rate which resulted in a good model calibration.

6.4.2 Model Discrimination

Overall the AUC results were poor, particularly for the models designed to target ever-smokers. The Bach Model reported 0.52 (95% CI [0.45, 0.59]) and there were only a slight improvements in the Hoggart Model 0.57 (95% CI [0.52, 0.62]), Pittsburgh Predictor 0.59 (95% CI [0.54, 0.64]) and LLP Model 0.63 (95% CI [0.59, 0.67]). It is surprising the LLP Model had a limited discriminative ability as this model was applicable to never-smokers and was expected to be more generalisible to the lower risk participants recruited in the ReSoLuCENT study. In contrast the $\text{PLCO}_{\text{M2014}}$ Model with an AUC of 0.74 (95% CI [0.71, 0.77]) had a good discrimination and reported the leading performance in the dataset by a significant margin.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**1078**	**47.41**	**78.91**	**0.2632**	**2.2479**
Bach	0.1	10	400	100.00	0.00	0.0000	1.0000
	0.25			100.00	0.00	0.0000	1.0000
	0.5			100.00	0.00	0.0000	1.0000
	1			99.88	0.72	0.0061	1.0061
	1.5			97.30	4.98	0.0228	1.0240
	2.5			85.01	20.65	0.0567	1.0714
PLCO 2014	0.1	6	1075	89.72	40.30	0.3002	1.5028
	0.25			83.17	55.20	0.3837	1.8563
	0.5			72.43	66.84	0.3927	2.1843
	1			53.86	77.08	0.3094	2.3501
	1.5			37.98	84.86	0.2285	2.5094
	2.5			19.83	92.73	0.1256	2.7287
PLCO 2012	0.1	6	854	96.07	15.85	0.1192	1.1417
	0.25			90.89	33.02	0.2392	1.3571
	0.5			79.40	48.32	0.2772	1.5364
	1			58.55	63.62	0.2217	1.6095
	1.5			45.63	76.46	0.2209	1.9387
	2.5			20.96	88.72	0.0967	1.8571
LLP	0.1	5	952	93.34	13.96	0.0730	1.0848
	0.25			82.48	33.44	0.1592	1.2392
	0.5			67.69	52.60	0.2029	1.4279
	1			49.66	68.68	0.1834	1.5858
	1.5			38.90	76.95	0.1585	1.6877
	2.5			25.76	84.74	0.1050	1.6883
Hoggart	0.1	1	798	86.08	17.50	0.0359	1.0435
	0.25			81.47	37.92	0.1939	1.3123
	0.5			54.00	51.99	0.0600	1.1249
	1			33.51	67.54	0.0105	1.0324
	1.5			28.84	71.89	0.0073	1.0258
	2.5			22.11	78.95	0.0105	1.0500
Pittsburgh	0.1	6	823	100.00	0.00	0.0000	1.0000
	0.25			100.00	0.00	0.0000	1.0000
	0.5			85.15	34.65	0.1979	1.3029
	1			59.83	54.27	0.1410	1.3084
	1.5			42.06	70.27	0.1233	1.4149
	2.5			15.32	84.00	-0.0068	0.9574

Table 6.1: ReSoLuCENT Dataset Prediction Rules

6.4.3 Prediction Rules

The NLST criteria reported a reasonable performance. This is despite the NLST criteria, which identifies heavy ever-smokers for screening, being applied in a lower risk sample population. Therefore the NLST criteria rejected a high proportion of disease free individuals from potential screening. However, the criteria would still identify 47% of individuals with lung cancer for screening, which is very impressive. The results

suggest that in younger populations, as observed in the ReSoLuCENT dataset, identifying heavy smokers for selective screening is a good criterion. The good sensitivity and specificity results equate to an impressive Youden index of 0.26 and the high specificity means the model reports a PLR of 2.25.

The Bach Model was limited in the dataset. The model is suboptimal in this sample population where participants would be considered of lower risk of developing lung cancer. This could be a result of considering predictors such as asbestos which was relevant in the model building dataset, whereas in other populations these variables may not be important in explaining lung cancer. The model was poorer in comparison to other models and the NLST criteria. The model was optimal at the 2.5% risk threshold and it could be argued that the model would improve at a higher risk threshold. However, the Youden index observed was very poor (Table 6.1) as the model failed to create a strong relationship between the sensitivity and specificity as observed in the other models.

In contrast, the $\text{PLCO}_{\text{M2014}}$ Model reported a good performance. It was expected that the model could be a leading performer in the dataset as the model was designed to be applicable to anyone aged over 20 years. At the 0.5% risk threshold the model reported a sensitivity of 72% and a specificity of 67% generated a Youden index of 0.39 and a PLR of 2.18; comfortably surpassing the other models and NLST criteria.

The $\text{PLCO}_{\text{M2012}}$ Model was the next leading model with a good performance at the 0.5% threshold, marginally below the NLST criteria. This was followed by the LLP Model which also reported a reasonable performance (Table 6.1) however; the NLST criteria offered a more robust performance. The simpler Hoggart and Pittsburgh models, considering fewer variables, had a poorer performance, as they failed to successfully classify high or low risk participants.

6.4.4 Concluding Statement

The models were poorly calibrated except for the Bach Model. The $\text{PLCO}_{\text{M2014}}$ Model had the strongest overall performance with a significantly improved discrimination in comparison to the other models and this was reflected in the best prediction rule results. The $\text{PLCO}_{\text{M2014}}$ Model offered a clear improvement upon the NLST criteria and reported its most robust performance at the 0.5% risk threshold.

6.5 CARET Dataset

The CARET dataset recruited high risk participants who were ever-smokers with a minimum 20 pack year smoking history and a high proportion of participants were asbestos exposed workers. The CARET dataset is a subset of the sample population from which the Bach Model was devised; this may result in the model having an elevated performance than would be observed in an external validation.

The PLCO, Bach, Hoggart, Pittsburgh and LLP models were evaluated using the dataset.

6.5.1 Model Calibration

There were varied calibration results using the CARET dataset. The LLP, $\text{PLCO}_{\text{M2014}}$ and Pittsburgh models all reported a p-value below 0.05 indicating the models were poorly calibrated. However, the Hoggart and Bach models both reported a good calibration. These models were both designed to predict risk in ever-smokers so were more likely to be calibrated to predict the high volume of incidences in this case-control dataset. Finally, the $\text{PLCO}_{\text{M2012}}$ Model, another model devised to target higher risk ever-smokers, did not accurately predict the observed data, although was close to the critical threshold with a p-value of 0.045. However, the calibration result was quite variable and reducing the volume of participants per group from 10 to 9 increased the p-value to 0.3314.

A comparison of the Brier results places all the models in a very narrow band between 0.3002 and 0.3197. The Bach Model reported the strongest performance. The success of the Bach Model was expected as the dataset is a subset of the population used to devise the model. Conversely, the results may indicate that models that can be applied to never-smokers are not well calibrated to predict incidence solely in

ever-smokers; highlighted by the LLP and $PLCO_{M2014}$ models, although testing should be conducted in cohort studies.

6.5.2 Model Discrimination

The models reported a poor discriminative ability. The AUC results obtained were very low; the Bach Model scored 0.55 (95% CI [0.52, 0.58]), $PLCO_{M2014}$ Model 0.61 (95% CI [0.58, 0.63]), $PLCO_{M2012}$ Model 0.60 (95% CI [0.58, 0.63]), LPP Model 0.58 (95% CI [0.56, 0.61]), Hoggart Model 0.56 (95% CI [0.53, 0.58]) and Pittsburgh Predictor 0.58 (95% CI [0.56, 0.61]).

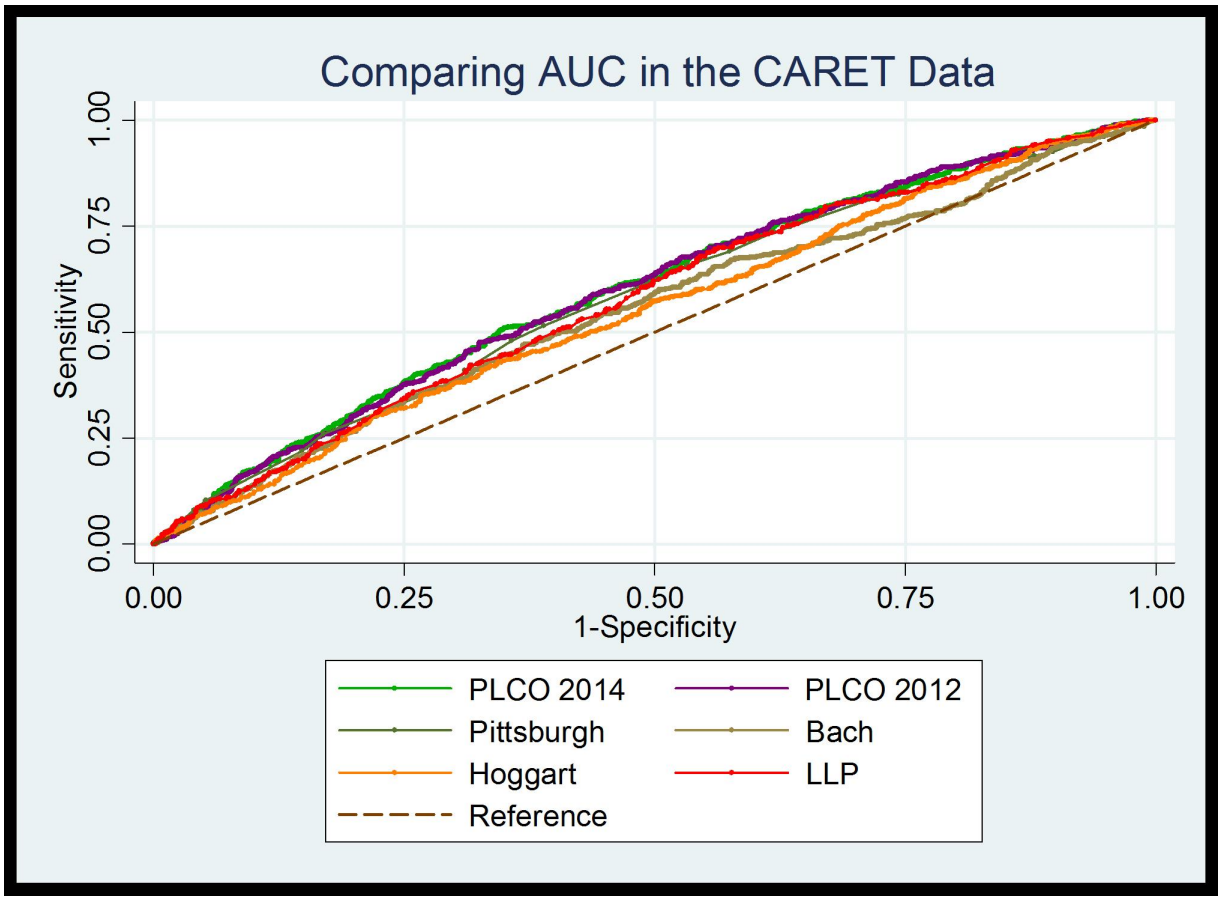

Figure 6.1: AUC of Models in the CARET Dataset

The poor results could be explained by the CARET dataset design. The study participants would be considered at high risk of developing lung cancer. Therefore, the models failed to successfully distinguish between the diseased and disease free individuals, as both were assigned high risk by the models.

6.5.3 Prediction Rules

The NLST screening criteria reported a limited performance in the CARET dataset. This was because of the high risk population collected for the CARET study. All participants were ever-smokers with a heavy smoking history therefore most participants, whether diseased or disease free, satisfied the NLST screening criteria. As a consequence, the NLST criteria reported a low specificity of 29% and a sensitivity exceeding 80%. The criteria did not report a strong relationship between the sensitivity and specificity with a Youden index of 0.1021 and PLR of 1.144. The CARET dataset highlights the high volume of ever-smokers who are lung cancer free and who would be unnecessarily screened using this selective screening criteria.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**2075**	**80.86**	**29.35**	**0.1021**	**1.1445**
Bach	0.1	10	2000	99.12	0.98	0.001	1.0010
	0.25			95.15	7.05	0.022	1.0237
	0.5			88.38	13.64	0.0202	1.0234
	1			81.32	18.79	0.0011	1.0014
	1.5			79.41	20.38	-0.0021	0.9974
	2.5			76.03	25.98	0.0201	1.0272
$PLCO_{2014}$	0.1	6	2075	99.86	0.43	0.0029	1.0029
	0.25			99.86	1.23	0.0109	1.0110
	0.5			99.28	4.2	0.0348	1.0363
	1			83.88	26.52	0.104	1.1415
	1.5			68.92	46.09	0.1501	1.2784
	2.5			43.17	70.87	0.1404	1.4820
$PLCO_{2012}$	0.1	6	2075	100	0.51	0.0051	1.0051
	0.25			99.86	0.94	0.008	1.0081
	0.5			99.71	2.39	0.021	1.0215
	1			91.22	16.45	0.0767	1.0918
	1.5			77.27	36.01	0.1328	1.2075
	2.5			55.97	58.99	0.1496	1.3648
LLP	0.1	5	2074	100	0.07	0.0007	1.0007
	0.25			99.71	1.45	0.0116	1.0118
	0.5			93.23	13.77	0.07	1.0812
	1			78.1	34.49	0.1259	1.1922
	1.5			68.88	43.7	0.1258	1.2234
	2.5			57.06	53.99	0.1105	1.2402
Hoggart	0.1	1	2074	96.12	8.33	0.0445	1.0485
	0.25			90.22	15.14	0.0536	1.0632
	0.5			83.17	24.57	0.0774	1.1026
	1			66.76	39.06	0.0582	1.0955
	1.5			58.42	48.48	0.069	1.1339
	2.5			47.63	59.86	0.0749	1.1866
Pittsburgh	0.1	6	2075	100	0	0	1.0000
	0.25			100	0	0	1.0000
	0.5			97.7	3.48	0.0118	1.0122
	1			92.23	12.03	0.0426	1.0484
	1.5			80.72	30.65	0.1137	1.1640
	2.5			62.45	50.65	0.131	1.2655

Table 6.2: CARET Dataset Prediction Rules

The Bach model reported a suboptimal performance. It could be argued the Bach Model suffered from our validation design; with 2.5% the highest risk threshold considered, which may not be sufficiently high. However, the evidence indicates the model does not improve upon the NLST criteria with a sensitivity below the 80% set by the NLST criteria and a lower specificity. The model fails to distinguish between diseased and disease free individuals in the CARET dataset.

The $PLCO_{M2014}$ Model reported the leading prediction rules. This is slightly surprising considering that

the model is applicable to everyone; it could be expected that a model that targets a higher risk population such as ever-smokers (Hoggart, $PLCO_{M2012}$, Pittsburgh) or heavy smoking ever-smokers (Bach) would be better suited to the CARET dataset. This indicates that the $PLCO_{M2014}$ Model is more versatile and can be applied across a range of different sample populations more successfully than other models. The model was optimal at the 2.5% risk threshold. This may be higher than where the model would be optimal in a sample population that did not recruit such high risk participants. The model reported the highest Youden index alongside the 2012 version of the PLCO model; but reported a higher PLR. While the improvement observed by the $PLCO_{M2014}$ Model in the CARET dataset is marginal, consistently reporting the leading criteria in the datasets will indicate the model could be considered as a selective screening tool in preference to any other model or criteria.

As reported, the $PLCO_{M2012}$ reported a similar Youden index to the 2014 version, although it had a poorer PLR. While, the model cannot be reported to be inferior to the 2014 version, it can be argued that the model should have notably offered an improved performance since the model was devised for ever-smokers. This could also be argued for the Hoggart and Pittsburgh models, both of which had an underwhelming performance in this ever-smoking dataset (Table 6.2). The LLP Model was limited and had a lower performance than the leading PLCO models. Although this model was designed to be applicable to everyone rather than this heavy smoking study.

In conclusion, no model or criteria demonstrated a robust performance in the dataset, as they failed to successfully distinguish between diseased and disease free individuals. There was no major difference between any of the models but the PLCO models had the leading performance.

6.5.4 Concluding Statement

The Hoggart and Bach models demonstrated a good model calibration with the Bach Model reporting the leading performance based upon the Brier Score. The model's discriminative ability was poor and this was reflected in the prediction rules. The $PLCO_{M2014}$ reported marginally the best performance. This, coupled with the leading performance in the ReSoLuCENT dataset, indicates the models ability to be applied in different environments, whether specific populations considered of higher or lower risk of developing lung cancer, if they were chosen for selective screening programmes.

6.6 UCLA Dataset

The UCLA dataset recruited a sample population that would be considered at low risk of developing lung cancer. The mean age is approximately 50, which would be considered quite young as a population for lung cancer selective screening. As a consequence the sensitivity may be lower and the specificity rate higher than observed in a more appropriate population for lung cancer selective screening. Finally, there is a high lung cancer incidence rate observed in this dataset. This is significantly higher than the prevalence rate observed worldwide for which the models were calibrated so they may be poorly calibrated.

The PLCO, Hoggart and Pittsburgh models were applied to the UCLA dataset.

6.6.1 Model Calibration

The models were poorly calibrated in the UCLA dataset. All four models reported a Hosmer-Lemeshow test p-value below 0.05, with the Hoggart model (0.0018) the only model not to report a p-value < 0.0001. This was to be expected in the UCLA dataset with the high lung cancer incidence rate. Additionally, the participants were estimated relatively low risks due to their young age (approximately 50) and reduced smoking history (low pack years).

The Brier score for the Hoggart, Pittsburgh and $PLCO_{M2012}$ models all reported a Brier Score between 0.46 and 0.48. In contrast, the $PLCO_{M2014}$ Model reported a Brier Score of 0.3718; this is a notable improvement than the other models, suggesting the model was better calibrated in the dataset over the remaining models but ultimately a poor calibration was still observed for this model.

6.6.2 Model Discrimination

The models exhibited a very strong discriminative ability. The $PLCO_{M2012}$ Model reported the best performance with an AUC 0.78 (95% CI [0.75, 0.81]); this was closely followed by the $PLCO_{M2014}$ Model with 0.77 (95% CI [0.75, 0.80]). The Hoggart (0.73 (95% CI [0.70, 0.77])) and Pittsburgh Predictor (0.75 (95% CI [0.72, 0.78])) also recorded exceptional results in the UCLA dataset (Figure 6.2). The exceptional discrimination results are slightly unexpected as this was a youthful population recruited for the dataset where the models may have failed to consistently assign a higher risk to individuals with lung cancer.

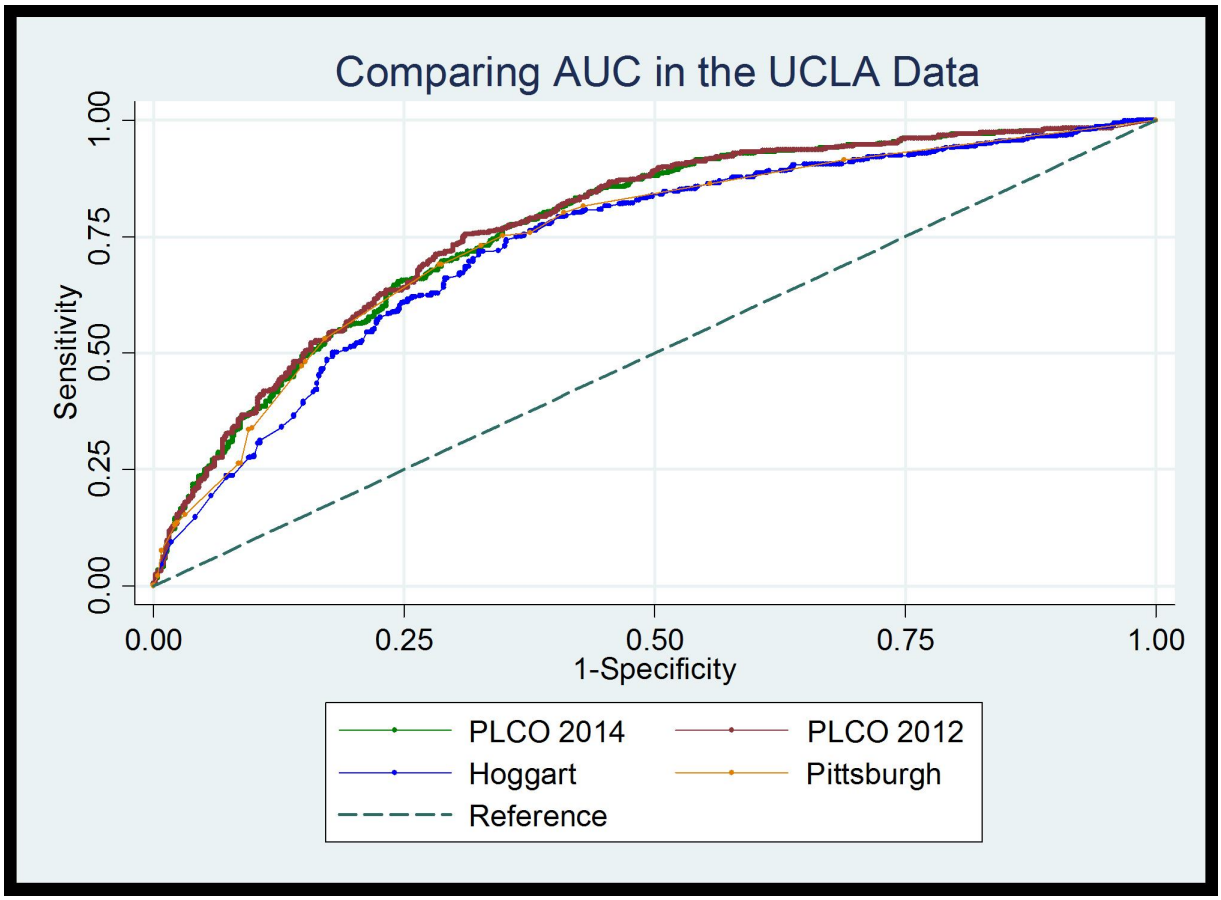

Figure 6.2: AUC of Models in the UCLA Dataset

6.6.3 Prediction Rules

In the UCLA dataset the NLST trial performed to a high standard. The criterion reported a specificity of 91% while still maintaining a sensitivity of 33%. This equates to a Youden index of 0.243 and a high PLR of 3.81 due to only screening approximately 9% of controls. Despite the good performance observed by the NLST criteria this is improved upon by all four models applied to the UCLA dataset.

The $PLCO_{M2014}$ Model demonstrated a clear improvement upon the NLST criteria and reported a strong performance (Table 6.3). The model was optimal at the 0.5% risk threshold where a 21% increase in the sensitivity was observed alongside a 5% improvement in specificity. This equates to an excellent Youden index exceeding 0.4 and an improved PLR of approximately 2.1. This highlights the benefits of the model. Despite the impressive performance this was not the leading model in the UCLA dataset.

The $PLCO_{M2012}$ offered a slight improvement on the 2014 version and reported a remarkable performance in the dataset at the 0.5% risk threshold. The model reported a sensitivity and specificity of both approximately 71.5%. This equated to a very high Youden index of 0.433 and a PLR of 2.536 in ever-smokers.

The Hoggart and Pittsburgh models also reported a high performance in the UCLA dataset and demonstrated a clear improvement upon the NLST criteria. However, these models reported a slightly poorer performance than the PLCO models, suggesting more complex models are better at distinguishing between diseased and disease free individuals especially in an unorthodox, low risk of developing lung cancer, population. The Hoggart and Pittsburgh models reported their optimal performance at the 0.25% and 0.5% risk thresholds, respectively. The models demonstrated a strong relationship between the sensitivity and specificity shown by Youden index scores of approximately 0.39 (Table 6.3). This is marginally poorer than the PLCO models but a significant improvement upon the NLST criteria.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**1577**	**33.00**	**91.33**	0.2433	3.8062
$PLCO_{2014}$	0.1	6	1577	75.54	70.2	0.4574	2.5349
	0.25			66.33	79.08	0.4541	3.1707
	0.5			53.94	86.84	0.4078	4.0988
	1			34.84	93.57	0.2841	5.4184
	1.5			20.44	97.24	0.1768	7.4058
	2.5			6.7	99.29	0.0599	9.4366
$PLCO_{2012}$	0.1	6	1016	92.87	43.24	0.3611	1.6362
	0.25			84.11	58.1	0.4221	2.0074
	0.5			71.49	71.81	0.433	2.5360
	1			50.51	84.76	0.3527	3.3143
	1.5			36.25	91.43	0.2768	4.2299
	2.5			17.11	97.14	0.1425	5.9825
Hoggart	0.1	1	988	92.65	24.21	0.1686	1.2225
	0.25			79.39	59.06	0.3845	1.9392
	0.5			66.94	69.49	0.3643	2.1940
	1			58.37	76.57	0.3494	2.4913
	1.5			51.22	80.31	0.3153	2.6013
	2.5			34.08	87.2	0.2128	2.6625
Pittsburgh	0.1	6	1016	100	0	0	1.0000
	0.25			100	0	0	1.0000
	0.5			75.76	63.43	0.3919	2.0716
	1			48.07	85.33	0.334	3.2768
	1.5			33.06	90.86	0.2392	3.6171
	2.5			13.24	97.9	0.1114	6.3048

Table 6.3: UCLA Dataset Prediction Rules

Overall in the UCLA sample population, all the models reported an exceptional performance with Youden index scores around 0.4. The results supported findings in other datasets that the PLCO models have the leading performance as a selective screening tool. In this case the 2012 version applied to ever-smokers reported the leading performance. All the models demonstrated a significant improvement upon the NLST criteria. The results indicate a model applied in an unorthodox population, which did not recruit participants who are specifically high or low risk of developing lung cancer, such as the UCLA dataset, can successfully distinguish between diseased and disease free individuals.

6.6.4 Concluding Statement

The models were poorly calibrated in the dataset. However, the discrimination and prediction rules were promising. Once again the $PLCO_{M2014}$ was the leading model, although the improvement in comparison to the other models was only marginal. The model was optimal at the 0.5% risk threshold.

6.7 New York Wynder Dataset

The New York Wynder dataset provided a large sample population that would be considered reflective of a target population for lung cancer screening, to validate the models. This should provide a strong indication into which models would be the leading performers in a selective screening programme. However, the case-control dataset design may heighten the discriminative ability of the models and will be considered when reviewing the results.

The Bach, Hoggart, Pittsburgh and PLCO models were applied to the dataset.

6.7.1 Model Calibration

In the NY Wynder dataset the models were poorly calibrated. The two PLCO models, the Hoggart and the Pittsburgh models all reported a Hosmer-Lemeshow p-value of < 0.0001. The Bach Model was also below the 0.05 critical level with a p-value of 0.045. While this was close to the critical value, changing the number of groups across the dataset to calculate the Hosmer-Lemeshow test never observed a p-value greater than 0.05.

The Brier Score results varied between 0.489 and 0.618 for the models. This indicates that the models were some way from predicting the observed incidences in this dataset. This can be attributed to applying the models in a dataset with a high lung cancer incidence rate whereas the models were designed in cohort studies with more realistic, lung cancer incidence rate.

6.7.2 Model Discrimination

There were mixed results in the NY Wynder dataset when evaluating the AUC. The Bach (0.63 95% [0.62, 0.65]), Hoggart (0.67 95% [0.65, 0.68]) and Pittsburgh (0.68 95% [0.67, 0.69]) models only reported reasonable discriminative ability and recorded the poorest performance of the models. This could be a result of the models only considering a small number of variables to distinguish between diseased and disease free individuals.

This was slightly improved on in the $PLCO_{M2012}$ Model (0.69 95% [0.68, 0.70]), although this too was only a reasonable discrimination. Once again the $PLCO_{M2014}$ Model reported the leading performance and there was a notable increase in AUC (0.77 95% [0.76, 0.78]). The results clearly indicate the $PLCO_{M2014}$ Model is the most successful at consistently assigning higher risks to diseased rather than disease free individuals.

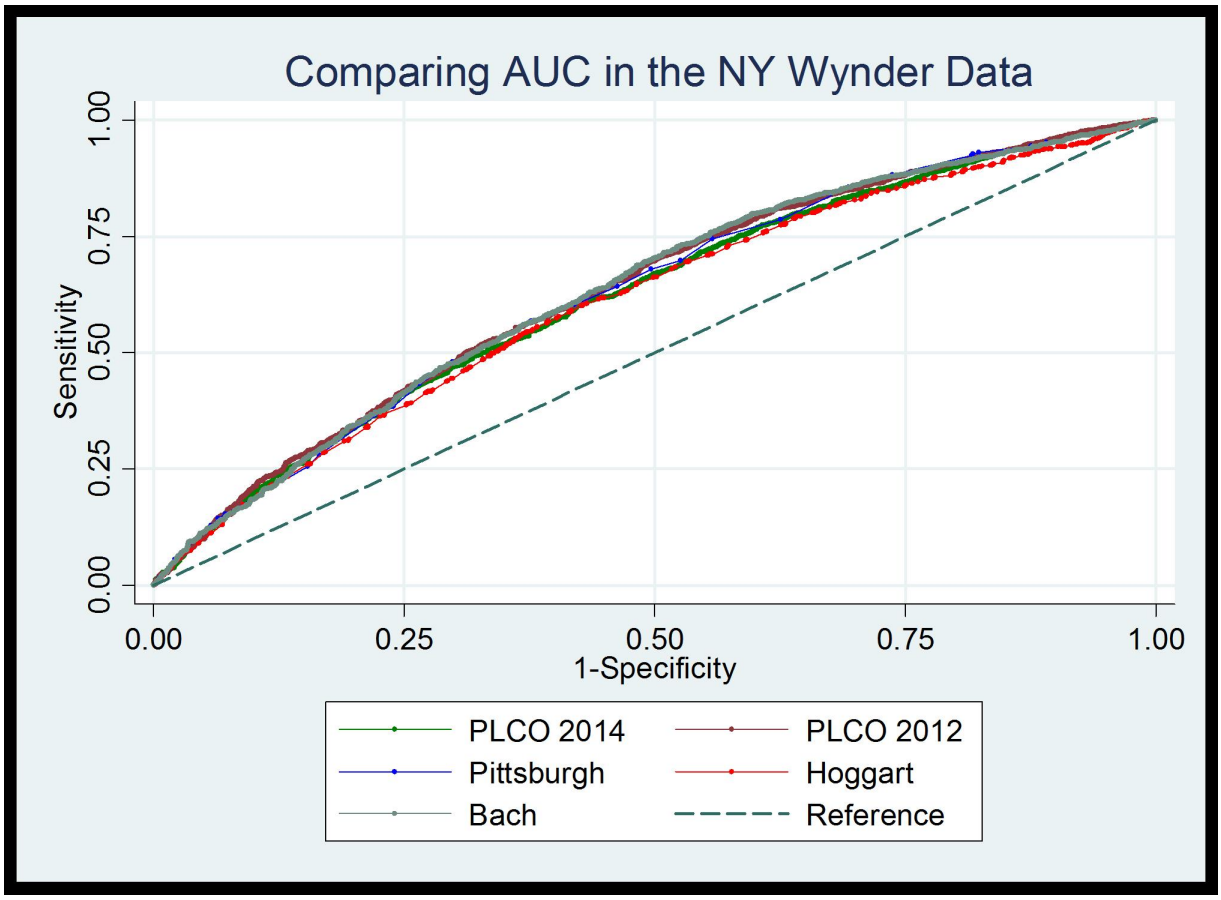

Figure 6.3: AUC of Models in the NY Wynder Dataset

6.7.3 Prediction Rules

The NLST criteria demonstrated a robust performance in the NY Wynder dataset with a sensitivity of 55% and specificity of 77%. This resulted in a Youden's index exceeding 0.3, which demonstrates the criteria could identify a high proportion of cases while still rejecting a sizeable proportion of controls. The high specificity and fairly high sensitivity enabled the criteria to report an impressive PLR of 2.418.

The Bach Model, despite the poor AUC result, demonstrated a reasonable performance. The model was optimal at the 0.5% or 1% risk thresholds as shown by the Youden Index (Table 6.4). The model performs slightly better at the 0.5% risk threshold; here the model reported a sensitivity of 80%. However, it would reject less than 40% of disease free participants, which would contribute to a large number of unnecessary screenings. This may lead to a poor benefit to harm ratio supported by a poor PLR of 1.32. It can be argued the 0.5% threshold is too low for a model that predicts absolute risk over 10 years, as shown by too many participants, both diseased and disease free, considered for screening. However, the model does not improve at higher risk thresholds such as 1.5% or 2.5% (Table 6.4. Additionally, despite a reasonable performance, the Bach Model did not improve upon the NLST criteria.

The $PLCO_{2014}$ Model performed best at the 0.5% risk threshold. Here the model demonstrated a strong performance with a sensitivity of 81% and specificity of 62%. This resulted in a Youden Index of 0.43, which is higher than the NLST criteria. This suggests the model offers a better criterion for classifying diseased and disease free individuals. This was supported by a remarkable PLR of 2.14.

The $PLCO_{2012}$ had a lower level of performance than the 2014 version. The model had a similar standard of performance at the 0.5% or 1% risk thresholds. The model failed to reject a high volume of controls unlike the Bach and $PLCO_{2014}$ models, which was reflected in a poorer specificity and a lower Youden Index. Overall, the model did not improve over the implemented NLST screening trial.

The Hoggart and Pittsburgh models, despite having a reasonable performance, performed poorer than the other models and the NLST criteria. The models were optimal at the 0.5% and 1% risk thresholds, respectively. The Youden index results were lower (Table 6.4) as the specificity was lower coupled with the sensitivity that did not improve upon the other models.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**9281**	**55.44**	**77.07**	**0.3251**	**2.4178**
Bach	0.1	10	4982	98.97	2.14	0.0111	1.0113
	0.25			91.13	19.24	0.1037	1.1284
	0.5			80.32	38.94	0.1926	1.3154
	1			68.72	51.57	0.2029	1.4190
	1.5			64.33	54.58	0.1891	1.4163
	2.5			61.09	57.58	0.1867	1.4401
$PLCO_{2014}$	0.1	6	9271	95.11	28.28	0.2339	1.3261
	0.25			89.23	49.42	0.3865	1.7641
	0.5			81	62.14	0.4314	2.1395
	1			66.07	73.28	0.3935	2.4727
	1.5			54.28	80.66	0.3494	2.8066
	2.5			37.45	88.52	0.2597	3.2622
$PLCO_{2012}$	0.1	6	7232	97.88	12.42	0.103	1.1176
	0.25			94.26	22.47	0.1673	1.2158
	0.5			87.89	35.88	0.2377	1.3707
	1			74.26	52.18	0.2664	1.5571
	1.5			63.32	64.82	0.2814	1.7999
	2.5			45.69	77.46	0.2315	2.0271
Hoggart	0.1	1	7198	94.66	13.84	0.085	1.0987
	0.25			87.19	30.82	0.1801	1.2603
	0.5			76.99	48.68	0.2567	1.5002
	1			59.17	65.24	0.2441	1.7022
	1.5			46.2	75.59	0.2179	1.8927
	2.5			36.27	81.95	0.1822	2.0094
Pittsburgh	0.1	6	7242	100	0	0	1.0000
	0.25			100	0	0	1.0000
	0.5			87.58	31.33	0.1891	1.2754
	1			71.91	56.57	0.2848	1.6558
	1.5			60.81	66.05	0.2686	1.7912
	2.5			44.33	78.23	0.2256	2.0363

Table 6.4: NY Wynder Prediction Rules

In summary the $PLCO_{2014}$ Model and the NLST criteria had the strongest performance. The NLST criteria would reject a high proportion of individuals without lung cancer from unnecessary screening, whereas the $PLCO_{2014}$ Model would capture a high proportion of diseased individuals for screening. While other models performed reasonably well they did not eclipse the $PLCO_{2014}$ Model or the NLST criteria.

6.7.4 Concluding Statement

The models reported a poor calibration in the dataset. The NLST trial performed well and no model exhibited a marked improvement. The $\text{PLCO}_{\text{M2014}}$ Model reported a similar standard to the NLST criteria and was the leading model in the dataset. The $\text{PLCO}_{\text{M2014}}$ Model was once again optimal at the 0.5% risk threshold.

6.8 Singapore Dataset

The Singapore dataset primarily collected information on smoking history. This restricted the dataset to the Hoggart and Pittsburgh models only. The dataset provided an opportunity to evaluate models in an Asian population. This can assess whether the models can replicate their results in distinct environments or whether a geographic specific model for an Asian population should be created to offer a more robust selective screening tool.

The high lung cancer incidence rate observed in the Singapore dataset may affect the model calibration as seen in other datasets. The small dataset size also creates uncertainty in the results, which will be reflected in a large confidence interval for the AUC.

6.8.1 Model Calibration

In the Singapore dataset both models were well calibrated with the Hoggart Model (0.2072) and Pittsburgh Model (0.1189) both exceeding the 0.05 p-value threshold. Despite this both models still reported Brier Score results exceeding 0.5. This indicates that for the majority of participants their estimated risk was low which was not representative of the high incidence rate observed in the dataset.

The slight reduction of lung cancer incidences observed in the dataset (30%), in comparison to approximately 50% in the other datasets, may be the factor that allowed the models to report a good calibration based on the Hosmer-Lemeshow test. However ultimately, the baseline risk predicted by the models is too low for the high incidence rate. This is a consequence of the case-control dataset design, and the high incidence rate is unlikely to be observed in the real world.

6.8.2 Model Discrimination

The two models recorded a reasonable discriminative ability. The Pittsburgh Predictor (0.65 95% CI [0.57, 0.72]) and Hoggart Model (0.63 95% CI [0.56, 0.71]) demonstrated a similar level of performance presented in Figure 6.4. However, if the models are to be considered as a selective screening tool they would be expected to perform to a higher standard than observed. The results could indicate that geographic specific devised models may be better suited as a selective screening tool because the Hoggart and Pittsburgh models reported lower results than that observed in other datasets.

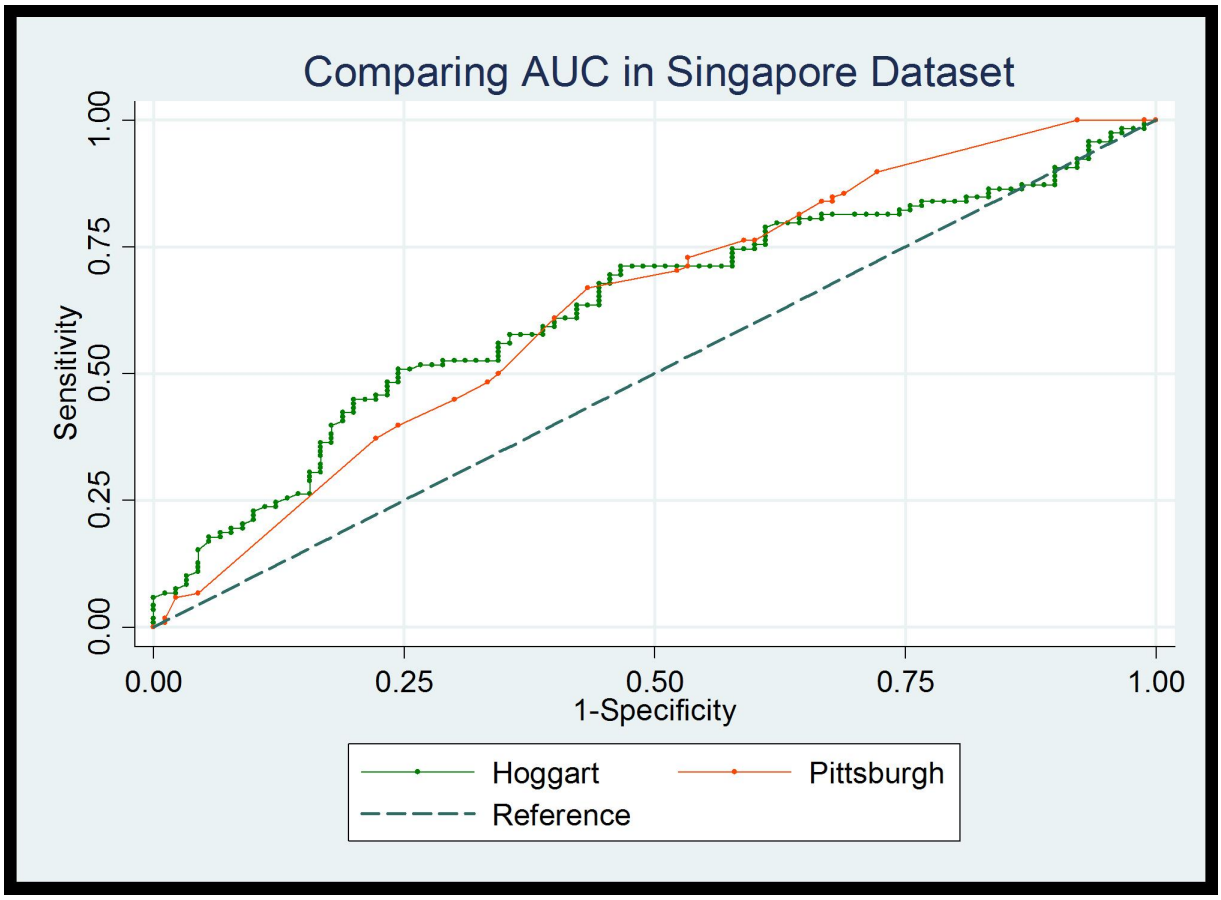

Figure 6.4: AUC of Models in the Singapore Dataset

6.8.3 Prediction Rules

The NLST criteria had a limited performance in the Singapore dataset. This is a consequence of the criteria rejecting a high proportion of participants, both diseased and disease free, from screening. The NLST criteria reported a sensitivity of 8.3% and a specificity of 98.5%. While removing the majority of disease free individuals from unnecessary screening is beneficial, it would be ineffective in identifying lung cancer incidences. The poor trade-off between the sensitivity and specificity is supported by the low Youden's index of about 0.07. The results highlights how incidence occur in never-smokers and more moderate ever-smokers, which would be automatically excluded by the NLST criteria, a clear limitation of this approach to identify a target population for selective screening.

In comparison the Hoggart and Pittsburgh models reported a reasonable performance, demonstrating a clear improvement upon the NLST criteria. Both models demonstrated a better relationship between the sensitivity and specificity shown in Youden index results around 0.25 at their optimal risk thresholds. The Hoggart Model at the 1.5% threshold reported a sensitivity of 49% for an impressive specificity of 75%. The Pittsburgh Model, with a slightly lower performance than the Hoggart Model but improvement over the NLST criteria, was also optimal at the 1.5% risk threshold. Here the model reported a sensitivity of 70% and a specificity of 50%. The Hoggart Model also reported the highest PLR exceeding 2 which was an improvement upon the Pittsburgh Model at 1.397.

Overall, both models improved upon the NLST criteria. The Hoggart model at the 1.5% risk threshold is the strongest performing model, slightly superior than the Pittsburgh Model. The results highlight how a prediction model can identify a larger proportion of participants with lung cancer for screening. The NLST criteria can be limited because it only targets heavy ever-smokers, and even more inclusive ever-smokers models, such as the Hoggart and Pittsburgh models, would improve lung cancer diagnosis rates.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**1070**	**8.28**	**98.54**	**0.0682**	**5.6712**
	0.1			83.9	20	0.039	1.0488
	0.25			79.66	35.56	0.1522	1.2362
	0.5			71.19	48.89	0.2008	1.3929
Hoggart	1	1	208	61.02	60.0	0.2102	1.5255
	1.5			49.15	75.56	0.2471	2.0110
	2.5			35.59	83.33	0.1892	2.1350
	0.1			100	0	0	1.0000
	0.25			100	0	0	1.0000
	0.5			84.87	34.74	0.1961	1.3005
Pittsburgh	1	6	412	75.63	43.16	0.1879	1.3306
	1.5			70.59	49.47	0.2006	1.3970
	2.5			49.58	67.37	0.1695	1.5195

Table 6.5: Singapore Prediction Rules

6.8.4 Concluding Statement

The two models were well calibrated in the Singapore dataset. Despite not reporting a convincing discriminative ability the two models improved upon the NLST criteria, with the Hoggart model having a slightly superior performance of the two models. The results highlight how only capturing high risk ever-smokers using the NLST criteria might not be successful because of the high number of diseased individuals rejected from screening. As a consequence the criteria would not improve lung cancer diagnosis rates.

6.9 New Zealand Dataset

The New Zealand dataset recruited participants who had developed lung cancer before they were 55 years old and would be considered low risk of developing lung cancer. This population is unlikely to be the target population considered for lung cancer screening, based on the target age of previously implemented selective screening trials. The low risk sample population may see the models being limited in correctly classifying high and low risk participants. Additionally, the sensitivity may be lower and the specificity higher than that which would be observed at the risk threshold in a more conventional, higher risk of developing lung cancer, target population for screening.

The Bach, Pittsburgh and Hoggart models were applicable to the dataset. However, only 42 participants were eligible from the dataset using the specified Bach Model criteria (50-75 years, 30+ PY smoking history) so this model was not considered in the dataset because of a lack of eligible participants.

6.9.1 Model Calibration

The models were well calibrated in the New Zealand dataset. The Hoggart and Pittsburgh models recorded a Hosmer-Lemeshow p-value exceeding 0.1. These models reported very similar Brier Scores of 0.283 and 0.284 for the Hoggart and Pittsburgh models, respectively. This supports both models being well calibrated in the New Zealand dataset. Despite being applied to a lower risk of developing lung cancer sample population; the models were well calibrated. This is likely a consequence of the lower observed incidence rate in this dataset, that is more appropriate to lung cancer incidence rates observed worldwide. The evidence suggests lung cancer prediction models, in particular the Hoggart and Pittsburgh models, may be well calibrated in real populations. Indeed, even the case to control ratio of approximately 1:5 is very high but the movement to a more realistic ratio has seen a notable improvement in the Brier Score (in

comparison to scores commonly exceeding 0.4) and the models reporting a good calibration when tested by the Hosmer-Lemeshow test.

6.9.2 Model Discrimination

The small proportion of participants in the dataset applicable to the Hoggart and Pittsburgh models is reflected in wide confidence intervals. The Hoggart Model with 180 participants had a reasonable discrimination of 0.662 (95% CI [0.574, 0.751]). This was similar to the Pittsburgh Model which used 183 participants and reported an AUC of 0.650 (95% CI [0.562, 0.739]). The results suggest that models have a limited discriminative ability in different populations. This is supported by the Singapore dataset where the same models also reported poor AUC results.

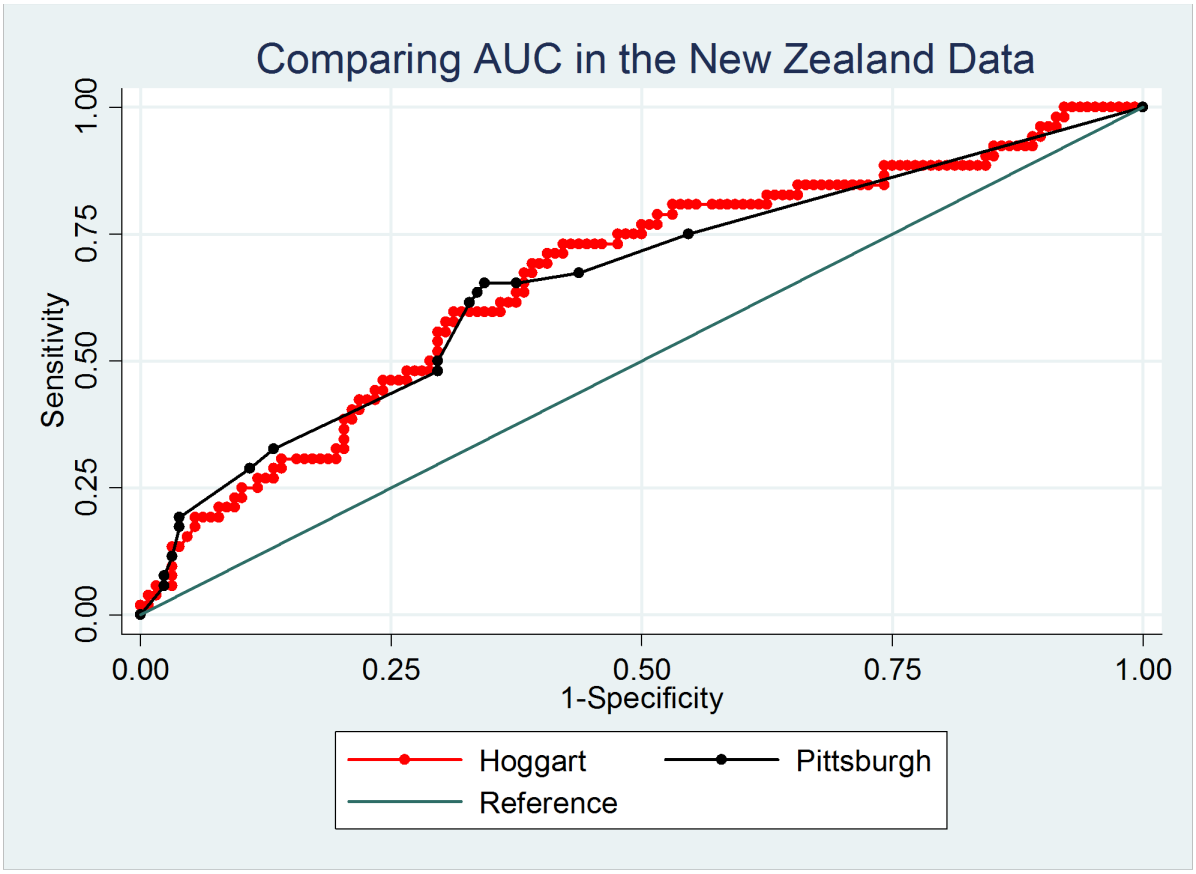

Figure 6.5: AUC of Models in the New Zealand Dataset

6.9.3 Prediction Rules

In the low risk of developing lung cancer sample population, the NLST criteria had a limited performance. This was expected because with an average age of 50 years for participants in the study (Table 5.3) the majority of participants would be rejected from screening. This resulted in a disproportionally low sensitivity (6.5%) and high specificity (99%). The inability for the criteria to identify high risk participants for screening is reflected in a weak Youden index of 0.054. While the criteria reported a high PLR exceeding 6 (Table 6.6) this is misleading because of the exceptionally high specificity. More significantly the NLST criteria only identifying 6% of individuals with lung cancer for screening would not be very beneficial and would not improve lung cancer diagnosis rates.

The Pittsburgh and Hoggart models demonstrated a reasonable performance in the dataset and a clear improvement upon the NLST criteria. The models were optimal at a low risk threshold in this lower risk

sample population, with the Hoggart Model optimal at the 0.25% risk threshold and the Pittsburgh Model optimal at the 0.5% risk threshold. The Hoggart Model reported a sensitivity of 61.5% and a specificity of 62.5% which is a good relationship between both measures, highlighted by a Youden index of 0.240. The Pittsburgh Model improved upon this with a slightly increased sensitivity to 62.3% and an increased specificity to 67% which resulted in a Youden index of 0.292. The Pittsburgh Model demonstrated the potential to identify participants who would benefit from screening in lower risk ever-smokers.

In this lower risk of developing lung cancer sample population, the models were able to surpass the NLST criteria. The results indicate how prediction models can identify diseased individuals in younger populations which are automatically disregarded by the NLST criteria. The prediction models showed a reasonable performance (Table 6.6). However, it is important to realise that the Hoggart and Pittsburgh models can only be applied to ever-smokers and in the ReSoLuCENT dataset, another lower risk of developing lung cancer sample population, the $PLCO_{M2014}$ Model was more successful and included never-smokers.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**348**	**6.45**	**98.95**	**0.054**	**6.1429**
Hoggart	0.1	1	180	84.62	25.78	0.104	1.1401
	0.25			61.54	62.50	0.2404	1.6411
	0.5			48.08	71.88	0.1996	1.7098
	1			42.31	76.56	0.1887	1.8050
	1.5			30.77	82.03	0.128	1.7123
	2.5			19.23	92.97	0.122	2.7354
Pittsburgh	0.1	6	183	100	0	0	1.0000
	0.25			100	0	0	1.0000
	0.5			62.26	66.92	0.2918	1.8821
	1			28.3	89.23	0.1753	2.6277
	1.5			16.98	96.15	0.1313	4.4104
	2.5			0	100	0	NA

Table 6.6: New Zealand Prediction Rules

6.9.4 Concluding Statement

The models were well calibrated in the New Zealand dataset. This is likely to be a reflection of the lower lung cancer incidence rate in the dataset which would be more likely observed in a real world population for which the models were calibrated. In contrast, the models had a relatively poor discriminative ability with AUC results of approximately 0.65. However, the models eclipsed the performance of the NLST criteria, when evaluating the prediction rules, as the heavy smoking NLST criteria was excessive for the majority of the participants in the New Zealand dataset. This indicates how the NLST criteria would not be a practical selective screening tool for lower risk populations as a high proportion of individuals with lung cancer would be rejected from screening.

6.10 CREST Dataset

The CREST dataset provides a large sample population, recruiting participants who would be the target for a lung cancer selective screening programme, with which to evaluate the models. However, the high lung cancer incidence rate observed is likely to affect the model calibration.

The CREST dataset allowed all the models excluding the two PLCO models to be tested. This included the first and only tests of the Spitz and African-American models. The African-American Model

is disadvantaged as this is an Italian dataset with only white participants. The model is being tested in a different environment to that of the model's design and intention, which may see the model under-perform.

6.10.1 Model Calibration

In the CREST dataset there was evidence the models were poorly calibrated. The Hosmer-Lemeshow test for the Spitz and African-American models reported a p-value < 0.0001. While the other models slightly increased the p-value, the Pittsburgh Model (0.0002), Bach Model (0.017) and the Hoggart Model (0.0211) were still below the critical 0.05 value. This was reflected in poor Brier Scores between 0.39 to 0.61. This was expected because of the high lung cancer incidence rate observed in the dataset. The baseline risks generated from the models are much lower than the observed incidence rate because the models were calibrated to reflect real world lung cancer incidence rates.

6.10.2 Model Discrimination

There was a large variability between the models' AUC results in the CREST dataset. The Bach Model had a very poor overall discrimination 0.56 (95% CI [0.48, 0.65]). This confidence interval included 0.5, which indicates the model may, at times, offer no better than chance at assigning a higher risk to diseased individuals.

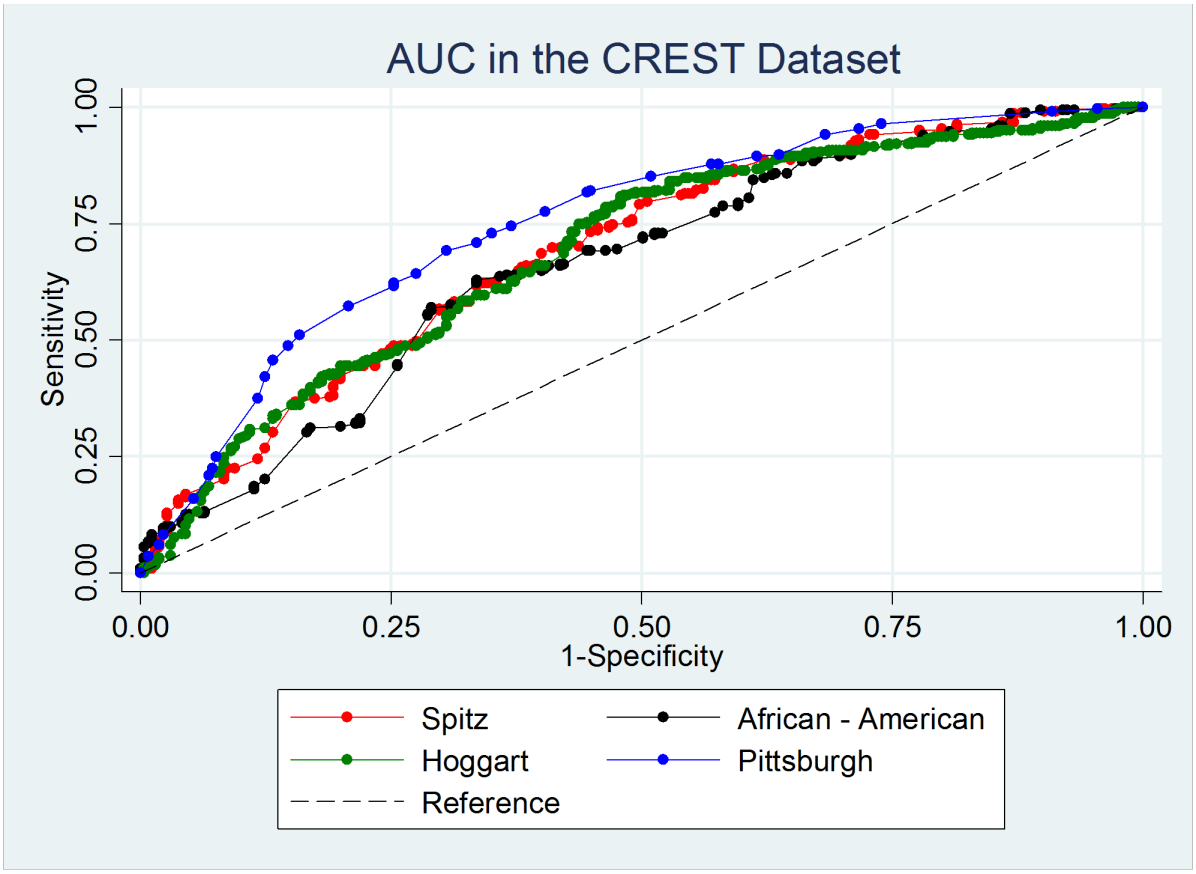

Figure 6.6: AUC of Models in the CREST Dataset

The Bach Model could only be validated in 215 participants who satisfied their criteria, which created uncertainty reflected in the wide confidence interval. Nonetheless, the poor AUC of 0.56 indicates a very poor overall discrimination. Give the small numbers considered the model is not presented in Figure 6.6 because the graph became distorted. The African-American Model had an improved performance (0.67

95% CI [0.64, 0.71]) and performed well considering that this was a different sample population than the target population for the model. However, other models did improve upon this performance indicating there are more appropriate models that can be applied in a predominately Caucasian population and potentially suggesting models are not robust if applied to different ethnic groups or populations other than originally proposed.

The Hoggart Model had a similar, reasonable discriminative ability to the African-American Model, (0.70 95% CI [0.66, 0.74]). However, the Spitz Model (0.76 95% CI [0.73, 0.79]) and the Pittsburgh Predictor (0.77 95% CI [0.74, 0.81]) were the leading models by a considerable margin in the dataset and demonstrated a very good overall discrimination.

6.10.3 Prediction Rules

The NLST criteria was successful in the CREST dataset, with a high sensitivity of 65% and a specificity of 69%. This equates to a Youden index exceeding 0.3, which indicates that there is a good trade-off between the sensitivity and specificity. The impressive sensitivity and specificity if also reflected in a good PLR result of 2.08. Overall, this is a good result in a sample population that would be considered a target population for lung cancer screening.

While the Bach Model performed reasonably at the 0.25% risk threshold the model was poorer in comparison to the NLST criteria, highlighted by a lower Youden index. The model would screen a high proportion of participants with a high sensitivity of 88% but this is mirrored with a low specificity of 32.5%. Considering the model at a higher risk threshold may be beneficial to reduce unnecessary screening of disease free individuals. However, the sensitivity heavily suffers as a consequence and the Youden index was lower at the higher risk thresholds 6.7.

The African-American Model did not improve on the NLST criteria and was one of the poorest models in the dataset. The model was limited when applied in a different sample population from the primary population of the model. The model reported its leading performance at the 0.5% threshold, with a sensitivity of 86% and a specificity of 38%. This reported a Youden index of 0.237, which is notably lower than the NLST criteria and other models. A fairer reflection of the model performance, would be obtained in an African-American sample population to review whether the model exhibits the leading performance.

The Hoggart Model reported a reasonable performance in the dataset. At the 0.5% risk threshold the Youden index was an impressive 0.3. This was a reflection of the high sensitivity (81%) and a still reasonable specificity (50%). The good model performance was reflected in a PLR result of 1.64.

The Spitz Model was optimal at the 0.5% risk threshold. The Youden index at this threshold was very impressive, exceeding 0.4. This was achieved through a sensitivity and specificity both around 70% (Table 6.7). A model that could consistently report this level of performance could be a valuable screening tool as it would identify a high proportion of individuals with lung cancer, although this should be evaluated in cohort studies for a more robust understanding into the model performance. The model offered an improvement over the NLST criteria and future testing should also evaluate if the high standard set by the model at the 0.5% risk threshold can consistently eclipse the NLST criteria.

The Pittsburgh Model reported the joint leading performance alongside the Spitz Model. The Pittsburgh Model was optimal at the 1.5% risk threshold, also reporting a Youden index over 0.4. This was achieved by a sensitivity exceeding 74% supported by a specificity of 67%. The model offered a large improvement over the NLST criteria and demonstrates how simpler models can perform robustly in new environments.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**908**	**64.92**	**68.82**	**0.3374**	**2.0821**
Bach	0.1	10	215	98.52	8.75	0.0727	1.0797
	0.25			88.15	32.50	0.2065	1.3059
	0.5			57.78	47.50	0.0528	1.1006
	1			26.67	80.00	0.0667	1.3335
	1.5			6.67	92.50	-0.0083	0.8893
	2.5			0.74	98.75	-0.0051	0.5920
Spitz	0.1	1	821	88.71	43.18	0.3189	1.5612
	0.25			82.15	57.27	0.3942	1.9225
	0.5			71.92	69.55	0.4147	2.3619
	1			56.43	78.64	0.3507	2.6419
	1.5			37.53	87.95	0.2548	3.1145
	2.5			11.29	98.41	0.097	7.1006
African-American	0.1	5	907	99.21	20.00	0.1921	1.2401
	0.25			94.24	28.76	0.23	1.3229
	0.5			85.60	38.10	0.237	1.3829
	1			71.47	48.76	0.2023	1.3948
	1.5			65.97	58.29	0.2426	1.5816
	2.5			58.90	67.43	0.2633	1.8084
Hoggart	0.1	1	619	95.94	11.68	0.0762	1.0863
	0.25			90.72	33.58	0.243	1.3659
	0.5			81.45	50.36	0.3181	1.6408
	1			64.35	62.41	0.2676	1.7119
	1.5			48.70	73.36	0.2206	1.8281
	2.5			40.87	82.85	0.2372	2.3831
Pittsburgh	0.1	6	640	100.00	0.00	0	1.0000
	0.25			100.00	0.00	0	1.0000
	0.5			94.20	38.31	0.3251	1.5270
	1			81.74	59.66	0.414	2.0263
	1.5			74.20	66.78	0.4098	2.2336
	2.5			62.03	77.29	0.3932	2.7314

Table 6.7: Prediction Rules in the CREST Dataset

6.10.4 Concluding Statement

The models were poorly calibrated in the CREST dataset. Evaluation of the discrimination and the prediction rules saw a large variability between the models' performance. The Bach and African-American models were not successful. The results suggest the African-American Model cannot be applied to different populations and the African-American lung cancer incidence and death rates cannot be applied in a European population. The Spitz and Pittsburgh models reported a strong and leading performance, improving upon the NLST criteria. The Spitz Model, optimal at the 0.5% risk threshold, should be validated at this threshold in additional datasets to confirm whether the very promising results can be replicated.

6.11 Israel Dataset

The study recruited participants that are representative of a population that would be considered in a selective screening trial. Unfortunately, only the Hoggart and Pittsburgh model can be evaluated alongside the NLST criteria as there is only a detailed smoking history and basic information provided in this dataset.

6.11.1 Model Calibration

The Hoggart and Pittsburgh models were both well calibrated in the Israel dataset with a Hosmer-Lemeshow p-value exceeding the critical value. This was surprising as 48.8% of participants in the dataset had lung cancer. However, the models had a poor Brier Score with the Hoggart Model reporting a score of 0.5833 and the Pittsburgh Model reporting a score of 0.5658.

6.11.2 Model Discrimination

The two models recorded a reasonable discrimination; the Hoggart Model marginally reported the highest AUC with a result of 0.66 (95% CI [0.60, 0.72]) in comparison the Pittsburgh Model reported an AUC of 0.63 (95% CI [0.57, 0.69]). The small volume of participants in the dataset is reflected in wide 95% confidence intervals.

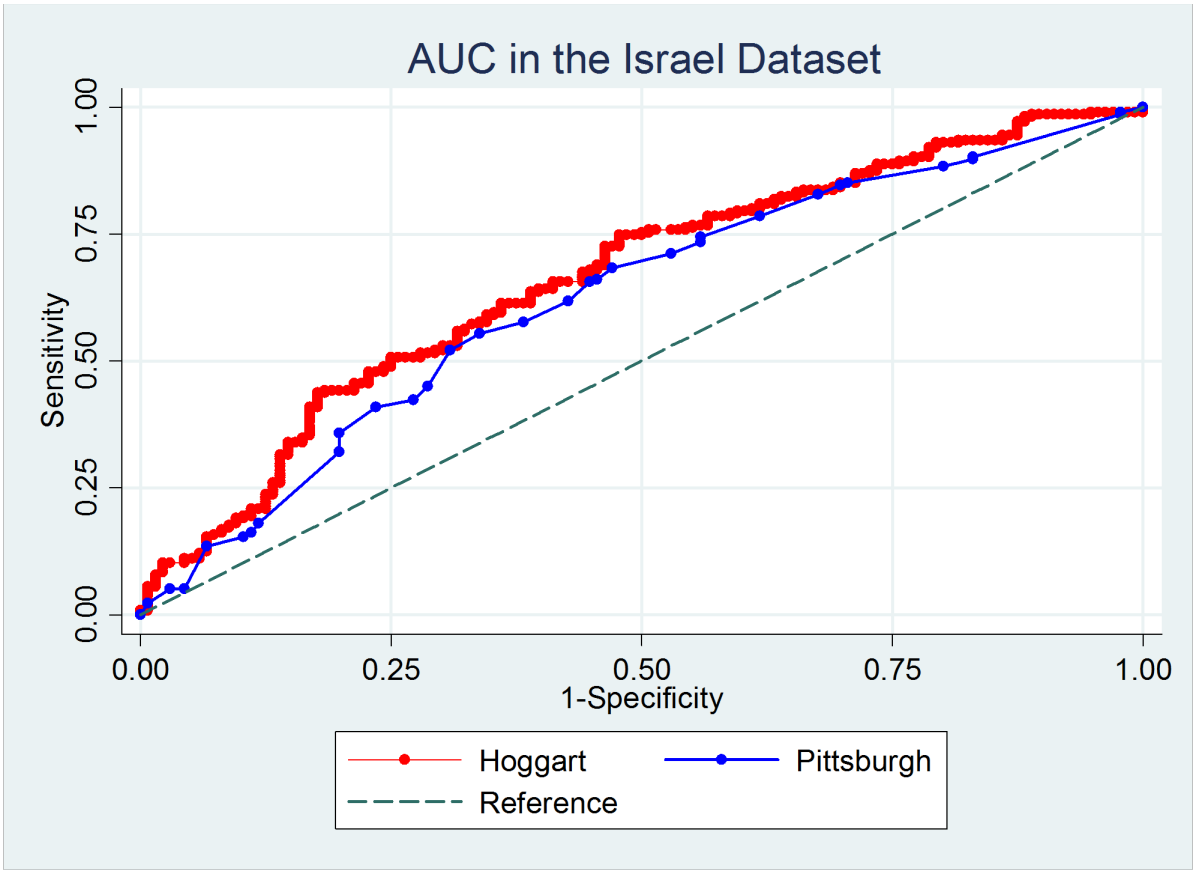

Figure 6.7: AUC of Models in the Israel Dataset

6.11.3 Prediction Rules

The NLST criteria was quite successful in the Israel dataset. The criteria combined a lower sensitivity (39%) with an impressive specificity (84%), this equates to a strong Youden index. The criteria would

capture a reasonable proportion of individuals with lung cancer which would improve lung cancer diagnosis rates, while the impressive specificity reduces the potential harms of unnecessary screening highlighted by a strong PLR result of 2.458.

The Hoggart Model had a slightly poorer performance than the NLST criteria. At the 1% risk threshold the Youden index was marginally lower; despite a sensitivity of 68% the specificity was reduced to 54%. The lower specificity meant the potential harms from unnecessary screening increased, which resulted in a lower PLR of 1.5.

The Pittsburgh Model had a similar overall performance as the Hoggart Model (Table 6.8). The model was optimal at the 2.5% threshold. In comparison to the Hoggart Model the sensitivity was lower at 55% and the specificity increased to 66%. This trade-off was reflected in a similar Youden index between the two models.

In summary, the models and the NLST criteria all demonstrated a reasonable selective screening ability in the Israel dataset. The NLST criteria reported a marginally leading performance and was successful at reducing unnecessary screening of individuals without lung cancer.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**615**	**39.00**	**84.13**	**0.2313**	**2.4575**
Hoggart	0.1	1	351	97.21	12.5	0.0971	1.1110
	0.25			92.09	20.59	0.1268	1.1597
	0.5			80	38.97	0.1897	1.3108
	1			68.37	54.41	0.2278	1.4997
	1.5			52.56	69.85	0.2241	1.7433
	2.5			39.07	83.09	0.2216	2.3105
Pittsburgh	0.1	6	352	100	0	0.0000	1.0000
	0.25			100	0	0.0000	1.0000
	0.5			88.37	20.44	0.0881	1.1107
	1			74.42	44.53	0.1895	1.3416
	1.5			68.37	53.28	0.2175	1.4665
	2.5			55.35	66.42	0.2177	1.6483

Table 6.8: Israel Dataset Prediction Rules

6.11.4 Concluding Statement

The two models were well calibrated in the Israel dataset. In contrast the AUC results were poor with the models reporting results of approximately 0.6. This was reflected in the prediction rules assessment where the models failed to improve upon the NLST criteria, which reported a marginally leading performance in the dataset.

6.12 ESTHER Dataset

The small case-control ESTHER study is only applicable to the Hoggart and Pittsburgh models. However, the study will allow a good indication of how the models would perform as the recruited sample population is representative of a target population for selective screening for lung cancer.

6.12.1 Model Calibration

The Hosmer-Lemeshow test demonstrated the Hoggart (0.0426) and Pittsburgh (0.006) models were not well calibrated in the ESTHER dataset. The Hoggart Model was close to being accepted at the 0.05

threshold and increasing the number of groups across the dataset saw the p-value rising above the 0.05 critical value. Evaluating the Brier score, both models reported a poor result with values of 0.57 and 0.55 for the Hoggart and Pittsburgh models, respectively. Overall, the models had a poor calibration in the ESTHER dataset. This is expected because the models are not calibrated to predict such high incidence rates as observed in the ESTHER study.

6.12.2 Model Discrimination

There were varied results for the models in the ESTHER dataset. The Hoggart Model reported a poor AUC of 0.63 (95% CI [0.56, 0.70]). However, there was a notable improvement observed for the Pittsburgh Model which reported an AUC of 0.72 (95% CI [0.66, 0.78]). This improvement, highlighted in Figure 6.8, suggests that the model was more successful at distinguishing between cases and diseased and disease free individuals.

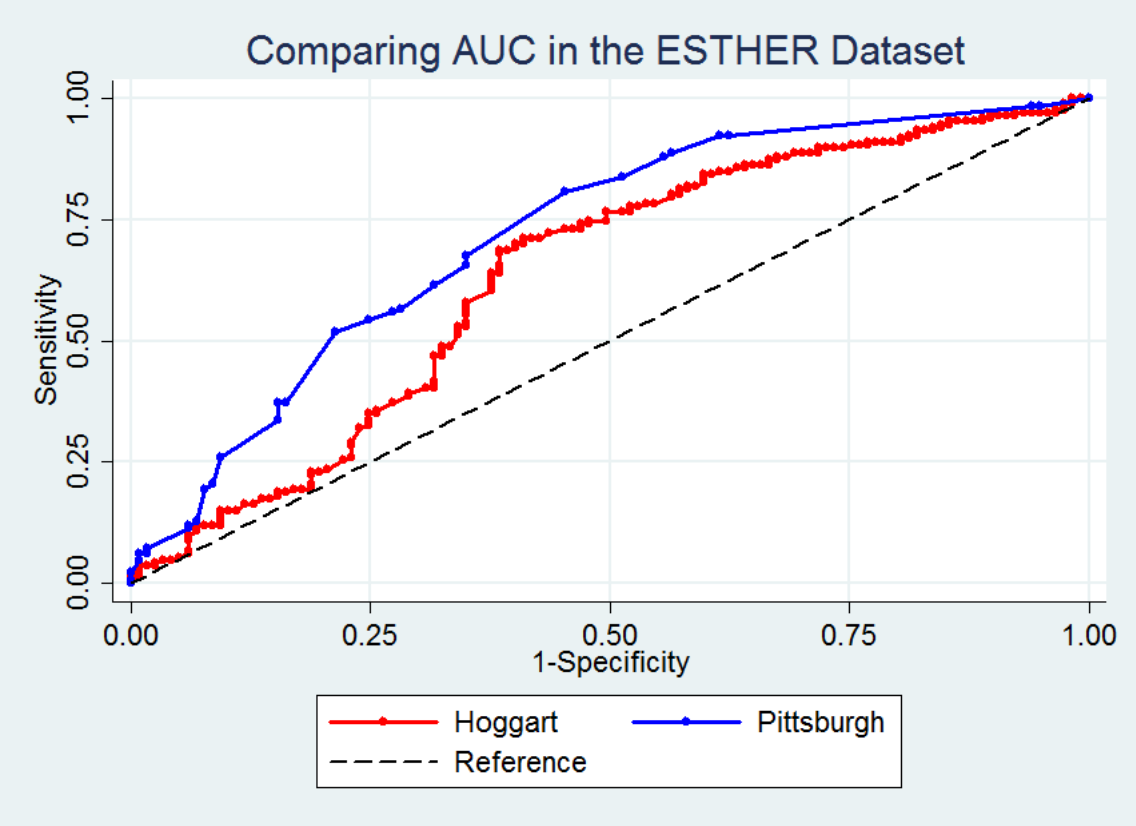

Figure 6.8: AUC of Models in the ESTHER Dataset

6.12.3 Prediction Rules

The NLST criteria reported a good performance in the ESTHER dataset. The criteria reported a sensitivity of 54% and an impressive specificity of 79%. This equates to a strong Youden index exceeding 0.3. Overall the model performs robustly in the dataset and would succeed in rejecting a high proportion of disease free individuals from unnecessary screening.

The Hoggart Model was optimal at the 0.5% risk threshold in the ESTHER dataset where it reported a similar performance to the NLST criteria. The model would capture a much higher proportion of diseased individuals (70%) while maintaining a specificity of 60%. This is a similar standard of performance to the NLST criteria as shown by the Youden index.

Similarly, the Pittsburgh Model does not improve upon the NLST criteria. The model was optimal at the 1% risk threshold with a sensitivity of 67% for a specificity of 65%. This equated to a Youden

index of 0.32, which is the same as the NLST criteria. The lower specificity results in a PLR of 1.93. The Pittsburgh Model did not translate the higher AUC into improved prediction rules in comparison to the Hoggart Model.

Ultimately, while the models have a good performance in the ESTHER dataset, they do not improve upon the NLST criteria. The models, similar to the NLST criteria, consider age and smoking history as the predominate predictors. The results indicate that by incorporating more predictors into the models, this may allow the models to more consistently improvement upon the NLST criteria.

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**369**	**54.05**	**78.80**	**0.3285**	**2.5495**
Hoggart	0.1	1	283	95.18	13.68	0.0886	1.1026
	0.25			88.55	29.91	0.1846	1.2634
	0.5			69.88	59.83	0.2971	1.7396
	1			35.54	74.36	0.0990	1.3861
	1.5			19.28	82.05	0.0133	1.0741
	2.5			16.27	88.03	0.0430	1.3592
Pittsburgh	0.1	6	283	100.00	0.00	0.0000	1.0000
	0.25			100.00	0.00	0.0000	1.0000
	0.5			92.17	38.46	0.3063	1.4977
	1			67.47	64.96	0.3243	1.9255
	1.5			56.63	71.79	0.2842	2.0074
	2.5			37.35	84.62	0.2197	2.4285

Table 6.9: ESTHER Dataset Prediction Rules

6.12.4 Concluding Statement

The Hoggart and Pittsburgh models were poorly calibrated in the dataset. When measuring the AUC the Pittsburgh Model was the stronger of the two models. This was also supported in the prediction rules where the prediction model performed well at the 0.5% risk threshold. However, despite the good performance the models do not improve upon the NLST criteria. This could suggest models considering more predictors would be more successful as a selective screening tool.

6.13 MSH-PMH Dataset

The MSH-PMH dataset provided a large case-control sample population to assess the performance of the models. The participants recruited for the study should allow the models to be thoroughly evaluated as it offers a sample population that is representative of a target population for selective screening.

The dataset was used to validate the Bach, PLCO, Pittsburgh, Hoggart and LLP models.

6.13.1 Model Calibration

The models were poorly calibration and recorded Hosmer-Lemeshow p-values < 0.0001, except the Bach Model, which although improved, still reported a p-value below 0.05. The Brier Scores were also poor because of the high lung cancer incidence rate in the MSH-PMH dataset.

6.13.2 Model Discrimination

The models demonstrated a strong discrimination in the dataset when combining the AUC results using Rubins Rules. The Bach (0.71 95% CI [0.65, 0.77]) and the LLP (0.74 95% CI [0.70, 0.78]) had the lowest

AUC but the results are still very strong. The remaining four models reported a very strong AUC with results for the Hoggart Model (0.79 95% CI [0.77, 0.81]), Pittsburgh Model (0.82 95% CI [0.80, 0.84]), $PLCO_{M2012}$ Model (0.80 95% CI [0.77, 0.82]) and $PLCO_{M2014}$ Model (0.79 95% CI [0.77, 0.81]).

Model	Risk Threshold (%)	Duration	Observations	Sens.	Spec.	Youden	PLR
NLST	**NA**	**NA**	**2744**	**41.48**	**93.85**	**0.3533**	**6.7447**
Bach	0.1	10	517	100.00	0.00	0.0000	1.0000
	0.25			100.00	0.00	0.0000	1.0000
	0.5			99.81	0.00	-0.0019	0.9981
	1			99.03	4.65	0.0368	1.0386
	1.5			98.44	10.47	0.0891	1.0995
	2.5			93.58	20.93	0.1451	1.1835
PLCO 2014	0.1	6	2423	94.22	33.88	0.2810	1.4250
	0.25			89.10	48.78	0.3788	1.7396
	0.5			81.15	63.18	0.4433	2.2040
	1			67.76	78.38	0.4614	3.1341
	1.5			55.04	85.04	0.4008	3.6791
	2.5			39.93	91.98	0.3191	4.9788
PLCO 2012	0.1	6	1299	83.84	49.88	0.3372	1.6728
	0.25			73.81	71.99	0.4580	2.6351
	0.5			65.06	82.29	0.4735	3.6736
	1			54.10	90.39	0.4449	5.6296
	1.5			43.98	93.35	0.3733	6.6135
	2.5			31.94	96.48	0.2842	9.0739
LLP	0.1	5	2096	95.49	15.35	0.1084	1.1281
	0.25			87.82	34.20	0.2202	1.3347
	0.5			75.82	55.47	0.3129	1.7027
	1			62.13	75.24	0.3737	2.5093
	1.5			54.74	82.07	0.3681	3.0530
	2.5			44.51	90.11	0.3462	4.5005
Hoggart	0.1	1	1639	93.44	19.33	0.1277	1.1583
	0.25			85.29	50.77	0.3606	1.7325
	0.5			73.12	74.48	0.4760	2.8652
	1			64.83	82.34	0.4717	3.6710
	1.5			48.35	89.75	0.3810	4.7171
	2.5			33.95	94.70	0.2865	6.4057
Pittsburgh	0.1	6	1639	100.00	0.00	0.0000	1.0000
	0.25			100.00	0.00	0.0000	1.0000
	0.5			90.63	50.69	0.4132	1.8380
	1			74.37	75.85	0.5022	3.0795
	1.5			64.66	83.95	0.4861	4.0287
	2.5			50.63	91.13	0.4176	5.7080

Table 6.10: MSH-PMH Dataset Prediction Rules

6.13.3 Prediction Rules

The NLST criteria reported an exceptional performance in the MSH-PMH dataset. The criteria reported a sensitivity of 41.5% supported by a very high specificity of 94%. This equates to a Youden index score exceeding 0.35 and while slightly higher scores have been observed in other datasets, this index score is exceptional in comparison to the other models validated in this dataset. Indeed, the ability of the model to still identify such a high proportion of lung cancer incidences despite rejecting 94% of disease free individuals, allows the criteria to report a PLR of 6.74, much higher than has previously been observed.

The Bach Model, optimal at the 2.5% risk threshold underachieved in comparison. The model appeared to require a higher risk threshold to yield a better performance because at this threshold a high proportion of both diseased and disease free individuals would still be screened. This resulted in a poor Youden index of 0.145 in comparison to the NLST criteria and the other models.

Next the LLP Model reported a similar performance to the NLST criteria. The model, optimal at the 1% risk threshold, had a sensitivity of 54.7% supported by a high specificity of 82%. While these results created an impressive Youden index exceeding 0.36 this was significantly lower than the remaining models.

The two versions of the PLCO models and the Hoggart and Pittsburgh models all reported an exceptional performance in the MSH-PMH dataset (Table 6.10) with Youden index scores between 0.46-0.50. This built upon the very high overall discriminative ability of the models. The Pittsburgh Model reported the leading performance at the 1% risk threshold. The results were exceptional with sensitivity and specificity scores of approximately 75%. While the level of performance shown by the models is unexpected the results support previous findings that the PLCO models and the Pittsburgh Model are the leading lung cancer prediction models and can improve upon the NLST criteria and the LLP and Bach models which were limited in the dataset.

6.13.4 Concluding Statement

The models recorded a poor calibration in the MSH-PMH dataset. However, all the models had a good overall discrimination which gave rise to strong prediction rules with some exceptional results for the PLCO, Hoggart and Pittsburgh models. These improved upon the NLST criteria, although the Bach and LLP models failed to offer an improvement.

6.14 Model Summaries

The results will now be summarised per model and criteria. This will consider the models' ability to predict the observed incidence rates across the datasets, the overall discrimination and provide an estimate of the optimal risk threshold for the models and a review of expected performance at this threshold. The model limitations will be discussed alongside recommendations on whether a model is a leading lung cancer prediction model and could be considered as a clinical tool.

6.14.1 Model Calibration and Discrimination

The models failed to consistently report a good calibration in the datasets (Figure 6.9). This is a limitation of the study because the poor results could be attributed to the datasets obtained, which had a high lung cancer incidence rate. This is higher than would be observed based on worldwide lung cancer prevalence rates for which the models are calibrated to predict lung cancer incidence.

Despite the overall poor results the Bach Model does perform reasonably well on evaluation of the calibration. The model reported a good calibration in 4 out of the 6 datasets in which the model could be evaluated. This could be explained by the high risk population to which the model was restricted to predict risk and applying the models to predict 10-year absolute risk. This will lead to a heightened predicted incidence rates in comparison to a model that could be applied to everyone to predict 1-year risk. This

Dataset	Model	Calibration (P-Value)	AUC (95% CI)
ReSoluCENT	Bach	0.380	0.52 [0.45, 0.59]
	PLCO(2012)	<0.001	0.67 [0.63, 0.71]
	PLCO(2014)	<0.001	0.74 [0.71, 0.77]
	LLP	<0.001	0.63 [0.59, 0.67]
	Hoggart	0.001	0.57 [0.52, 0.62]
	Pittsburgh	<0.001	0.59 [0.55, 0.64]
CARET	Bach	0.105	0.55 [0.53, 0.58]
	PLCO(2012)	0.045	0.60 [0.58, 0.63]
	PLCO(2014)	0.007	0.61 [0.58, 0.63]
	LLP	0.006	0.58 [0.56, 0.61]
	Hoggart	0.596	0.56 [0.53, 0.58]
	Pittsburgh	0.039	0.58 [0.56, 0.61]
UCLA	PLCO(2012)	<0.001	0.78 [0.75, 0.81]
	PLCO(2014)	<0.001	0.77 [0.75, 0.80]
	Hoggart	0.002	0.74 [0.70, 0.77]
	Pittsburgh	<0.001	0.76 [0.73, 0.78]
NY Wynder	Bach	0.045	0.63 [0.62, 0.65]
	PLCO(2012)	<0.001	0.77 [0.76, 0.78]
	PLCO(2014)	<0.001	0.69 [0.68, 0.70]
	Hoggart	<0.001	0.67 [0.66, 0.68]
	Pittsburgh	<0.001	0.68 [0.67, 0.69]
Singapore	Hoggart	0.207	0.63 [0.56, 0.71]
	Pittsburgh	0.119	0.65 [0.57, 0.72]
New Zealand	Bach	0.664	0.56 [0.38, 0.74]
	Hoggart	0.132	0.66 [0.57, 0.75]
	Pittsburgh	0.142	0.65 [0.58, 0.74]
CREST	Bach	0.017	0.56 [0.48, 0.65]
	Spitz	<0.001	0.76 [0.73, 0.79]
	African-Americar	<0.001	0.67 [0.64, 0.71]
	Hoggart	0.021	0.71 [0.66, 0.73]
	Pittsburgh	<0.001	0.77 [0.75, 0.81]
Israel	Hoggart	0.28	0.66 [0.60, 0.73]
	Pittsburgh	0.196	0.63 [0.57, 0.69]
ESTHER	Hoggart	0.043	0.63 [0.56, 0.70]
	Pittsburgh	0.006	0.72 [0.66, 0.78]
MSH-PMH	Bach	0.38	0.71 [0.65, 0.77]
	PLCO(2012)	<0.001	0.79 [0.77, 0.82]
	PLCO(2014)	<0.001	0.80 [0.77, 0.81]
	LLP	<0.001	0.74 [0.71, 0.75]
	Hoggart	0.001	0.79 [0.77, 0.81]
	Pittsburgh	<0.001	0.82 [0.80, 0.84]

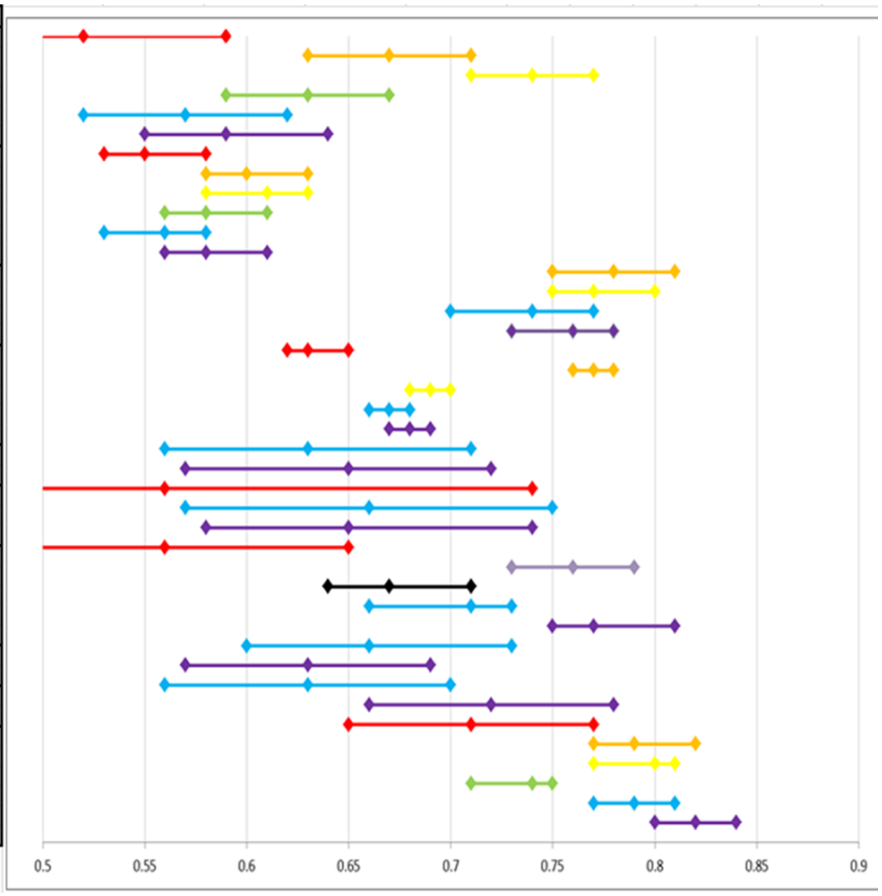

Figure 6.9: Calibration and AUC for the Prediction Models

allowed the Bach Model to predict the incidence rates nearer to the observed prevalence rate. Additionally, the model always reported a good calibration when previously validated (Systematic Review).

The remaining models underperformed with the Hoggart and Pittsburgh models sometimes reporting a good calibration, although this was sporadic across the 10 datasets. The other models consistently reported a Hosmer-Lemeshow p-value below 0.05 as they failed to predict the high observed lung cancer incidence rate. It is recommended the calibration is reviewed in more appropriate studies, that reflect real world lung cancer prevalence rates, to allow a more comprehensive review of the model calibration.

The discrimination results, measured through the AUC, provided more conclusive evidence than the calibration results. The $\text{PLCO}_{\text{M2014}}$ Model reported the highest or very close to highest AUC result of all the prediction models that were applicable in same datasets. The model demonstrated the strongest discriminative ability including in differing sample populations that would be considered lower risk (ReSoLuCENT) or higher risk (CARET) of developing lung cancer.

In contrast, the Bach Model was unable to distinguish between diseased and disease free participants. This may be a limitation of the high risk target population to which the model is applied, so the model fails to consistently assign a higher risk to the individuals with lung cancer than disease free participants. The $\text{PLCO}_{\text{M2012}}$ and Pittsburgh models reported a reasonable discriminative ability. The Pittsburgh results demonstrate how a simple model that only considers four basic factors can still be practical to distinguish between individuals with or without lung cancer. Finally, the Spitz Model reported a good discriminative ability in the only dataset to which the model was applicable, further testing of the Spitz Model should evaluate if the strong AUC of 0.76 (95% CI [0.73, 0.79]) can be replicated.

6.14.2 Prediction Rules

The models reported a mixed performance when evaluating the prediction rules. There were some clear leading models, with the $PLCO_{M2014}$ Model reporting the leading performance at the 0.5% risk threshold. The results are explored in more detail in the summary for each model.

Model	Threshold (%)[1]	Sensitivity (%)	Specificity (%)	PLR[2]
NLST	NA	30-40	80	1.75
Bach	0.5	80	45	1.455
LLP	1	75	50	1.5
Spitz	0.5	72	70	2.4
African-American	0.25	94	29	1.324
PLCOM2012	0.5	80-85	55-60	1.941
PLCOM2014	0.5	70	75	2.8
Hoggart	0.5	70	55	1.556
Pittsburgh	1	65	65	1.857
[1]Risk Threshold where the model performed robustly. [2]PLR is Positive Likelihood Ratio				

Table 6.11: Summary of Prediction Rules for the Models

6.15 NLST Trial Criteria

The NLST criteria has been previously implemented to identify high risk participants for annual LDCT screening. The criteria selected people for screening if they were aged between 55-74 and had a 30+ pack year smoking history.

The high risk criteria was designed to reject a large proportion of disease free individuals from unnecessary screening. This was observed across the datasets with an average specificity of 80%. The high risk criteria resulted in a lower sensitivity of approximately 30-40%. However, the criteria still offers some benefit while the potential harms from unnecessary screening are heavily restricted, which is observed in a PLR of 1.75.

Overall the criteria performed consistently with a reasonable performance. The criteria would be optimal it the objective of a screening programme was to limit unnecessary screening. The criteria reported a performance in sample populations that would be considered lower or higher risk of developing lung cancer as it failed to distinguish between diseased and disease free individuals. Here, prediction models may offer a more robust performance.

6.16 Bach Model

The Bach Model can be applied to predict either incidence or absolute risk over any duration. The model was validated for predicting absolute risk over 10 years because this was how the model was presented in the original article. The model has limitations as it is restricted to a small subset of high risk ever-smokers who are aged 50-75 years and with a minimum 30 pack year smoking history. This restricts the impact of the model to improve lung cancer diagnosis rates as the target population is already restrictive. The Bach Model would be expected to identify the majority of cancers in this subset to positively influence lung cancer diagnosis and survival rates. Indeed, the NLST criteria would screen all participants who are eligible for the Bach Model.

The model had a low discriminative ability in comparison to other models and this was reflected in poorer results when evaluating the prediction rules. The prediction rules for the Bach Model had only been presented in one previous article. In this article, presented in the Systematic Review (Section 3),

the model's prediction rules performance was evaluated at 2.5%, 5% and 7.5% risk thresholds. At these thresholds the results were poor because the model reported a low sensitivity in an already highly restrictive target population. This series of external validations reported more promising results. Except for the high risk CARET dataset the model was strongest in the remaining 4 datasets at 1.5%, 0.5%, 0.25% and 0.5%. Evaluating the prediction rules at the thresholds the model recorded the most consistent performance at the 0.5% risk.

At the 0.5% risk threshold the Bach Model sensitivity was around 80% and the specificity was approximately 45%. This model would only be considered as a selective screening tool if the programme wanted to identify the highest risk cases. If considering selective screening to everyone in a geographical area the Bach Model at the 0.5% risk threshold would limit screening costs in comparison to the NLST criteria of any of the remaining models. However, other models, by considering a larger target population, would have a stronger impact in identifying lung cancer incidences and improving lung cancer diagnosis and survival rates in comparison.

6.17 LLP Model

The LLP Model was applied to predict 5-year risk of lung cancer incidence in four ILCCO datasets. The model can be applied to never-smokers but is restricted to participants aged 40-80 years because the age and gender specific incidence rates are only provided for this age range. This is a reasonable age range as would be expected for a lung cancer screening programme.

The model performed robustly at the 1% risk threshold, which was supported by the optimal performance of the model at the 0.91% threshold in a previous external validation [74]. At this threshold the sensitivity was around 75%, although the specificity was lower at 50%. While the model reported a reasonable performance when evaluating the prediction rules this was not one of the leading models. The $\text{PLCO}_{\text{M2014}}$ Model reported a much improved performance and was also applicable to a large population by having more flexible age restrictions (20+). Additionally, the Pittsburgh and $\text{PLCO}_{\text{M2012}}$ models reported an improved performance, although these were restricted to ever-smokers. In summary, the results indicate other models would offer a more robust selective screening tool in comparison to the LLP Model.

6.18 Spitz Model

The Spitz Model was applied to predict 1 year absolute risk of lung cancer. Unfortunately, the model could only be applied to the CREST dataset so any promising results would have to be replicated in other populations. Without replicating the results in new external validations there can be no confidence the Spitz Model would be consistent and successful as a selective screening tool.

The impressive discriminative ability was reflected in the prediction rules; where the model was optimal at the 0.5% risk threshold. In previous external validations the model had been reviewed at the 2.5% risk threshold because the model was run recursively over 5 years rather than 1 year in this study. The model at the 0.5% threshold reported a sensitivity of 72% and specificity of 70%. The high sensitivity would improve lung cancer diagnosis rates and the equally high specificity would limit unnecessary screening. A model which could consistently replicate these results at the same risk threshold could be a successful selective screening tool. Therefore, future testing should evaluate the Spitz Model at the 0.5% risk threshold.

6.19 African-American Model

The African-American Model, applied to predict 5-year absolute risk, was reviewed in the CREST dataset. While the model can be applied to anyone aged 20+ years, the model was designed to predict risk in African-Americans. Unfortunately, the model could not be applied to an African-American sample population, instead a European population in this external validation.

The model was optimal at the 0.25% risk threshold, with a sensitivity of 94%, although the specificity is low at 29%. This was poorer than other models in the same dataset and the NLST criteria. This suggests that the model cannot be robustly applied in different populations from the original model purpose.

No conclusive results can be determined from the external validation. The African-American Model appears to fail in different populations. Future work should evaluate this model and the other lung cancer prediction models in an African-American population. This will evaluate if this is then the leading model or if there is an universal model that can be applied in differing populations.

6.20 $\text{PLCO}_{\text{M2014}}$ Model

The $\text{PLCO}_{\text{M2014}}$ Model was applied to five ILCCO datasets to predict 6-year risk of lung cancer incidence. The model could be applied to never-smokers as well as ever-smokers provided they were aged at least 20 years. Since people under 20 years would not be considered for lung cancer screening, there are no concerns with the model restrictions.

The impressive discriminative ability was reflected in the prediction rules. The model was optimal at the 0.5% risk threshold where the model reported a sensitivity of approximately 70% and a specificity of 75%. This is an exceptional performance with a Youden index of 0.45 and a PLR of 2.8. There are many advantages to this prediction model because the high sensitivity would improve lung cancer diagnosis rates if applied as a selective screening tool. Additionally, the strong specificity would limit potential harm by reducing unnecessary screening.

The results demonstrate that if a selective screening programme aims to identify a high proportion of individuals with lung cancer, then the $\text{PLCO}_{\text{M2014}}$ Model should be considered at the 0.5% risk threshold. This offered a significant improvement over the NLST criteria and would double the proportion of diseased individuals identified, however an extra 5% of disease free individuals would be screened. If a selective screening programme preferred to identify cases amongst ever-smokers then the $\text{PLCO}_{\text{M2012}}$ or Pittsburgh models may be preferred, although the impressive discriminative ability of the model, including in ever-smokers studies, suggests that the model would still be successful in these specific target population.

6.21 $\text{PLCO}_{\text{M2012}}$ Model

The $\text{PLCO}_{\text{M2012}}$ Model predicted 6-year risk of lung cancer in ever-smokers aged 20+ years. The model was applied to 5 ILCCO datasets.

The $\text{PLCO}_{\text{M2012}}$ Model had only been evaluated at one risk threshold in previous validation studies. That study found 1.51% was the optimal risk threshold for the model and it would improve on the NLST criteria. The optimal 1.51% risk threshold was not supported in this study. The 1.51% threshold was identified in a heavy 30+ pack year ever-smoker cohort where participants would be considered at high risk of developing lung cancer. This external validation found the model performed robustly at the 0.5% risk threshold across the datasets.

At this low optimal risk threshold the model would screen a high proportion of ever-smokers. The sensitivity was between 80-85% with results of 83%, 87%, 71% and 81%, which are consistently high. The specificity was more variable across the datasets and was approximately 55-60%. This may be deemed too high a proportion of disease free participants unnecessarily screened in a clinical setting, although if this is acceptable, then the high sensitivity would improve lung cancer diagnosis rates.

This model should be preferred in comparison to the NLST criteria if the selective screening programme aimed to capture a high proportion of lung cancer incidences in ever-smokers. This model marginally had the better performance over the Pittsburgh Model; the other leading ever-smokers model. The results across the datasets demonstrated that the model was robust and successful in many different populations.

6.22 Hoggart Model

The Hoggart Model was applied to all 10 ILCCO datasets to predict 1-year absolute risk in ever-smokers aged 35+ years. Across the 10 datasets the Hoggart Model reported mixed results.

The prediction rules had not been previously validated. The results across 9 datasets, after excluding the higher risk CARET dataset, provided clear evidence that 0.5% was the optimal risk threshold. The Hoggart Model generated reasonable results at this risk threshold. Across the datasets the sensitivity was around 70% with a slightly low specificity around 55%. This resulted in a reasonable Youden index of approximately 1.25 and a PLR of 1.56.

Across the datasets the Hoggart Model reported a similar performance to the NLST criteria. However, this was not the leading model as the Pittsburgh and $\text{PLCO}_{\text{M2012}}$ models demonstrated a stronger performance. Additionally, the validation results indicate the $\text{PLCO}_{\text{M2014}}$ Model would be a more successful selective screening tool and could also be applied to never-smokers.

6.23 Pittsburgh Model

The Pittsburgh Model was applied in all the ILCCO datasets to predict 6-year risk of incidence in ever-smokers.

The Pittsburgh Model reported the strongest prediction rules performance at the 1% risk threshold. Here the model reported a sensitivity and specificity of approximately 65% which allowed for a Youden index of 0.30 and a PLR of approximately 1.85. This was a very good performance amongst ever-smokers and one of the leading models alongside the $\text{PLCO}_{\text{M2012}}$ Model. This may be preferred to the NLST criteria because the model would capture a higher proportion of lung cancer incidences, improving lung diagnosis rates. However, the model would have to screen additional disease free individuals.

In summary, if a selective screening programme was implemented to identify lung cancer cases in ever-smokers then the Pittsburgh Model could be considered as this was the leading model alongside the $\text{PLCO}_{\text{M2012}}$ Model. Applying the model at the 1% risk threshold would identify approximately 65% of lung cancer individuals although this would require screening approximately 35% of all disease free ever-smokers.

6.24 Summary

The external validations show that the previously implemented NLST criteria can be improved by considering different prediction models to select a target population for screening. The PLCO models and the Pittsburgh Model had the best performance. These models reported the highest AUC results and the strongest prediction rules results. The models were poorly calibrated in these datasets due to the high lung cancer incidence rates, and they should be evaluated in cohort studies with more appropriate incidence rates to gain a better understanding of the model calibration.

The $\text{PLCO}_{\text{M2014}}$ Model was the strongest performing model with an AUC result between 0.68-0.79 in most datasets. The model had an optimal performance at the 0.5% risk threshold with a sensitivity of 70% and a specificity of 75%. In comparison to the implemented NLST criteria, the $\text{PLCO}_{\text{M2014}}$ Model would capture 30-40% more individuals with lung cancer, although it would require screening 5% more disease free individuals. This is a considerable amount of additional screening when considering the lung cancer prevalence rate. Health decision makers may prefer to utilise this model with the improved lung cancer capture rate at the determinate of extra unnecessary screening of disease free individuals.

The $\text{PLCO}_{\text{M2012}}$ and Pittsburgh models, which both target ever-smokers, reported a similar performance to the NLST criteria. These models performed robustly at the 0.5% and 1% risk thresholds respectively for 6-year risk. While they reported a high lung cancer capture rate they would require screening a greater volume of disease free individuals (10-15%). Therefore, the costs of applying this criteria in a selective screening trial would be increased in comparison to the NLST trial criteria or the PLCO_{2014} Model.

The Spitz Model displayed a promising performance at the 0.5% risk threshold. However, this was only evaluated in one dataset so further testing of the model is required to assess if the promising results can be replicated. The remaining models had a poorer performance so should not be considered as a selective screening tool.

There can be some difficulty interpreting the results as to what constitutes a good compromise between the sensitivity and specificity. The next validation, presented in the subsequent chapter, tries to address this by fixing the specificity to a rate that seems appropriate based on prior lung cancer screening. Then the sensitivity is measured and to review which models are more successful.

CHAPTER 7

External Validation of Models in comparison to UKLS Guidelines: Part 2

7.1 Introduction

The previous validation identified the $PLCO_{M2014}$ as a leading model. This model at the 0.5% risk threshold reported a sensitivity of 70% and a specificity of 75%. This would capture a high proportion of individuals with lung cancer however, screening 25% of disease free individuals could be unrealistic and may not be economically viable. Therefore, a second validation was conducted to evaluate the models while restricting the specificity in the target population. The models were compared to each other and the UKLS screening guidelines, which aims to restrict unnecessary screening by only identifying extremely high risk participants. This proposed validation will identify the models' risk thresholds to maintain a high specificity yet still be a successful selective screening tool but identifying individuals with lung cancer for screening.

7.2 Objectives

To conduct the external validation and identify economical screening criteria the chapter has the following objectives;

1. Identify an appropriate specificity level based on the UKLS guidelines.
2. Assess the performance of the prediction rules of the models and the UKLS criteria.
    - Identify the risk threshold that allows the model to maintain the desired specificity level.
    - Assess the sensitivity, Youden Index and PLR, at this risk threshold.
3. Identify the leading model and subsequent risk threshold that would be the optimal selective screening criteria.

7.3 Methodology

A second validation was conducted to evaluate the models in a target population and attempt to consider the economic costs and feasibility of screening a large volume of disease free individuals. This can be observed in previous lung cancer programmes, such as the NLST and UKLS criteria, which target high risk individuals. The objective of this validation is to identify if a prediction model can be utilised to improve these selective screening programmes while heavily restricting unnecessary screening.

The first stage of the validation was to determine an acceptable specificity. There is no minimum acceptable standard for a screening programme in terms of budgeting, sensitivity and specificity. Therefore, the minimum standard was inferred from what has been acceptable in previous screening programmes; this

was obtained from the UKLS trial criteria. The UKLS trial applied the LLP Model in a target population of 50-74 year olds and participants were sent for annual screening if their 5-year risk of lung cancer exceeded 5% [34]. The high LLP Model threshold was chosen because this constrained the proportion of disease free individuals that would be screened. The crude average specificity of the LLP Model in the validation datasets will be determined. This will become the minimum acceptable specificity for the remaining lung cancer prediction models. If a models screen the same proportion of disease free individuals as seen in UKLS trial, then the costs incurred from false positives could be deemed acceptable. This requires the assumption that any true positive identified through screening, while requiring a screening cost, is a benefit and a willing cost. Since the specificity is fixed across the models, a model with a higher sensitivity will demonstrate an improved benefit to harm ratio.

The first objective is to identify a target population that would most benefit from a lung cancer screening programme. The UKLS trial restricted the target population to participants aged 50-75 years, with no additional restrictions. This population was selected because lung cancer risk below 50 years is negligible [1] and participants over 75 years may not be healthy for screening or for any subsequent treatment. The age range is also used by the NLST screening trial which restricted the target population to participants between 55-74 years, although the trial was further restricted to ever-smokers with a minimum 30 pack year smoking history. The additional ever-smoker restriction will not be enforced. Therefore, the target population has been defined to 50-75 years old, all participants outside this age range in the datasets will not be considered in the external validation.

Some models have further restrictions, such as not being applicable to never-smokers. To allow all the models to be compared, any participant who is in the target population (50-75 years old) but cannot have their risk calculated using a specific model, will still be considered when validating the performance of that model. They would automatically be rejected from screening, as their risk has not been calculated by the model. Therefore, when reporting the sensitivity and specificity they will be considered as not high risk or identified for screening.

The next stage is to define the minimum specificity. This is determined by the performance of the LLP Model at the 5% risk threshold for 5-year risk in the datasets, as prescribed by the UKLS specifications. The LLP Model reported an average specificity of 90% in the three applicable datasets and the results are presented in Section 7.4. This is an expected result as previously published validations found the LLP to have a specificity of 85% at the 5% risk threshold [69]. The remaining models will be evaluated in the target sample populations at the risk threshold that allows the model to achieve this specificity of 90%. This can restrict a high proportion of unnecessary screening while acknowledging some people who will not develop cancer will have to be screened while identifying lung cancer incidences.

In summary, for each model the risk threshold that allows the model to achieve a specificity of 90%, in the target population of 50-75 year olds, will be reported. At this risk threshold the sensitivity and PLR will also be reported. Any leading models will be identified by a higher sensitivity at the fixed specificity.

7.4 Liverpool Lung Project Model

The LLP Model was applied in three datasets at the 5% risk threshold. The model was applied in the lower risk of developing lung cancer ReSoLuCENT dataset, the higher risk of developing lung cancer CARET dataset and the MSH-PMH dataset. As highlighted in Table 7.1 the LLP Model reported a crude average specificity of 90%. As expected, applying the model at the high risk threshold allowed the model to report a high specificity. This is replicated in previous studies, where applying the model at the 5% threshold reported a specificity of 85%.

Dataset	Sensitivity	Specificity	PLR
ReSoLuCENT	11.79	90.48	1.238
CARET	27.01	80.78	1.405
MSH-PMH	23.56	96.37	6.490

Table 7.1: Liverpool Lung Project Model Validation Results

This high risk threshold, that allows the high specificity, is counter balanced by a poor sensitivity. The results were variable in the different datasets. There is a lower sensitivity in the ReSoLuCENT dataset and the converse is observed in the higher risk of developing lung cancer sample population in the CARET dataset. Across the three datasets the average sensitivity is approximately 20-22%.

In summary, the high risk threshold for the LLP Model would reject a high proportion of participants from screening. The model would avoid screening 90% of disease free individuals. However, this limits the proportion of lung cancer incidences that would be identified. The sensitivity was approximately 20-22% across the three datasets.

7.5 Pittsburgh Model

The Pittsburgh Model was applicable to all the ILCCO datasets. The Pittsburgh Model is a logistic additive model [85] where the generated risks are not continuous. As a result the participants are grouped at different risks and the sensitivity and specificity rates also increased in groups. The model is only applicable to ever-smokers; all never-smokers who were aged between 50-75 years old were still considered in the sensitivity and specificity rates although they would not be selected for selective screening.

To maintain a specificity of approximately 90% a risk threshold around 2.9-3% would be required. A conservative risk threshold option of 3% would limit screening of disease free individuals aged 50-75 years, in most real populations, to a maximum of 1 in 10 disease free individuals.

Dataset	Risk Threshold (%)	Sensitivity	Specificity	PLR
ReSoLuCENT	3.17	5.85	93.53	0.904
CARET	5.63	10.2	94.92	2.008
UCLA	1.45	35.48	91.85	4.353
NY Wynder	3.84	31.63	90.78	3.431
Singapore	0.4	35.44	90.87	3.882
New Zealand	0.98	32.35	93.43	4.924
CREST	3.84	41.9	91.47	4.912
Israel	4.66	23.11	90.26	2.373
ESTHER	2.61	33.51	90.22	3.426
Canadian	1.08	58.65	91.42	6.836

Table 7.2: Pittsburgh Model Validation Results

The sensitivity was an improvement upon the UKLS criteria and was consistently around 30-33%. This is a notable improvement over the UKLS criterion which reported a sensitivity of approximately 20-22%. However, in a direct comparison of the Pittsburgh Model and the UKLS criteria in the ReSoLuCENT, CARET and MSH-PMH datasets, the Pittsburgh Model failed to report an improved sensitivity in two of these datasets. This may suggest that the Pittsburgh Model would not improve upon the NLST criteria when applied to the same target populations.

To summarise, the Pittsburgh Model at the 3% risk threshold should allow a specificity of 90%. Here a sensitivity of 30-33% would be observed. The model should be evaluated at this 3% risk threshold in

some cohorts in direct comparison to the UKLS criteria as the current findings suggest the model would not offer an improved selective screening criteria, despite the overall improved sensitivity.

7.6 Hoggart Model

The Hoggart Model was applied to all the ILCCO datasets to predict 1-year absolute risk. The model was only applicable to ever-smokers but never-smokers were considered in the sensitivity and specificity rates where they would be rejected from selective screening.

The results indicate a risk threshold around 3% would allow the model to achieve a specificity of 90% in most screening programmes. However, the results were quite varied across the different datasets (Table 7.3). As expected the higher risk sample population in the CARET dataset reported a very high threshold (4.68%) in comparison to the New Zealand and Singapore datasets alongside a surprisingly low risk threshold in the MSH-PMH dataset (0.60%). The remaining results were contained between 2.15-3.67%; based on these results it was concluded a risk threshold of 3% should allow the Hoggart Model to maintain a specificity of 90% in most real world screening programmes in this target population.

Dataset	Risk Threshold (%)	Sensitivity	Specificity	PLR
ReSoLuCENT	3.67	10.72	90.95	1.185
CARET	4.68	11.81	90.28	1.215
UCLA	2.6	35.25	91.01	3.921
NY Wynder	3.18	32.09	90.59	3.410
Singapore	0.87	34.47	90.11	3.485
New Zealand	1.31	38.24	90.51	4.030
CREST	3.24	32.7	90.1	3.303
Israel	3.1	30.66	90.26	3.148
ESTHER	2.15	16.76	90.22	1.714
Canadian	0.6	56.69	90.21	5.791

Table 7.3: Hoggart Model Validation Results

Across the datasets the Hoggart Model reported a specificity around or slightly exceeding 30%. Table 7.3 shows consistent results between 30-35% in the majority of datasets. The Hoggart Model reports similar results to the UKLS criterion in a direct dataset comparison. Assessing the datasets applicable to the UKLS and Hoggart criteria there is little improvement in the sensitivity except for a large improvement in the MSH-PMH study. However, applying this model at the 3% risk threshold, as recommended, in the MSH-PMH study would dramatically lower the sensitivity. Overall, an approximate specificity of 30% would offer an improvement over the UKLS criteria. However, when comparing the Hoggart Model to other models applicable to the dataset there were more successful models identified.

The Hoggart Model should be applied at the 3% risk threshold for 1-year risk to allow the model to obtain a specificity around 90% in most populations. This allowed the model to obtain an approximate sensitivity of 30%, which is an improvement over the UKLS sensitivity. However, the results were not as conclusive when comparing the UKLS and Hoggart guidelines in the same dataset where there was no apparent improvement. Further testing in the same dataset of the Hoggart Model at the 3% and the UKLS criteria could assess which is a more successful selective screening tool.

7.7 $\text{PLCO}_{\text{M2014}}$ Model

The $\text{PLCO}_{\text{M2014}}$ Model was applicable to 5 of the ILCCO datasets. The model was applicable to the same datasets as the UKLS criteria so a direct comparison between their performances can be conducted. The

model was applicable to all the participants aged 50-75 years.

The risk threshold was quite variable across the 5 datasets and as observed in the UKLS, Pittsburgh and Hoggart validations the MSH-PMH dataset reported an unexpectedly low risk threshold. Evaluating the model in the remaining 4 datasets the results (Table 7.4) indicate a risk threshold of 3% would ensure 90% of diseases free individuals would be rejected from screening. This is based on the ReSoLuCENT dataset and the NY Wynder datasets which reported a threshold close to 3%. This is slightly higher, as expected, in the heavy ever-smoker CARET dataset, but we would expect that in a more real world population a lower risk threshold would be required to maintain a specificity of 90%. The UCLA slightly deviates from the other results with a lower risk of 1.02% however in the Pittsburgh and Hoggart validations the UCLA dataset also reported a lower risk threshold than observed in the other datasets. Taking into account these factors a risk threshold of 3% should allow a specificity of approximately 90%.

Dataset	Risk Threshold (%)	Sensitivity	Specificity	PLR
ReSoLuCENT	3.04	15.98	90.09	1.613
CARET	4.17	17.49	90.06	1.760
UCLA	1.02	45.16	90.02	4.525
NY Wynder	2.95	34.99	89.98	3.492
Canadian	0.86	59.75	90.1	6.035

Table 7.4: $\text{PLCO}_{\text{M2014}}$ Model Validation Results

The sensitivity results were quite variable but also very promising across the datasets (Table 7.4). The average sensitivity was approximately 35%, which is exemplary and a notable improvement in comparison to the UKLS trial and the Pittsburgh and Hoggart models. In a direct comparison to the UKLS guidelines the model showed a clear improvement, in the ReSoLuCENT and MSH-PMH datasets. Although this was not observed in the CARET dataset it is important to recall the UKLS criteria only reported a specificity of 80% in the dataset rather than the fixed 90% for the $\text{PLCO}_{\text{M2014}}$ Model. Indeed, this is supported by the higher PLR the $\text{PLCO}_{\text{M2014}}$ Model reported in comparison to the UKLS criteria in the CARET dataset. Additionally, the model improves upon the previous Hoggart and Pittsburgh validation results. In direct comparison of the sensitivity in the five datasets applicable to all the models the $\text{PLCO}_{\text{M2014}}$ Model reported a higher sensitivity.

In summary, the $\text{PLCO}_{\text{M2014}}$ Model is the most successful model currently evaluated. The model should be considered at the 3% risk threshold. This should allow the model in most screening populations to report a sensitivity of 35% for a specificity of 90%. This should be reviewed in cohort studies to assess if the impressive results can be replicated. In a direct comparison between models and the UKLS criteria in the datasets this model alongside the 2012 version consistently reported the leading performance and as such should be considered as a selective screening tool.

7.8 $\text{PLCO}_{\text{M2012}}$ Model

The $\text{PLCO}_{\text{M2012}}$ Model is applicable to five datasets similar to the $\text{PLCO}_{\text{M2014}}$ Model. Unlike the 2014 version the $\text{PLCO}_{\text{M2012}}$ Model is not applicable to never-smokers. All never-smokers excluded by the model will be considered by automatically excluding them from selective screening when reporting the sensitivity and specificity results.

Across the five datasets the mean risk threshold is 2.98% for 6-year risk and would generate a specificity of 90% in the datasets. However, in real world populations the risk threshold should be increased as the mean is distorted by the very low MSH-PMH dataset risk threshold. Indeed, across the remaining 4 datasets 3.5% is a reasonable estimate to allow the model to consistently achieve a specificity of approximately 90%. However, the risk thresholds were quite variable across the models so it is recommended that the

Dataset	Risk Threshold (%)	Sensitivity	Specificity	PLR
ReSoLuCENT	3.91	12.67	90.09	1.279
CARET	5.29	17.2	90.06	1.730
UCLA	1.25	45.62	90.02	4.571
NY Wynder	3.48	35.56	90.01	3.560
Canadian	0.96	60.49	90.1	6.110

Table 7.5: $PLCO_{M2012}$ Model Validation Results

$PLCO_{M2012}$ Model is evaluated at the 3.5% in additional validations to assess the creditability of this recommendation.

The $PLCO_{M2012}$ Model reported a very promising sensitivity. The crude average for the sensitivity across the datasets is just below 35%. A direct comparison between the PLCO models per dataset shows that they report similar sensitivity and PLR results. This should be expected because the two models consider the same predictors with just some recalibration and the inclusion of one additional variable to incorporate never-smokers in the 2014 version [82].

In summary, the two versions of the PLCO Models currently share the leading performance of all the models. In a direct dataset comparison they offer a clear improvement over other models and the UKLS guidelines. The $PLCO_{M2012}$ Model should be considered at the 3.5% risk threshold for 6-year risk to maintain a sensitivity of 35% and specificity of 90%, although it would be beneficial to evaluate the model in cohorts to assess if the impressive performance can be replicated.

Decision makers would need to decide between the two PLCO models. The $PLCO_{M2014}$ Model may probably be preferable in screening programmes because it incorporates never-smokers. Although incorporating never-smokers in the risk prediction the $PLCO_{M2014}$ Model was unable to identify any additional lung cancer incidences, so there should be no preference towards the $PLCO_{M2014}$ Model. However, if the selective screening programme wanted to identify lung cancer only in ever-smokers then the $PLCO_{M2012}$ Model should be implemented at the 3.5% risk threshold because more individuals with lung cancer would be identified in comparison the 2014 version.

7.9 Bach Model

The Bach Model was applicable to five datasets. This included the MSH-PMH datasets which generated unusual results in comparison to the other datasets. The Bach Model was applied to predict 10-year absolute risk as presented in the original article [77]. The model is very restrictive and could only be applied to ever-smokers with a minimum 30 pack year smoking history. This resulted in a large volume of participants being automatically excluded by the model criteria. These will still be considered when reporting the prediction rules.

The risk thresholds for the Bach Model were extremely variable across the datasets (Table 7.6). There were 2 datasets with a very low risk threshold for 10 year risk around 0.25%; this may be a result of the high automatic exclusion using the Bach Model criteria which boosts the specificity. Another 2 datasets generated a risk threshold of approximately 6.5% and finally in the CARET dataset a very high risk threshold of 12.25% was required to maintain a specificity of 90%. The results were so variable an accurate prediction for a risk threshold to maintain a specificity of approximately 90% could not be estimated.

Dataset	Risk Threshold (%)	Sensitivity	Specificity	PLR
ReSoLuCENT	6.81	11.31	90.09	1.141
CARET	12.25	13.56	89.99	1.355
NY Wynder	6.05	36.1	90.01	3.614
CREST	0.25	30.79	89.76	3.007
Canadian	0.26	58.77	90.1	5.936

Table 7.6: Bach Model Validation Results

Further evaluation of the Bach Model is required if considering this model as a selective screening tool in order to assess if a consistent risk threshold can be identified. However, based on the results observed in the validation the Bach Model would be too variable in different populations to recommend the model as a universal screening tool.

7.10 Spitz Model

The Spitz Model was only applicable to the CREST dataset. The model can be run recursively but is applied for 1-year risk because the model was applied over this duration in the original article [75]. The Spitz Model can be applied to all participants in the 50-75 years age range which may improve the model performance because no participants are automatically excluded.

The Spitz Model exhibited promising results in the CREST dataset. The model was applied at the 1.31% risk threshold to achieve a specificity of 90%. The model should be validated at this threshold in new environments to assess if this will maintain the specificity of 90%. The sensitivity was very promising at 40%; if this high standard can be replicated in other studies then the model would be very successful. This sensitivity is higher than the two PLCO models, which currently exhibited the leading performance. Unfortunately, the models cannot be directly compared because the PLCO Models were not applicable to the CREST dataset. It would be advisable to compare these models in the same datasets at the identified risk thresholds to assess which would be more successful as a selective screening tool.

Dataset	Risk Threshold (%)	Sensitivity	Specificity	PLR
CREST	1.31	40	90.1	4.040

Table 7.7: Spitz Model Validation Results

In summary, the initial results were very promising and showed the best performance of all the models currently validated. Further analysis is required in external datasets to review if the sensitivity and specificity results can be replicated at the 1.31% risk threshold for 1-year risk. The Spitz Model should also be compared to the PLCO Models in the same dataset because these were the other leading models. A direct comparison of the models' performance would indicate conclusively which model would be more successful as a screening tool. Currently an inference cannot be made between the models.

7.11 African-American Model

Like the Spitz Model the African-American Model was only applicable to the CREST dataset. The model could be applied to all participants in the age range which could help improve the model performance. However, the model was applied in a different dataset to the model's target population which could negatively influence the performance of the model. The African-American model predicted lung cancer risk for 5 years.

The African-American Model required a high risk threshold at 5.19% to report a specificity of 90%. At this threshold the model reported a sensitivity of 27%. This was similar to the UKLS guidelines and any future validations should assess the African-American Model at the 5% or 5.19% risk threshold. The African-American model did not report a high sensitivity in comparison to other models in the CREST dataset. However, testing the model in a more appropriate sample population of African-Americans may improve the model performance.

Dataset	Risk Threshold (%)	Sensitivity	Specificity	PLR
CREST	5.19	26.67	88.4	2.299

Table 7.8: African-American Model Validation Results

The African-American Model underperformed in the CREST dataset in comparison to other models. However, the model was applied in a different sample population rather than the target population. Validating the model in a correct sample population may allow the model to excel and this may still be the most appropriate model for an African-American population considered for selective screening.

7.12 Summary

The models were validated in comparison to the UKLS screening guidelines. The UKLS programme was designed to identify high risk participants. There was an average specificity of 90% which allowed the criteria to reduce unnecessary screening. As a result of the high risk individuals identified for screening the sensitivity suffered at 20-22%. The validations discovered that some models could improve upon the UKLS screening guidelines by increasing sensitivity without reducing the specificity below 90%.

The PLCO models offered the largest improvement, with a sensitivity of around 35% for a specificity of 90%. The $\text{PLCO}_{\text{M2012}}$ Model should be applied at the 3.5% risk threshold and the $\text{PLCO}_{\text{M2014}}$ Model requires a 3% risk threshold. While other models also improved upon the UKLS guidelines they did not report the same level of performance as the PLCO models across all the datasets and in a direct comparison between models in the datasets. The results indicate the PLCO models would be more successful as a selective screening tool. It is recommended the PLCO models should be evaluated in cohort studies as these case-control datasets may lead to optimistic results. The PLCO models should be evaluated at the risk thresholds identified in this validation to review if the models can consistently perform to the same standard. If they can report a similar performance level then it would be beneficial to consider these models as a selective screening tool.

The next stage of the project is to update prediction models to assess if an improved lung cancer model can be devised.

CHAPTER 8

Literature Review of Updating and Aggregating Prediction Models

8.1 Introduction

Single prediction models can be updated or multiple models aggregated into a meta-model in an attempt to create an improved prediction model. Single model updating can involve recalibrating the model, re-estimating the model parameters or extending the model with new parameters. This aims to improve the model's predictive and discriminative ability. An improved model is desirable as it combines the evidence in the model building dataset and the model updating dataset which can create a model that is robust in differing populations and a more successful selective screening tool.

Methods to update a single model and aggregate multiple prediction models are presented and discussed in a literature review. The methodology is presented with examples of where they have been used while considering the advantages and limitations of the methods.

The practicality of the methods for lung cancer prediction models will then be discussed. The appropriate methods will then be applied to lung cancer models to evaluate their ability to create a more robust model.

8.2 Objectives

The chapter will conduct a literature review on methods to update a single prediction model or aggregate multiple prediction models. To achieve this the chapter will;

1. Identify model updating methods and present their methodology.
2. Demonstrate where the methods have been applied previously to update prediction models and report on their success in creating a more robust model.
3. Discuss the potential benefits, limitations, and concerns of each method.
4. Assess the practicality of the methods for lung cancer prediction models.
    - Identify which methods will be applied to the lung cancer prediction models.

8.3 Introduction to Review of Single Updating Methods

There are several approaches to update a single prediction model using a model updating IPD. These include updating the intercept, recalibrating the parameters, and extending the model. The different methods can be utilised in different scenarios depending on how the original model performed in the dataset. A model that had a successful discrimination and prediction rules but a poor calibration may benefit from minor updating whereas a model that was less successful with a poor calibration and discrimination may

require more extensive model updating [45, 138]. The available methods to update a single model in a new dataset are listed as follows;

1. No updating
2. Model Recalibration
   (a) Update Intercept
   (b) Recalibration of the intercept and slope
3. Model Re-estimation
   (a) Recalibration and selective re-estimation
   (b) Re-estimation
4. Model Extension
   (a) Re-estimation and extension
   (b) Selective re-estimation and selective extension with recalibration
   (c) Re-estimation and selective extension without recalibration

8.4 No Model Updating

When a model exhibits a strong overall performance when externally validated then model updating may not be required. Unnecessary updates could result in over-fitting if additional parameters are only included to explain the minor difference between predicted and observed risk in the dataset. Additionally, the new model may have a poorer performance in external sample populations in comparison to the original model. This will unnecessarily create a new model to the existing prediction model literature generating confusion about which model should be used to predict an individual's risk or identify high risk participants in a selective screening programme.

8.5 Model Recalibration

The first updating methods are recalibration techniques [58]. These aim to improve the accuracy of the model estimate based on the observed incidence rates in a sample population. These methods can improve the model calibration, while the discrimination and prediction rules remain unaltered because all participants' risks are adjusted equally [45]. Therefore, these methods are beneficial for a model with a strong discriminative ability but needs recalibration. These methods are most likely to be applied to successful models which are now being applied in a different target populations and require recalibration based on a sample population of the new target population.

8.5.1 Updating the Intercept

This method updates the intercept of the logistic regression equation. This centers the mean risk estimated by the model for all participants to reflect the mean incidence rate observed in the sample population. All participants are centered around the mean by having their risk scaled up or down equally. While the discrimination and prediction rules remain unaltered, the risk threshold at which the model reports a certain sensitivity and specificity will change.

To update the intercept the correlation factor is calculated, which is based on the observed incidence rate and the predicted risk in across the dataset;

$$\text{Correlation Factor} = \ln\left(\frac{\text{Observed Rate}}{1-\text{Observed Rate}} \times \frac{1-\text{Mean Predicted Risk}}{\text{Mean Predicted Risk}}\right) \tag{8.1}$$

$$\text{where Observed Rate} = \frac{\text{Number of Participants With The Condition}}{\text{Total Participants}} \tag{8.2}$$

This is added to the original intercept to create the final model.

While this method can improve the calibration it could be seen as a quick fix solution. The model could require repeated updating in different target populations before being used to estimate an individual's risk. Before a model would be considered for implementation as a clinical utility in multiple differing populations, a final baseline incidence rate would need to be determined [45].

This method has been commonly applied to update clinical prediction models. Studies found the calibration was improved when the original model exhibited a poor calibration in the new populations [139, 140, 141]. The studies also concluded that this method was beneficial when the model updating sample population was small [139]. However, it is important that the sample population used to update the model reflects the target population where the model will be applied. Otherwise, model updating may be detrimental and the new model may underperform when implemented [140].

In summary, the method improves the model calibration while the discriminative ability and prediction rules are unaltered. Therefore, this method is most appropriate for a robust model that needs minor adjustment of the estimated baseline risk to reflect the incidence rate observed in a sample population that reflects the new target population.

8.5.2 Recalibration of Intercept and Slope

This approach updates the intercept and the variable coefficients [142]. The recalibration factor can be determined by the line of best fit in a plot of the observed and predicted risk as follows;

$$ObservedRisk = \alpha_{Calibration} + \left(\beta_{Calibration} \times (LinearPredictor)_{Original}\right) \tag{8.3}$$

For an original log odds model the line of best fit is plotted between the linear predictor and the natural logarithm of the observed risk. A perfect original model has an intercept (Alpha) of zero and gradient (Beta) of one. The logistic regression is updated using this information. The intercept is the sum of the original intercept and the alpha (calibration) and the coefficients are scaled by beta (calibration) [45, 138].

$$LinearPredictor_{New} = \alpha_{New} + \left(\beta_{calibration} \times (LinearPredictor)_{Original}\right) \tag{8.4}$$

The method can improve the model calibration as the estimated risks are modified to reflect the observed incidence rate in the sample population. Therefore, this method is most effective for a model with a good discriminative ability which is applied in a new population to provide accurate risks. If the model has other deficiencies, such as a poor discriminative ability, then this method may not be preferred; scaling all the variables by a constant will not address any weaker variables which should be removed or constrained in the model.

Model recalibration has been applied to update clinical prediction models with mixed success. The method is most successful when there are only small differences between the original model building and model updating datasets [141] and the mode updating dataset is relatively small. If the model exhibits a poor calibration, potentially due to over-fitting, then this can be corrected by adjusting the calibration slope [143]. However, some studies have suggested that updating the intercept led to a more robust model when applied in external populations [140, 141] rather than recalibrating the intercept and the slope.

8.6 Model Re-estimation

The next series of model updating methods re-estimate variables in the prediction model. These methods identify which variables are significant to the model's predictive ability in the validation dataset. The variables that are not significant have their regression coefficients reduced or shrunk in the model. This means these variables are not weighted as heavily in the updated prediction model and therefore do not have as large an impact in predicting ones' risk of developing the disease. The model re-estimation techniques combine the original evidence, in the model building dataset, and evidence in the external sample population. Combining the evidence from the original dataset and an external sample population may allow a more robust model to be devised that could be applied across several different target populations.

These methods can improve the model calibration and also influence the model discrimination and prediction rules. Therefore, these methods can be advantageous when an external validation identifies that the model's discriminative ability is poor.

There are two methods to re-estimate the model. To apply these methods a technique called shrinkage is required which is introduced below.

8.6.1 Shrinkage

Shrinkage is applied to a prediction model such that the "regression coefficients can be shrunk either towards the mean incidence rate or towards the recalibrated values" [138]. This is applied to avoid over-fitting the model to the external data. This is particularly beneficial for large models with many variables that are updated on small datasets where over-fitting could be prevalent [144].

The shrinkage factor is a value between [0, 1], and is dependent on the calibration of the original model in the external data. A well calibrated model will require a smaller revision towards the mean or null model. Whereas a model with a poor calibration, will have the regression coefficients heavily shrunk towards the mean or more predictive null model. This means the weaker variables have less of an impact in the updated model. This means an individual's risk cannot deviate as far as originally from the observed dataset incidence rate or the more predictive null model based on an inadequate variable the model.

8.6.2 Recalibrated Null Model and Selective Re-estimation

This method identifies the variables in the original model that are most predictive of the disease in the model updating dataset, the model only including the parameters that significant to the model predictive ability is defined as the null model. This identified using either a forward or backward selection process. Prior to identifying the null model, the original model should be recalibrated using the techniques identified in Section 8.5.2. It is important to recalibrate the original model otherwise for a poorly calibrated model the majority of variables will show a significant improvement in the model goodness of fit and be included in the null model. Once the recalibrated null model has been identified then shrinkage is required between the recalibrated null model and recalibrated original model. The shrinkage reduces the weighting of the parameters that were identified as not being significant in the original model. Therefore, these parameters do not have as large an impact in deviating ones' risk from what was predicted in the null model.

The shrinkage coefficient is calculated using an F-test which evaluates the difference in performance between the null model and recalibrated original model. The factor, denoted as ĉ, is calculated as follows;

$$\hat{c} = \frac{\left|F_{1|0} - 1\right|}{F_{1|0}} \tag{8.5}$$

Then the null model and the recalibrated original model can be combined using the shrinkage factor to create the selectively re-estimated model as follows;

$$LP_{Recalibrated\&SelectiveRe-estimated} = ((1 - \hat{c}) \times LP_{NullRecalibratedModel}) + \hat{c}LP_{RecalibratedModel} \tag{8.6}$$

The final model is a weighted average of the null recalibrated model and the recalibrated original model [138]. If the shrinkage factor is zero, indicating the variables not selected in the null model offered no improvement in the model's predictive ability, this is reflected in the final model being simply the null model. The variables identified for the recalibrated null model remain unaltered while the remaining variables are selectively re-estimated. Therefore, if the null model includes all the variables then this method is redundant and identical to simple recalibration (Section 8.5.2).

This method can improve the model calibration, discrimination, and prediction rules. By evaluating the variables in the second dataset where model updating occurs, this could create a more robust model as the key variables associated with the disease are identified. It is important to identify any distinct characteristics in the model updating dataset that may influence whether variables are considered predictive.

Model re-estimation has been applied to update clinical prediction models. Some studies found the updated model failed to offer a large improvement in discrimination or improve the calibration in comparison to more simplistic methods [141]. Additionally, if the original model has a good discriminative ability then simple recalibration may be more appropriate. Conversely, other studies found this was the best method when a large dataset is available to update the model [139]. Additionally, when variable effects are heterogeneous between the development and validation samples and calibration plots show inconsistent predictions across the whole range of predicted probabilities, then re-estimation of individual predictors or even inclusion of additional predictors is the most beneficial method [143]. Re-estimation towards the null model is most successful in a large dataset where the model building and model updating datasets are heterogeneous.

8.6.3 Shrinkage and Re-estimation

The second re-estimation method does not recalibrate the original model. All the parameters are re-estimated and then shrunk towards the mean incidence rate observed in the external sample population. The evidence from the external sample population is incorporated through the mean rates which may create a more robust model in new target populations.

Firstly, the shrinkage factor is calculated using the F-statistic of the original model as follows;

$$\hat{c} = \frac{|F_{LPModel} - 1|}{F_{LPModel}} \tag{8.7}$$

This factor allows the variable coefficients to be re-estimated towards the mean observed incidence. The method modifies the intercept, $\bar{Y}$ as the mean incidence in the sample population. Participants are then distributed around the mean using their value for the variable in comparison to the mean variable value, $\bar{X}_i$. These factors are then scaled using the shrinkage coefficient.

$$LP_{Re-estimated} = \bar{Y} + \hat{c}\beta_1 \left(X_1 - \bar{X_1}\right) + \ldots + \hat{c}\beta_p \left((X_p - \bar{X_p}\right) \tag{8.8}$$

The method improves the model calibration because the predicted risks are distributed around the observed incidence in the sample population. The method can also improve the discrimination and prediction rules. However, a biased dataset (collection based on unusual occurrences for one or more of the variables) could hinder the model in new populations because the observed mean for the variable coefficient could be unrealistic. Re-estimation towards the mean does not evaluate each variable independently therefore will not identify any underlying problems to accurately estimate risk caused by irrelevant variables included in the prediction model.

Re-estimation towards the mean has been applied to update prediction models. This method has been successful when the original model was poorly calibrated and also reported a poor discriminative ability [139]. Additionally, for a model with a poor discriminative ability and a large model updating dataset available, then extending the model further improved the model performance rather than re-estimation [145].

In summary, this method is most successful for a model with a strong discriminative ability but requires recalibration for accurate predictions in a new target population where the model updating dataset is a

sample population of the new target population. If the model has a poor discriminative ability, then model extension or the other model re-estimation method (Section 8.6.2 may be more appropriate methods.

8.7 Model Extension

The final set of single model updating methods are defined as model extension. These methods incorporate additional variables in the prediction model. A more robust model could be created because additional key variables that explain risk can be incorporated.

Although, there are potential risks with model extension including over-fitting by attempting to explain the difference between the predicted and observed risks rather than the underlying condition [144]. The model may then underperform if applied in different target populations. These concerns can be avoided by only considering variables with a significant association with the disease. Additionally, the original model should be recalibrated to avoid all additional variables being included because they improve the model goodness of fit due to an original model calibration deficiency.

To identify additional predictors to be included in a model the hazards or odds ratios can be evaluated. If a statistically significant ratio is observed then the variable should be considered. However, a variable associated with the disease does not guarantee an improvement in model performance as "Ware and Pepe showed simple examples in which enormous odds ratios were required to meaningfully increase the AUC" [139, 146, 147]. It is important to consider not including too many additional predictors which may not significantly improve the model performance and could result in over-fitting and the model may then be unsuccessful in new populations.

8.7.1 Re-estimation and Extension

To apply this method, the first stage is to recalibrate the original model. The model variables are re-estimated towards the recalibrated null model using the techniques previously presented in Section 8.6.2. The model is then extended to include all additional variables available in the model updating dataset or identified as being associated with the disease through previous testing.

As previously discussed, including all additional parameters may not improve the model performance, and could limit the model in new populations. This approach can cause over-fitting by considering too many parameters especially if the model updating dataset is small [144]. This could create a final model that while being successful when internally validated may perform poorly in new populations. A more selective model extension technique that identifies the appropriate variables, rather than including all available different information, could create a more robust model.

Previous studies have been critical of this method when applied to clinical prediction models. While this method has been shown to be beneficial if the original model had a poor overall performance [143, 145] the updated models have been criticised. This method should be applied with caution in small datasets, because the method places a comparatively large significance on the updating data, and hence would be prone to peculiarities of the updating dataset [141]. In a study updating a prostate cancer prediction model, while additional markers have the potential to improve discrimination these should be selected using forward selection rather than all being included as this created an inferior prediction model when externally validated [145].

8.7.2 Selective Re-estimation and Selective Extension with Recalibration

The first stage is to recalibrate and re-estimate the original model which has been presented in Section 8.6.2 before the model is extended. New variables can be included if they are shown to have an association with the condition in the model updating dataset. A second approach is to include the new variables using forward selection from the selectively re-estimated model if they offer a significant improvement to the model goodness of fit, measured by chi-squared. Incorporating the new parameters using forward selection

is preferred as previous studies have shown incorporating new variables based on their association with the condition may not lead to an improved calibration or discriminative ability [139, 146, 147].

This method can create a robust model with an improved calibration, discrimination, and prediction rules. Including only parameters that improve the model goodness of fit limits the risk of over-fitting caused by including all additional variables. The main concern with this method is that if the model updating dataset has an unusual participant recruitment and variables may be included in the final model that do not assist in estimating the likelihood of a disease developing. One such example would be if all diseased individuals are collected dependant on having a different prior existing condition; this is likely to show a significant association with the condition which may not be observed in real populations. Precautions can be taken by conducting a review of the dataset recruitment process and analysing the population demographic. The model should also be externally validated in comparison to the original model to assess if a more robust model has indeed been developed.

Models have commonly been extended and is likely to be an increasingly popular method as genetic markers are incorporated into existing models, as has been observed in lung cancer (Section 3). One study found when predictor effects are heterogeneous between the development and validation samples then extensive updating was required [143]. Re-estimating the existing variables and then selectively including additional variables was a successful method [143]. A separate study compared different strategies for model extension for a new variable. The study found that when the dataset used to extend the prediction model was small, simple re-estimation methods led to the largest increase in discriminative ability of the prediction model, but as the available sample population increased more extensive extension methods outperformed re-estimation techniques [145].

8.7.3 Re-estimation and Selective Extension without Recalibration

This method is similar to the model extension method described in Section 8.7.2, however the preliminary work differs. The original model is not recalibrated because the model is shrunk around the mean incidence rate observed in the model updating dataset, as presented in Section 8.6.3, then extended using forward selection to identify new variables that improve the model goodness of fit, as presented in Section 8.7.2.

Selective extension limits the risk of over-fitting [144] because only variables associated with the disease are included rather than including all available variables. However, the method could still be affected by an unusual recruitment of diseased and disease free individuals in the model updating dataset. Additionally, a poor original model calibration may see a large quantity of new variables included to rectify the original model calibration deficiency. This could result in an extended model that underperforms in new populations because the included variables explained the variance between the predicted and observed risk in the model updating dataset rather than the underlying condition. Therefore, it is important to assess how the model performs in new populations, and whether the new variables included in the extended model created a more robust model in comparison to the original model.

In summary, this method has been successful when the original model is poorly calibrated and reported a poor discriminative ability [143]. Indeed, with a large dataset available to perform the model updating, for a model with a poor original performance, model extension outperformed more simplistic model updating methods [145]. Additionally, using stepwise forward selection created a more robust model than including all available additional variables [145]. However, any peculiarities in the model updating dataset will be reflected in the new model, which then could under-perform in new populations [141].

8.8 Summary of Methods to Update a Single Prediction Model

There are a number of available methods to update a single prediction model which can be advantageous in different scenarios. Model recalibration only alters the model calibration. These methods would be preferable for a model with a strong discrimination and prediction rules performance but a poor calibration in a sample population, which is reflective of a new target population in which the model will be applied.

However, for a model with a poor overall performance, more extensive updating methods may be beneficial. This can include re-estimating or extending the model to incorporate new variables. The updated models will improve the original model calibration in an internal dataset but the calibration, discrimination, and prediction rules should be evaluated in external datasets to evaluate if a more robust model has been produced in comparison to the original model.

8.9 Aggregating Multiple Prediction Models

Multiple prediction models can be aggregated into a meta-model. The different methods are presented and evaluated. These methods aggregate models built in different sample populations and combine their evidence in an attempt to create a more robust model. This avoids disregarding models in preference to creating a new model based exclusively on the available dataset. These methods can improve model prediction as "Madigan and Raftery note that averaging over all the models in this fashion provides better average predictive ability than using any single model" [151].

Currently, there has not been a large volume of meta-modelling methods reported. This is because of an emphasis on creating new models rather than aggregating existing models. However, there has been an advance in methods to aggregate models in recent years. These methods are designed to either use the models and their standard errors in the model building dataset or use an external dataset to aggregate the models.

The next section of the review will present the methods to aggregate multiple prediction models.

8.9.1 Model Averaging

Model averaging, presented by Debray et al (2014) [148], combines multiple models based on their calibration in an external dataset. The models are weighted with better calibrated models assigned a large weight in the final model.

To apply this method, the original models are initially recalibrated in the external dataset, using the method presented in Section 8.5.2. This ensures that poorly calibrated models, potentially due to being applied in a different sample population, are not nullified by being assigned a low weight in the meta-model.

Next the models are applied to the external dataset and a risk for each patient, p, can be generated from every model (1, . . . , M). The final risk for each individual is the weighted sum of the individual model predictions as follows;

$$\bar{p_i} = \sum_{1}^{M} w_i p_i, \tag{8.9}$$

$$\text{where } \sum_{1}^{M} w_i = 1 \tag{8.10}$$

The weights for each model $(w_1, \ldots, w_M)$ now needs to be determined. The simplest solution assigns each model an equal weight $(\frac{1}{M})$. However, this does not reward stronger performing models. Debray proposed to assign a weight to each as follows;

$$w_n = \frac{e^{-0.5BIC_m}}{\sum_{l=1}^{M} e^{-0.5BIC_l}}, \tag{8.11}$$

$$\text{where } BIC_m = -2l_m + u_m ln(N) \tag{8.12}$$

Here, N is the number of patients in the model aggregation IPD and u the number of estimated parameters used to recalibrate the model. Since all the methods will be recalibrated using an intercept

and scaling parameter u will be fixed at 2. However, if models require more extensive updating then this is reflected with a higher u as a penalty [148]. Then, for each model the log-likelihood, l, is determined as follows;

$$l_m = \sum_{i=1}^{N} y_i ln(p_{im}) + (1 - y_i) ln(1 - p_{im}), \tag{8.13}$$

$$\text{where } y_i = \begin{cases} 0 \text{ if disease not present} \\ 1 \text{ if disease present} \end{cases} \tag{8.14}$$

In a hypothetical scenario, a model with perfect calibration would assign a probability of 0 to disease free individuals and 1 to all diseased individuals. This would result in a log-likelihood of zero. As the model accuracy lowers from this perfect scenario, the log-likelihood increases and as a result the model has a lower BIC in the final weighting.

This method is advantageous because it rewards better calibrated models by assigning them a higher weight in the meta-model which could create a more robust meta-model. The method is straightforward to apply and applicable to a range of diverse model forms. Additionally, this method has reported some favourable results. One study found the meta-model had an improved calibration and discriminative ability [141]. This was supported by an additional study finding the aggregated meta-model outperformed the existing models [150]. Another study found model averaging created an improved model, although only reporting a minor improvement it did consistently improve upon the original models when validated [149].

However, there are some disadvantages to model averaging. Updating before aggregating the models could create a meta-model that is then successful in the sample population but has a poor performance if the target population differs from the sample population. Secondly, to use the final meta-model the original models are run separately to generate a participant's risk and then combined. As a result of this the meta-model can be large, complex and cumbersome which health professionals and the public may be reluctant to use [148]. This was highlighted in a study in which the final meta-model was too complicated to present [151], although this can be averted by creating an easy to use model interface. The method also assigns extreme weights to some models so the final model is heavily determined by a combination of a few models rather than all the models considered. This is a consequence of the method being similar to Bayesian model selection and assumes "only one of the models is correct" [148]. Therefore, it will not utilise the full wealth of evidence available by rejecting some of the models. Finally, the weightings are based upon the model calibration. As a result, the final model may have an improved discriminative ability, as models with a poorer discriminative ability may be minimised in the final model if they also have a poor calibration. Additionally, other studies found meta-model approaches are a relatively new area of research and needed to be applied to more clinical prediction models before they can be recommended [141].

8.9.2 Stacked Regression

The next method, developed by Debray et al. [148], called stacked regression, is "a method for forming linear combinations of different predictors to give improved prediction accuracy" [152]. The method aggregates the models' linear predictors based on their calibration performance in an external dataset using a minimising likelihood function.

This method does not recalibrate the models using the external dataset before aggregating the models. The model variables are identified by calculating the maximum likelihood of $\alpha_0 \dots \alpha_M$ by minimising the following function in the dataset with N participants for $M + 1$ unknown constraints;

$$= \sum_{i=1}^{N} (y_i ln(1 + e^{-a_0 - \sum_{m=1}^{M} a_m LP_{im}})) + ((1 - y_i) ln(1 + e^{a_0 + \sum_{m=1}^{M} a_m LP_{im}})) \tag{8.15}$$

$$\text{With constraint} \quad a_m \geq 0 \tag{8.16}$$

Here y_i is a binary variable which takes the value 0 if the disease is not present in the individual and 1 if it is present. For each model $(1, \ldots, M)$ for an individual, i, their risk is defined as LP_{iM}. α_0 is the intercept parameter for the "optimal baseline risk for the validation study" [148]. The constraint for the alpha values to be greater than zero eliminates co-linearity in the meta-model. This inhibits the inclusion of two similar models, such as a model and a recalibrated version of the model, in the meta-model because they would negate each other.

The weighting of the variables in the final model is determined by minimising the likelihood function for all the variables as follows;

$$\text{Intercept: } \hat{\beta}_0 = \hat{\alpha}_0 + \sum_{m=1}^{M} \hat{\alpha}_m LP_{0m} \tag{8.17}$$

$$\text{Parameters: } \hat{\beta}_i = \sum_{m=1}^{M} \hat{\alpha}_m LP_{im} \tag{8.18}$$

These variables are the different variables in the original models and if a variable is not present in a specific model then the alpha is zero so as to not influence the final aggregated model. Then the final logit model can be calculated as follows;

$$logit^{-1} = \hat{\beta}_0 + \sum_{i=1}^{K} \hat{\beta}_i x_{jk} \tag{8.19}$$

There are advantages to the stacked regressions approach to aggregate multiple prediction models. Firstly, this method creates a final formula that is neat and simpler to apply rather than applying multiple distinct models and weighting these. Additionally, the method does not assign an extreme weight to individual models in the final meta-model so the method is a fairer combination of their respective evidence. This method has been implemented and devised a more robust model in comparison to the original models in one study [141]. This was supported by other independent studies which discovered stacked regressions decreased calibration error rates and yielded an improved predictive performance [152, 153].

However, there could be issues with a large quantity of variables in the model designed to measure the same exposure. This can be seen in lung cancer prediction models such as smoking history (i.e. pack years, CPD, smoking duration, quit duration) or family history of cancer (i.e. any cancer, lung cancer, present in more than 2 family members) and these different variables will all need to be incorporated separately in the meta-model. Additionally, the method can be difficult to apply and is more restrictive as to which original models can be incorporated into the meta-model. Different model designs will not have a similar linear predictor or a single coefficient for each variable; some may have cubic, splines, or conditional coefficients. For example, for one variable age, i, considered in two distinct models, X and Y, could be incorporated into each model as follows;

$$i_x = \alpha \times Age \tag{8.20}$$

$$i_y = (\beta_0 \times Age) + (\beta_1 \times Age) \tag{8.21}$$

$$\text{where } \beta_0 = \begin{cases} x, \text{ if age } < 65 \\ 0, \text{ if age } \geq 65 \end{cases} \tag{8.22}$$

$$\text{where } \beta_1 = \begin{cases} 0, \text{ if age } < 65 \\ y, \text{ if age } \geq 65 \end{cases} \tag{8.23}$$

These two variables could not be combined because of the different model forms. The systematic review (Section 3) supports these concerns as distinct model forms were identified. Additionally, some models consider scaling factors in the linear predictor which is observed for lung cancer in the form of age, gender,

and smoking status specific incidence rates. Therefore, combining the models using stacked regressions is problematic. In the original article Debray et al. [148] assumes that there is a set of core variables across all the models and the models have the same form. However, this is not commonly the case and many distinct models have been designed to predict the same outcome. A study highlighted this concern that the quantity of models averaged is generally low because of difficulties combining the distinct clinical models [152]. A solution is to reduce the pool of models to models with a similar form. However, this could exclude successful models and their evidence from the final meta-model. Finally, not recalibrating the original models may result in some successful models being unfairly assigned a lower scaling factor if applied in a different sample population. It has been argued that not recalibrating the original models means the meta-model borrows less information from the model aggregation dataset in comparison to model averaging [150]. This is a relatively new field and more research is required before stacked regressions can be recommended [141].

8.9.3 Bayesian Model Averaging

The next approach to aggregate models is Bayesian Model Averaging (BMA), a similar method to the weighted averaging method (Section 8.9.1) [148].

Bayesian model selection methods can be used to identify the strongest model, often from a series of nested models, in an external dataset. However, BMA combines multiple models into a meta-model [154]. The methodology for BMA has been available for a considerable time [155].

The models are weighted based on their calibration in an external dataset but the methods are applied without model recalibration. This may result in successful models being unnecessarily assigned a low final weighting if the dataset is different from the model building dataset. Each model is assigned a weight based on their calibration in comparison to the other models, the final meta-model is the sum of the weighted averages from the different models. For example, the aggregated probability for a patient, j, using models $1, \ldots, M$ is expressed as;

$$Y_j = \sum_{i=1}^{M} w_i \hat{Y}_j \tag{8.24}$$

$$\text{Where } \hat{Y}_j \quad \text{is the risk from one model for person } J \tag{8.25}$$

The weights for BMA are determined by calculating the posterior model which is expressed as $Pr\left(M_j|Data\right)$. This approach then compares different models together using the Bayes factor. The following equation is used to assess the 'evidence' of preferring Model 2 in comparison to Model 1;

$$\text{Bayes Factor: } B_n = \frac{Pr(M_2|Data)}{Pr(M_1|Data)} \div \frac{Pr(M_1)}{Pr(M_2)} \tag{8.26}$$

Setting all prior odds as equal, $Pr(M_1) = Pr(M_2) = Pr(M_M)$, means that there is no initial bias towards assigning a higher weight to any model. This is the case in most instances, where there is no information to determine whether a model should be assigned a higher weight. As a result, the Bayes Factor becomes a measure of posterior odds between the models.

However, a second version of BMA could be applied which offers a bias towards models based on their discriminative ability. This aims to allow models that are successful at identifying individuals with a disease to be higher rewarded. Models with a stronger AUC performance can be assigned a higher weight by assigning each model prior odds as follows;

$$\Pr(M_j) = \frac{Pr(M_j)}{\sum_{i=1}^{K} Pr(M_i)} \tag{8.27}$$

$$\text{where} \sum_{i=1}^{K} Pr(M_i) = 1 \tag{8.28}$$

The posterior odds can be calculated based on the model calibration and the final model weights can be determined;

$$w_j = Pr(M_j|Y_n = y_n) = \frac{m_j \times Pr(M_j)}{\sum_{i=1}^{K} m_i \times Pr(M_j)} \tag{8.29}$$

$$\text{where} \sum_{j=1}^{K} w_j = 1 \tag{8.30}$$

Here $Pr(M_j)$ is either equal for all models when considering a non-informative prior or calculated using equation 8.9.3 to incorporate the discriminative ability into the weights as an additional weighting. To calculate the weightings m_j needs to be calculated as follows;

$$m_j = \int L_j(\theta_j) p_j(\theta_j) d\theta_r \tag{8.31}$$

This integral can be converted, and was presented as follows [155];

$$log(m_j) = log(L_j(\theta_j)) - \frac{d_j}{2} log(n) \tag{8.32}$$

$$\text{where } l_j = \sum_{i=1}^{N} y_i ln(p_{il}) + (1 - y_i) ln(1 - p_{il}), \tag{8.33}$$

$$\text{where } y_i = \begin{cases} 0 \text{ if disease not present} \\ 1 \text{ if disease present} \end{cases} \tag{8.34}$$

This is the log likelihood minus the dimension of the model, d_j, where n is the number of participants in the datasets. Then the individual model weights can be determined.

This method is easy to apply and all models of different forms to be included in the meta-model. Additionally, the discriminative ability of the models can be incorporated into the meta-model, by considering an additional weighting, which may allow a more robust selective screening tool to be devised. It may prevent models with a good discriminative ability but poor calibration being harshly penalised in the meta-model. When previously implemented BMA has outperformed the original model in terms of calibration but also discriminative ability. While the benefit is typically minor it has been shown to be consistent [149]. An independent study concluded BMA improved the model predictive performance and when there is prior information this should be included to further improve the model predictive performance [151]. Finally, BMA was argued to provide a coherent approach for incorporating uncertainty due to variable selection and model form [157].

There are some concerns with BMA. The final meta-model formula can be complex as each model is run independently discouraging pubic application [151]. Bayesian model selection operates under the assumption that one model is the correct model; with this assumption BMA will assign a high weight to one or two leading models and minimise the remaining models. This may not effectively combine the evidence across the models. Additionally, with this approach the calculation is more challenging and "the number of terms can be enormous rendering exhaustive summation infeasible and the integrals implicit can in general be hard to compute" [151]. Finally, BMA does not initially recalibrate the models this could negatively influence more successful models which have a poor calibration in a differing external dataset. Some studies argued that while BMA marginally improved upon the original models there was no leading method to aggregate the models [149].

8.9.4 Univariate Meta-Analysis

The next method is univariate meta-analysis and was published by Debray et al [158]. To apply this method for each model the regression variable coefficients and standard errors are required, although an external dataset is not required. Univariate meta-analysis considers the variable coefficients as sample statistics with sample standard deviations.

The least squares approach is to combine the variable coefficients [158] as follows;

$$w_{ij} = \frac{1}{\sigma_{ij}^2 + \tau_j^2} \tag{8.35}$$

"Where τ_j^2 is the between-study variance of β_j" [158] which is the variable coefficient in the model and sigma the standard error in the $\mathrm{i_{th}}$ model for the $\mathrm{j_{th}}$ variable. These weights are then applied to the model specific variable coefficients assuming a random effects model. Each variable is calculated separately to devise the final meta-model.

The method expects the models to have a large overlap of variables and "that identical model formulations are available for the published prediction models" [158]. This allows the variable coefficients to be combined in the meta-model [158]. As identified in the systematic review (Section 3) distinct prediction models often incorporate a large range of variables, including different measures for the same exposure (smoking history and family history of cancer) and the model form differs. This could limit the practicality of univariate meta-analysis as models may be removed to satisfy these conditions. This may reduce the pool of applicable models that can be combined and unnecessarily exclude successful models. A study found implementation can be difficult when the literature models greatly differ in terms of included variables [150]. Additionally, the method does not utilise evidence in the external datasets. The weighting is based on the strength of the coefficients in the model building datasets. However, these can have unusual dataset designs, such as high or low risk of developing a disease sample population. With this approach, models are not penalised if they are unsuccessful in new populations. When this approach has been used the meta-model did not report an improved performance in comparison to the original models [150].

In summary, while univariate meta-analysis may create a presentable final model there are concerns with this method. The models require the same form and a core set of variables. Models with different forms will be unnecessarily excluded reducing the pool of evidence included in the meta-model. The evidence is further reduced by not using an external dataset to base the meta-model, which may also identify poorer or stronger performing models which should be penalised or enhanced in the final model respectively.

8.9.5 Multivariate Meta-Analysis

The next method presented by Debray et al. [158] is multivariate meta-analysis. This combines multiple models by estimating the with-in prediction model variance to capture correlation between variables within the prediction model, and the between study covariance, to capture the heterogeneity between the studies [158]. To apply this method the models require the same model form with a core set of variables to combine the different model variable coefficients into a single estimate. Unfortunately, it is unlikely a large quantity of prediction models will have the same form with a core set of predictors. Thus, many models may be excluded from the meta-model because they have differing model forms.

The method evaluates the variables to assess if they are highly correlated [158]. Variables which are not highly correlated across the models are minimised in the meta-model [158]. However, if there are a large range of distinct variables then these potentially successful variables are negatively affected in the meta-model.

This method is based on the performance of the models in the original dataset, which omits the evidence available in an external dataset. Although, by considering the between study covariance, the parameters that successfully predict risk in multiple models are rewarded in the meta-model.

Multivariate meta-analysis has previously been implemented. One study found the multivariate meta-model is not without limitations. In particular, there can be difficulty estimating between-study correlations

[159]. As a result, models can be excluded from the meta-model and their evidence is lost in the final meta-model. The study also concluded the meta-model only reported a marginal improvement upon the original models [159].

8.9.6 Bayesian Inference

The final model aggregation method is Bayesian Inference [158]. This method combines multiple models by creating a prior and posterior for each model using an external IPD. When there is sufficient prior evidence, which comes from the prediction models, then there is less necessity to include the external evidence, which forms the posterior [158]. These prior and posterior are then used to assign a weight to each model in the final meta-model.

When using Bayesian Inference, it is important to consider there may be a problem with bias when "literature models are derived using data-driven selection with stepwise methods" [158]. For Bayesian Inference to be successfully applied the models require a set of common variables across the models, which can be pooled into a meta-model [158]. This is not observed in many scenarios, including lung cancer prediction models. Other methods which do not exclude models or variables are preferable as these synthesise all the evidence of the original models in an attempt to make a model with a better predictive ability [151]. This makes the method impractical for most prediction models considered for model aggregation as models are likely to be unnecessarily disregarded if they do not have a universal model form or core set of predictors.

8.10 Summary of Methods to Aggregate Multiple Models

There were six identified methods to aggregate prediction models. The methods vary, some evaluate the model performance in an external IPD while other methods update the variable coefficients using their values and standard errors. There are also differences between updating the coefficients or combining the models by running them simultaneously and weighting the separate results.

BMA and Weighted Averaging assess the performance of the original models in an external validation, which are then weighted based on their calibration. This combines the evidence of the original models, their robustness in a different sample population, and the evidence in the sample population in devising the meta-model. This method is advantageous because different model forms can be combined so no models are unnecessarily excluded and their evidence lost. Additionally, the model discrimination can be considered when weighting the models by assigning a prior to each model during BMA, based on the AUC results. However, the methods commonly assign a higher weight to one or two models and the remaining models are minimised. Also, the final meta-model can be lengthy and complicated to apply by running multiple models simultaneously, this could see a successful prediction model not being considered as a clinical tool. However, creating a useable calculator would address this concern.

Univariate and multivariate meta-analysis aggregate models using the variable coefficients and standard errors in the model building dataset. This does not require an external validation dataset, an advantage if one is not available, but loses the evidence available in this dataset if one is available. By not reviewing the models, this does not reward models that perform robustly in new populations. Additionally, these methods do not consider if a model was devised using an unusual dataset, this can include a high risk dataset with a high exposure to a particular condition. These methods also require a set of core variables and the same model form to be aggregated into a meta-model. Models with distinct forms may be disregarded despite being a successful prediction model.

8.11 Updating and Aggregating Lung Cancer Prediction Models: Applicability and Concerns

The reviewed methods will be applied to selected lung cancer prediction models to evaluate their practicality and success in creating a model with an improved calibration and discriminative ability. Before the methods can be applied the lung cancer prediction models need to be selected and which methods are appropriate for the identified lung cancer models.

The single updating methods will be applied to the $\text{PLCO}_{\text{M2014}}$ Model. This model was selected as it was identified as the leading prediction model in the external validation (Section 6).

The $\text{PLCO}_{\text{M2014}}$, $\text{PLCO}_{\text{M2012}}$, Bach, Hoggart, Pittsburgh, and LLP models are all considered for the model aggregation at this stage. The Spitz and African-American models were excluded from the model aggregation as they do not share an ILCCO dataset with the PLCO models that can be used to perform the model aggregation. The PLCO models were preferred as they performed robustly across a range of datasets whereas the Spitz and African-American models were only evaluated in one dataset.

The concerns that need to be considered before aggregating prediction models are presented to determine which models should be included in the model aggregation.

8.11.1 Model Prediction

The first concern to consider is what the model predicts. Prediction models can predict either incidence or absolute risk. Combining multiple models that predict different outcomes would be problematic. This would combine conflicting risks for an individual and there would be uncertainty into what outcome the meta-model predicts.

The PLCO, LLP, and Pittsburgh models predict lung cancer incidence, whereas the Hoggart and Bach models can be applied to predict either lung cancer incidence or absolute risk because they combine two different equations. By only considering the lung cancer incidence equation of the Bach and Hoggart models then no models need to be removed from the lung cancer model aggregation.

8.11.2 Model Duration

Prediction models predict an outcome developing over a specific duration which can vary between models. Models that predict risk over different durations cannot be combined as the estimated risk for an individual is influenced by the duration the outcome can occur within. Combining estimates over different durations would give a conflicting final risk estimate. Additionally, the final meta-model should predict risk over a defined period, which cannot be determined if this varies across the aggregated model.

The PLCO and Pittsburgh models predict risk over a fixed 6-year duration. The Hoggart and Bach models predict risk for 1 year, but can be run recursively to estimate 6-year risk. However, the LLP Model is fixed to predict lung cancer risk over 5 years. Therefore, this cannot be combined with the other models and thus excludes the LLP Model from the lung cancer model aggregation.

8.11.3 Target Population

The next consideration is the target population for the prediction models. Models that predict risk for distinct target populations cannot be combined. The aggregated models require a specific target population that is applicable to all the original models.

For lung cancer prediction models the target populations are restricted by smoking status, age, and a minimum smoking history. The six prediction models have a different target population, as presented in table 8.1.

	No Age	20+	35+	40-80	50-75	
Everyone		$PLCO_{M2014}$		LLP		No Smoking Restrictions
Ever-Smokers	Pittsburgh	$PLCO_{M2012}$	Hoggart			No Smoking Restrictions
Ever-Smokers					Bach	30+ PY

Table 8.1: Lung Cancer Prediction Models Target Populations

Combining the models will necessitate that they are restricted to a population that is applicable to all the models considered in the aggregation. The Bach Model is the most restrictive model and if this is considered in the aggregation then the models would be confined to predicting risk in $50-75$ year olds with a minimum 30+ pack year smoking history. Excluding this model, and the LLP Model already excluded, would allow the aggregated model to be applied to ever-smokers aged at least 35 years. While this would still not screen never-smokers, this allows all ever-smokers likely to be considered in a selective screening programme to be evaluated in the meta-model.

It may be beneficial to run two versions of the model aggregation for lung cancer prediction models, one with the Bach Model included and a second version excluding the Bach Model allowing a larger range of participants to be evaluated.

8.11.4 Model Form

The same model design is important for some of the model aggregation techniques. Unfortunately, the same model form is not always present for distinct models created by different authors. Distinct model forms will limit the available methods to aggregate prediction models as variables across the models cannot be combined into a single variable as desired. This may be a result of considering specific incidence scaling rates, considering splines for a variable, or different degree polynomials for the models.

The lung cancer prediction models have distinct model forms. The PLCO models consider a logistic regression equation centered on the mean for each variable, the Pittsburgh Model considers an additive logistic regression, the LLP Model considers age-gender lung cancer specific incidence rates from the Liverpool, UK area, the Bach Model is a cubic spline model, and the Hoggart Model considers Weibull incidence and death rates specific to smoking status and smoking duration in a stratified survival model. These models cannot be combined into a meta-model with a single coefficient for each variable as desired by some model aggregation methods.

The lung cancer prediction models cannot be combined using the stack regressions, univariate and multivariate meta-analysis, or Bayesian inference approaches. However, the literature review identified additionally methods which do not require a similar model form across the models. These methods, model averaging and BMA, allow each method to be run separately and then the estimates combined by assigning a weight to each model. These methods can still be considered for model aggregation.

8.11.5 Model Variables

To aggregate the models, some methods require a set of core variables that are present in all original models.

For the lung cancer prediction models there is not a core set of variables. There are many different variables considered between the models as there are different measures to quantify a similar condition in participants. For example, smoking history is measured by CPD, pack years, smoking duration, or quit duration and family history of cancer is measure by early or late onset, specifically lung cancer or smoking related cancer, or any type of cancer. This excludes stack regressions, univariate and multivariate meta-analysis, and Bayesian inference but these were rejected previously. Model averaging and BMA are still acceptable methods as these do not require a core set of variables to be present in the models.

8.11.6 Model Co-Linearity

Co-linearity can be problematic in model aggregation. The final meta-model would bias towards one of the models and minimise the other model [148]. This can be avoided by only considering one of any of the similar models.

There was a concern with model aggregation for lung cancer prediction models where the $PLCO_{M2012}$ and $PLCO_{M2014}$ models are similar. The 2014 model was a recalibrated and extended version of the 2012 model. The $PLCO_{M2012}$ Model will not be considered in the model aggregation in preference of the 2014 version. Although the meta-model can only be applied to an ever-smokers population and the 2012 version is applicable to ever-smokers, the 2014 version was preferred as this had the universal leading performance including in the UKLS target population.

8.12 Summary

This chapter identified methods for model updating and model aggregation in a literature review. A range of different methods were identified for single updating. These include model recalibrated, re-estimation of the model parameters, and extending the model to include additional parameters. These methods all use an external dataset to perform the model updating, provided by the ILCCO datasets. The methods will be applied to the $PLCO_{M2014}$ Model, which was selected for single model updating because it was the leading model in the external validations.

The literature review also identified six different methods for model aggregation. The methods varied including methods which combine models by assigning them all a weight based on their performance in an external dataset, combining the evidence of the original models and the external dataset in an attempt to devise a more robust model. Other methods combine the variable coefficients that commonly occur across the models, requiring the same model form and common variables across the models considered in the aggregation.

The model aggregation methods could be applied to the Bach, $PLCO_{M2014}$, Pittsburgh, and Hoggart models. The identified methods that were appropriate for the lung cancer prediction models are Model Averaging and BMA (with and without an additional weighting based on the AUC).

The literature review evaluated methods to update and aggregate prediction models. Appropriate methods to update lung cancer prediction models were identified which will now be applied and the new models validated to assess if they create an improved prediction model.

CHAPTER 9

Updating and Aggregating Lung Cancer Models: Methodology and Dataset Analysis

9.1 Introduction

The literature review identified methods to update and aggregate prediction models. These will be applied to lung cancer prediction models to review their practicality and their ability to devise a model with an improved calibration, discriminative ability, and potential as a selective screening tool. This chapter will present the IPD that will be used to update the models and the external validation dataset to validate the newly devised models. The methodology of how the models will be validated will be presented, and a validation of the original models will be conducted which will provide a baseline performance for the new models to improve upon.

9.2 Objectives

This chapter will perform a dataset analysis and present the guidelines to review the original models and the new updated models. Next it will conduct an external validation of the original models that will be subsequently updated. To achieve this the chapter will;

1. Present a methodology to review the model calibration, discrimination, and prediction rules for the original models and the new updated models to identify a leading model.

2. Perform a dataset analysis on the IPD and the external validation dataset to identify any concerns that could influence the model updating or validation results.

3. Conduct a validation of the original prediction models, as per the methodology, to provide a baseline performance to compare with the new models in the subsequent chapters.

9.3 Methodology

The literature review identified methods to update a single prediction model and aggregate multiple prediction models, using an external dataset. Single model updating will be conducted on the $\text{PLCO}_{\text{M2014}}$ Model as it had the leading performance in the external validation, a good basis on which to attempt to devise a leading prediction model for lung cancer.

For the identified model aggregation methods two versions of each method will be conducted. The first version will aggregate the $\text{PLCO}_{\text{M2014}}$, Bach, Hoggart, and Pittsburgh models and the second version will exclude the Bach Model. Exclusing the Bach Model will allow the meta-model to be applicable to a larger population.

To apply the model updating methods an external IPD was created from the ILCCO datasets. The $PLCO_{M2014}$ Model was applicable to five datasets; these are the ReSoLuCENT, UCLA, CARET, MSH-PMH, and the NY Wynder datasets. The ReSoLuCENT, UCLA, CARET and MSH-PMH datasets were combined into one dataset to conduct the model updating methods. These four datasets were chosen because it pools two sample populations that would be reflective of a target population for lung cancer screening and both a sample population that would be considered lower and higher risk of developing lung cancer. The NY Wynder dataset will provide an external validation dataset to evaluate the updated models. This approach was preferred rather than combining 80% of each dataset into a model updating dataset and the remaining 20% forming the validation dataset, as this allowed the external validation dataset to be distinct from the model updating datasets. This avoids the potential for optimistic results [46]. In the single updating datasets anyone aged over 20 years can be included because this is the only restriction when using the $PLCO_{M2014}$ Model.

The same datasets are considered for the model aggregation methods. The datasets will be reduced in size because the models have a more restrictive target population. By excluding the Bach Model from the model aggregation the dataset can include all ever-smokers aged at least 35 years. When including the Bach Model this is further restricted to ever-smokers aged 50-75 years with a minimum 30 pack year smoking history. Additionally, the Bach Model requires asbestos exposure, which is not reported in the UCLA dataset therefore this dataset is excluded and the remaining three datasets (ReSoLuCENT, CARET, MSH-PMH) will form the model aggregation dataset, with the NY Wynder dataset remaining as the external validation dataset.

When validating the updated models, it is important to remember these will be recalibrated to reflect a high lung cancer incidence rate in the dataset. Therefore, evaluating the models at the 0.1, 0.25, 0.5, 1, 1.5, and 2.5% risk thresholds may not be appropriate. Instead, the optimal risk threshold for each model will be identified in an internal validation using the model updating and aggregation dataset. The model will then be evaluated at the identified risk threshold in the external validation to assess if a consistent performance is identified. The second validation will assess the models while maintaining a specificity of 90%. Here the models will be reviewed in a reduced sample population of anyone aged between 50-75 years in the dataset, as per the UKLS screening trial guidelines.

9.4 Dataset Analysis

The three differing model updating datasets and differing NY Wynder datasets (caused by model restrictions) which are applicable for the separate model updating techniques are presented.

The ReSoLuCENT and MSH-PMH datasets are included in the model updating datasets. Missing information in these datasets was previously imputed, as presented in Sections 4.10 & 4.11. The missing information was imputed in 11 separate iterations. The average for the imputed value will be considered in the final dataset.

Single model updating will be conducted using a combination of case-control datasets which resulted in a high proportion of individuals with lung cancer in the final dataset (41.7%). Since model updating recalibrates the model to reflect the observed incidence rate in the dataset the models may be recalibrated to predict a lung cancer incidence rate that would be considered excessive for a real world population. The calibration results may also be misleading in the external validation because the NY Wynder dataset reports a high lung cancer incidence rate. Therefore, it is expected the recalibrated updated models will be more suited to the NY Wynder dataset in comparison to the original $PLCO_{M2014}$ Model.

A review of the population demographic for the model updating and NY Wynder datasets (Table 9.1) demonstrated a sample population similar to what which would be expected if a lung cancer screening target population has been obtained. The ages of the participants are centered around 60 years, although the disease free individuals in the model updating dataset are slightly lower at 55 years. There is a mixture of ethnicities and education levels; however, the NY Wynder dataset only reported 5 education levels. A high proportion of the individuals with lung cancer are ever-smokers and have a heavy smoking history.

Variable		Single Updating					
		IPD			NY Wynder		
		Case	Control	P-value	Case	Control	P-Value
	Total	2,903 (41.72%)	4,056 (58.28%)		4,783 (51.59%)	4,488 (48.41%)	
Gender	Male	1,573 (54.19%)	2,264 (55.82%)	0.177	2,915 (60.95%)	2,565 (57.15%)	0.000
	Female	1,330 (45.81%)	1,792 (44.18%)		1,868 (39.05%)	1,923 (42.85%)	
	Age (s.e.)	59.14 (0.171)	55.99 (0.162)	0.000	60.56 (0.142)	60.28 (0.151)	0.173
	BMI (s.e.)	26.13 (0.098)	26.95 (0.082)	0.000	25.28 (0.062)	26.35 (0.073)	0.000
Ethnicity	White	2,442 (84.12%)	3,477 (85.72%)	0.002	4,236 (88.56%)	4,051 (90.26%)	0.008
	Black	139 (4.79%)	163 (4.02%)		471 (9.85%)	376 (8.38%)	
	Hispanic	84 (2.89%)	226 (5.57%)		57 (1.19%)	53 (1.18%)	
	Asian	222 (7.65%)	177 (4.36%)		19 (0.40%)	8 (0.18%)	
	American-Indian	16 (0.55%)	13 (0.32%)		0 (0%)	0 (0%)	
	Hawaiian	0 (0%)	0 (0%)		0 (0%)	0 (0%)	
Education	1	240 (8.27%)	92 (2.27%)	0.000	5 (0.10%)	1 (0.02%)	0.000
	2	354 (12.19%)	233 (5.74%)		678 (14.18%)	432 (9.63%)	
	3	396 (13.64%)	311 (7.67%)		3,045 (63.66%)	2,634 (58.69%)	
	4	598 (20.60%)	709 (17.48%)		0 (0%)	0 (0%)	
	5	640 (22.05%)	1,041 (25.67%)		622 (13.00%)	759 (16.91%)	
	6	675 (23.25%)	1,670 (41.17%)		433 (9.05%)	662 (14.75%)	
Smoking Status	Never	366 (12.61%)	1,352 (33.33%)		257 (5.37%)	1,782 (39.71%)	
	Former	1,020 (35.14%)	1,145 (28.23%)		1,614 (33.74%)	1,617 (36.03%)	
	Current	1,517 (52.26%)	1,559 (38.44%)		2,912 (60.88%)	1,089 (24.26%)	
CPD	Former (s.e.)	21.18 (0.334)	17.73 (0.360)	0.000	27.19 (0.346)	22.18 (0.350)	0.000
	Current (s.e.)	23.33 (0.246)	22.11 (0.248)	0.001	26.76 (0.215)	21.83 (0.340)	0.000
Duration	Former (σ)	32.83 (0.365)	22.66 (0.384)	0.000	33.32(0.287)	24.55 (0.314)	0.000
	Current (s.e.)	41.03 (0.213)	37.64 (0.249)	0.000	40.066 (0.194)	36.54 (0.365)	0.000
Quit	Duration (s.e.)	10.92 (0.353)	17.11 (0.378)	0.000	12.48 (0.234)	18.04 (0.282)	0.000
Family History of LC	0	2,458 (84.67%)	3,478 (85.75%)	0.109	4,587 (95.90%)	4,357 (97.08%)	0.001
	1	379 (13.06%)	513 (12.65%)		175 (3.66%)	120 (2.67%)	
	2+	64 (2.27%)	65 (1.60%)		21 (0.44%)	11 (0.25%)	
Prior Tumour	No	2,667 (91.90%)	3,514 (86.64%)	0.000	4,635 (96.91%)	4,297 (95.74%)	0.003
	Yes	235 (8.10%)	542 (13.36%)		148 (3.09%)	191 (4.26%)	
COPD	No	2,437 (83.95%)	3,733 (92.04%)	0.000	4,298 (89.86%)	4,300 (95.81%)	0.000
	Yes	466 (16.05%)	323 (7.96%)		485 (10.14%)	188 (4.19%)	
Asbestos	No	2,067 (89.64%)	2,822 (91.68%)	0.010	4,773 (99.79%)	4,483 (99.89%)	0.242
	Yes	239 (10.36%)	256 (8.32%)		10 (0.21%)	5 (0.11%)	

Table 9.1: Population Demographic for Single Model Updating

Variable		Model Aggregation without the Bach Model					
		IPD			NY Wynder		
		Case	Control	P-value	Case	Control	P-Value
	Total	2,532 (48.85%)	2,651 (51.15%)		4,505 (62.67%)	2,683 (37.33%)	
Gender	Male	1,450 (57.27%)	1,640 (61.86%)	0.001	2,810 (62.38%)	1,816 (67.69%)	0.000
	Female	1,082 (42.73%)	1,011 (38.14%)		1,695 (37.62%)	867 (32.31%)	
	Age (s.e.)	59.47 (0.174)	57.94 (0.161)	0.000	60.67 (0.141)	60.28 (0.180)	0.086
	BMI (s.e.)	26.18 (0.104)	27.25 (0.098)	0.000	25.27 (0.064)	26.27 (0.092)	0.000
Ethnicity	White	2,241 (88.51%)	2,350 (88.65%)	0.325	3,991 (88.59%)	2,408 (89.75%)	0.115
	Black	117 (4.62%)	106 (4.00%)		452 (10.03%)	242 (9.02%)	
	Hispanic	50 (1.97%)	120 (4.53%)		50 (1.11%)	30 (1.12%)	
	Asian	109 (4.30%)	65 (2.45%)		12 (0.27%)	3 (0.11%)	
	American-Indian	15 (0.59%)	10 (0.38%)		0 (0%)	0 (0%)	
	Hawaiian	0 (0%)	0 (0%)		0 (0%)	0 (0%)	
Education	1	232 (9.16%)	67 (2.53%)	0.000	5 (0.11%)	0 (0%)	0.000
	2	295 (11.65%)	145 (5.47%)		655 (14.54%)	278 (10.36%)	
	3	354 (13.98%)	219 (8.26%)		2,881 (63.95%)	1,616 (60.23%)	
	4	539 (21.29%)	566 (21.35%)		0 (0%)	0 (0%)	
	5	578 (22.83%)	786 (29.65%)		581 (12.90%)	463 (17.26%)	
	6	534 (21.09%)	868 (32.74%)		383 (8.50%)	326 (12.15%)	
Smoking Status	Never	NA	NA		NA	NA	
	Former	1,020 (40.28%)	1,124 (42.40%)		1,610 (35.74%)	1,609 (59.97%)	
	Current	1,512 (59.72%)	1,527 (57.60%)		2,895 (64.26%)	1,074 (40.03%)	
CPD	Former (s.e.)	21.18 (0.334)	17.89 (0.364)	0.000	27.20 (0.346)	22.18 (0.352)	0.000
	Current (s.e.)	23.31 (0.245)	22.30 (0.248)	0.004	26.78 (0.216)	21.89 (.343)	0.000
Duration	Former (s.e.)	32.83 (0.365)	22.88 (0.388)	0.000	33.38 (0.287)	24.63 (0.314)	0.000
	Current (s.e.)	41.11 (0.211)	38.20 (0.233)	0.000	40.20 (0.192)	36.86 (0.360)	0.000
Quit	Duration (s.e.)	10.92 (0.353)	17.36 (0.381)	0.000	12.50 (0.234)	18.10 (0.283)	0.000
Family History of LC	0	2,122 (83.81%)	2,274 (85.78%)	0.056	4,318 (95.85%)	2,604 (97.06%)	0.004
	1	345 (13.63%)	324 (12.22%)		167 (3.71%)	74 (2.76%)	
	2+	65 (2.57%)	53 (2.01%)		20 (0.44%)	5 (0.19%)	
Prior Tumour	No	2,317 (91.54%)	2,325 (87.70%)	0.000	4,364 (96.87%)	2,587 (96.42%)	0.303
	Yes	214 (8.46%)	326 (12.30%)		141 (3.13%)	96 (3.58%)	
COPD	No	2,095 (82.74%)	2,402 (90.61%)	0.000	4,031 (89.48%)	2,545 (94.86%)	0.000
	Yes	437 (17.26%)	249 (9.39%)		474 (10.52%)	2,545 (94.86%)	
Asbestos	No	NA	NA		NA	NA	
	Yes	NA	NA		NA	NA	

Table 9.2: Population Demographic for Model Aggregation without the Bach Model

Variable		Model Aggregation with Bach Model					
		IPD			NY Wynder		
		Case	Control	P-value	Case	Control	P-Value
	Total	1,350 (49.65%)	1,369 (50.35%)		3,028 (72.42%)	1,153 (27.58%)	
Gender	Male	856 (63.41%)	937 (68.44%)	0.006	2,002 (66.12%)	868 (75.28%)	0.000
	Female	494 (36.59%)	432 (31.56%)		1,026 (33.88%)	285 (24.72%)	
	Age (s.e.)	61.01 (0.163)	59.91 (0.149)	0.000	62.41 (0.120)	62.58 (0.189)	0.443
	BMI (s.e.)	26.57 (0.135)	27.50 (0.137)	0.000	25.48 (0.078)	26.54 (0.148)	0.000
Ethnicity	White	1,295 (95.93%)	1,311 (95.76%)	0.686	2,723 (89.93%)	1,063 (92.19%)	0.014
	Black	16 (1.19%)	24 (1.75%)		266 (8.78%)	81 (7.03%)	
	Hispanic	6 (0.44%)	8 (0.58%)		31 (1.02%)	9 (0.78%)	
	Asian	25 (1.85%)	18 (1.31%)		8 (0.26%)	0 (0%)	
	American-Indian	8 (0.59%)	8 (0.58%)		0 (0%)	0 (0%)	
	Hawaiian	0 (0%)	0 (0%)		0 (0%)	0 (0%)	
Education	1	144 (10.67%)	32 (2.34%)	0.000	1 (0.03%)	0 (0%)	0.001
	2	129 (9.56%)	43 (3.14%)		490 (16.18%)	143 (12.40%)	
	3	232 (17.19%)	145 (10.59%)		1,950 (64.40%)	741 (64.27%)	
	4	247 (18.30%)	332 (24.25%)		0 (0%)	0 (0%)	
	5	289 (21.41%)	477 (34.84%)		361 (11.92%)	172 (14.92%)	
	6	309 (22.89%)	340 (24.84%)		226 (7.46%)	97 (8.41%)	
Smoking Status	Never	NA	NA		NA	NA	
	Former	449 (33.26%)	315 (23.01%)		962 (31.77%)	576 (49.96%)	
	Current	901 (66.74%)	1,054 (76.99%)		2,066 (68.23%)	577 (50.04%)	
CPD	Former (s.e.)	26.60 (0.454)	28.32 (0.617)	0.022	32.58 (0.414)	33.10 (0.546)	0.442
	Current (s.e.)	25.69 (0.279)	25.60 (0.253)	0.805	28.46 (0.239)	26.49 (0.416)	0.000
Duration	Former (s.e.)	39.05 (0.326)	41.21 (0.194)	0.000	37.61 (0.259)	33.77 (0.379)	0.000
	Current (s.e.)	43.42 (0.208)	36.50 (0.383)	0.000	43.57 (0.168)	42.37 (0.316)	0.001
Quit	Duration (s.e.)	6.50 (0.313)	6.97 (0.352)	0.329	9.70 (0.214)	12.55 (0.346)	0.000
Family History of LC	0	1,142 (84.59%)	1,192 (87.07%)	0.032	2,904 (95.90%)	1,121 (97.22%)	0.038
	1	176 (13.04%)	155 (11.32%)		109 (3.60%)	29 (2.52%)	
	2+	32 (2.37%)	22 (1.60%)		15 (0.50%)	3 (0.26%)	
Prior Tumour	No	1,231 (91.19%)	1,228 (89.70%)	0.188	2,947 (97.32%)	1,115 (96.70%)	0.281
	Yes	119 (8.81%)	141 (10.30%)		81 (2.68%)	38 (3.30%)	
COPD	No	1,089 (80.67%)	1,203 (87.87%)	0.000	2,677 (88.41%)	1,067 (92.54%)	0.000
	Yes	261 (19.33%)	166 (12.13%)		351 (11.59%)	86 (7.46%)	
Asbestos	No	1,184 (87.70%)	1,191 (87.00%)	0.580	3,022 (99.80%)	1,152 (99.91%)	0.431
	Yes	166 (12.30%)	178 (13.00%)		6 (0.20%)	1 (0.09%)	

Table 9.3: Population Demographic for Model Aggregation with the Bach Model

Asbestos exposure will be considered as an additional variable for model extension. In the model updating dataset, asbestos exposure is significantly associated with an elevated lung cancer risk. However, in the NY Wynder dataset asbestos exposure is not significantly associated with an increased lung cancer risk. Therefore, while it may be included as an additional parameter in the prediction model this may not improve the model discrimination in the external validation.

The model aggregation, excluding the Bach Model, included ever-smokers aged at least 35 years. This removed a higher proportion of disease free individuals. In the final dataset around 50% of individuals had lung cancer (Table 9.2) so the models are likely to be poorly calibrated. Since the model aggregation is based upon the model calibration this is a limitation of the datasets obtained. There were no additional concerns with the model aggregation or NY Wynder datasets.

Finally, including the Bach Model in the model aggregation, resulted in smaller aggregation and external validation datasets, as these were restricted to ever-smokers aged between 50-75 years with a minimum 30 pack year smoking history. There are similar concerns in this model aggregation datasets as previously, with a high ratio of individuals with lung cancer to disease free individuals (Table 9.3), which is likely to result in poor model calibration on which to base the model aggregation. This is exacerbated in the NY Wynder dataset where 72% of participants reported lung cancer incidence.

In summary, the main concern with the datasets is the high proportion of diseased to disease free individuals. The original models are likely to be poorly calibrated. This may result in extensive model updating being required, and a tendency to extend the model to include all available variables as these can significantly improve the model goodness of fit. Finally, the recalibrated models may generate unrealistic risk estimates based on the high incidence rate in the model updating dataset, and may therefore not be appropriate in real world populations with substantially lower incidence rates. While, the discriminative ability and prediction rules will remain unaffected and an improved model can be created, the selected risk thresholds are likely to be higher as the models are recalibrated to estimate higher risks.

9.5 Validation of the Original Models

The original models are validated to provide a baseline performance for the updated and meta-models to be compared. The model updating datasets and the NY Wynder dataset are both external datasets for the original models. In contrast the newly updated models will be internally validated in the model updating dataset. Therefore, it is expected that the updated models will have an improved performance when internally validated. Therefore, the NY Wynder dataset will offer a more rigorous assessment of the updated models to assess if a more robust model has been devised.

9.5.1 Single Model Updating

There were concerns the high lung cancer incidence rates would hinder the $\text{PLCO}_{\text{M2014}}$ Model calibration. Indeed, the Hosmer-Lemeshow test in the IPD returned a p-value below 0.001 and a χ^2 exceeding 1,250. The $\text{PLCO}_{\text{M2014}}$ Model also reported a poor calibration in the NY Wynder dataset. The poor calibration demonstrated through the Hosmer-Lemeshow test is supported by the Brier Score which exceeded 0.4.

The original model reported a reasonable discrimination in the model updating dataset, with an AUC of 0.68, although this decreased to 0.66 when restricted to a sample population of participants aged 50-75 years. The model was optimal at the 0.316% risk threshold; here, the model reported a sensitivity of 78.33% and a specificity of 50%. The high sensitivity and specificity resulted in a Youden index just below 0.3 and a PLR of 1.58. When applying the model to the UKLS target population the prediction rule performance, similar to the AUC, was lowered. The $\text{PLCO}_{\text{M2014}}$ Model should be applied at the 3.16% risk threshold to maintain a specificity of 90%. In the IPD the sensitivity dropped below 20%, a poor capture rate highlighted by a Youden index below 0.1 and a PLR of only 1.96 despite the high specificity.

Validation	Internal/External	Hos-Leme.	Brier Score	AUC [95% CI]	Threshold (%)	Sens.	Spec.	Youden	PLR
1	Internal	0 (1266.26)	0.4007	0.6813 [0.669, 0.694]	0.32	78.33	50.32	0.2865	1.5768
	External	0 (3428.03)	0.4887	0.7667 [0.757, 0.776]	0.32	86.85	54.68	0.4153	1.9163
2	Internal	0 (726.84)	0.4173	0.6567 [0.642, 0.671]	3.19	19.59	90.00	0.0960	1.9601
	External	0 (9793.05)	0.4906	0.7834 [0.773, 0.794]	2.95	35.02	89.99	0.2502	3.5002
Validation 1 at the optimal risk threshold					Validation 2 using the UKLS guidelines				

Table 9.4: External Validation of $PLCO_{M2014}$ Model in the Single Updating Datasets

The $PLCO_{M2014}$ Model performed better in the NY Wynder dataset. The AUC was 0.767 and reported an encouraging prediction rules performance. At the optimal risk threshold, identified in the model updating dataset as 0.316%, the model reported an increased sensitivity of 86.85% and specificity of 54.68%. Therefore the Youden index was above 0.4. In the target UKLS population the performance improved further. The AUC increased to 0.783 and at a risk threshold of 2.95% the sensitivity was 35% for a specificity of 90%.

Overall, the model was poorly calibrated as was expected. It is likely that the recalibrated models will report a higher performance because they will predict the higher incidence rate. Additionally, the $PLCO_{M2014}$ Model had a limited discriminative ability in the model updating dataset. However, the AUC and prediction rules performance improved in the NY Wynder dataset such that the new models will be required to produce an exceptional performance to demonstrate an improved model.

9.5.2 Model Aggregation without the Bach Model

The original models reported a poor performance in the model aggregation dataset. The high incidence rates resulted in all the models reporting a poor calibration measured by the Hosmer-Lemeshow tests and supported by a Brier score which was higher than 0.4 for all the models.

The models also reported a low discriminative ability in the aggregation dataset with all the models reporting an AUC of approximately 0.6. This was reflected in poor prediction rules results where none of the models achieved a Youden index exceeding 0.2 and the PLR was low with most results below 1.5. This poor performance was also observed in the UKLS target population.

Model	Validation	Internal/External	Hosmer-Lemeshow	Brier Score	AUC [95% CI]	Threshold (%)	Sens.	Spec.	Youden	PLR
$PLCO_{2014}$	1	Internal	0 (687.51)	0.4666	0.6222 [0.607, 0.637]	0.82	72.00	45.34	0.1734	1.3172
		External	0 (1059.44)	0.5917	0.6806 [0.668, 0.693]	0.82	75.58	49.57	0.2515	1.4988
	2	Internal	0 (563.05)	0.4623	0.6148 [0.598, 0.631]	3.62	17.85	89.99	0.0784	1.7829
		External	0 (1292.11)	0.5891	0.7016 [0.688, 0.716]	3.94	24.99	89.98	0.1498	2.4948
Hoggart	1	Internal	0 (948.39)	0.3874	0.6040 [0.589, 0.619]	2.24	81.20	37.19	0.1839	1.2929
		External	0 (748.86)	0.4801	0.6778 [0.665, 0.691]	2.24	81.42	42.86	0.2428	1.4250
	2	Internal	0 (795.28)	0.3792	0.5907 [0.574, 0.607]	29.05	12.58	90.03	0.0261	1.2615
		External	0 (589.00)	0.4678	0.6819 [0.668, 0.696]	24.84	25.80	89.94	0.1574	2.5642
Pittsburgh	1	Internal	0 (744.99)	0.4638	0.5911 [0.576, 0.607]	0.72	83.14	33.69	0.1682	1.2537
		External	0 (370.90)	0.5887	0.6793 [0.667, 0.692]	0.72	81.51	43.91	0.2542	1.4531
	2	Internal	0 (672.30)	0.4579	0.5751 [0.558, 0.592]	5.12	14.93	89.33	0.0426	1.3988
		External	0 (292.67)	0.5842	0.6951 [0.681, 0.709]	4.66	24.13	89.98	0.1411	2.4085
Validation 1 at the optimal risk threshold					Validation 2 using the UKLS guidelines					

Table 9.5: External Validation of Models in the Datasets without the Bach Model

The performance improved in the NY Wynder dataset although the models were still poorly calibrated. The AUC increased to between 0.67-0.70 for the models and this improvement was observed in the prediction rules performance. At their optimal risk threshold, the models reported a Youden index of 0.25. The models reported a sensitivity of approximately 25% when the specificity was fixed at 90% in the UKLS target population.

In summary, the models were poorly calibrated and reported a similar discriminative ability. This is likely to reduce the difference in weightings between the BMA methods because the additional weighting

will not allow a drastic change from the non-informative prior. Additionally, basing the aggregation on the calibration may be unreliable since the models were poorly calibrated, particularly for methods that do not originally recalibrate the model. In the NY Wynder datasets the models will aim to eclipse an AUC of 0.70, a Youden index of 0.25 at the optimal risk threshold, and a sensitivity of 25% using the UKLS criteria.

9.5.3 Model Aggregation with the Bach Model

In the final validation the models were poorly calibrated in both datasets as a result of the high lung cancer incidence rate. The poor calibration is supported by the Brier score, which is above 0.4 for all the models.

The models also had a poor discriminative ability in the aggregation dataset. The $\text{PLCO}_{\text{M2014}}$ Model marginally reported the highest AUC with a result of 0.59, highlighting the low discriminative ability standard of the models in the dataset. Further the Bach Model reported an AUC below 0.5.

Model	Validation	Internal/External	Hosmer-Lemeshow	Brier Score	AUC [95% CI]	Threshold (%)	Sens.	Spec.	Youden	PLR
PLCO2014	1	Internal	0.1283 (296.53)	0.4682	0.589 [0.568, 0.611]	2.26	51.85	62.24	0.1409	1.3731
		External	0.0890 (455.37)	0.6757	0.597 [0.578, 0.616]	2.26	57.69	55.59	0.1329	1.2993
	2	Internal	0.1283 (296.53)	0.4682	0.589 [0.568, 0.611]	4.42	18.96	89.99	0.0896	1.8949
		External	0.0890 (455.37)	0.6757	0.597 [0.578, 0.616]	5.40	18.43	90.03	0.0845	1.8476
Hoggart	1	Internal	0 (460.43)	0.3725	0.529 [0.508, 0.551]	24.96	29.70	76.33	0.0604	1.2551
		External	0.0001 (251.50)	0.5146	0.609 [0.590, 0.628]	25.01	31.14	81.01	0.1215	1.6396
	2	Internal	0 (460.43)	0.3725	0.529 [0.508, 0.551]	32.29	12.37	90.07	0.0244	1.2452
		External	0.0001 (251.50)	0.5146	0.609 [0.590, 0.628]	31.25	18.76	90.03	0.0878	1.8807
Pittsburgh	1	Internal	0 (550.24)	0.4631	0.526 [0.504, 0.547]	4.66	26.37	82.18	0.0855	1.4796
		External	0.0001 (251.50)	0.6696	0.611 [0.592, 0.630]	4.66	20.94	88.81	0.0975	1.8714
	2	Internal	0 (550.24)	0.4631	0.526 [0.504, 0.547]	5.63	10.81	94.30	0.0512	1.8981
		External	0.0001 (251.50)	0.6696	0.611 [0.592, 0.630]	5.63	15.62	91.85	0.0747	1.9161
Bach	1	Internal	0.1529 (293.8)	0.4577	0.554 [0.533, 0.576]	2.12	76.30	32.36	0.0866	1.1280
		External	0.0001 (251.50)	0.6614	0.601 [0.582, 0.620]	2.12	76.39	36.51	0.1290	1.2032
	2	Internal	0.1529 (293.8)	0.4577	0.554 [0.533, 0.576]	6.87	13.56	89.99	0.0355	1.3546
		External	0.0001 (251.50)	0.6614	0.601 [0.582, 0.620]	7.01	18.53	90.03	0.0855	1.8575
Validation 1 at the optimal risk threshold						Validation 2 using the UKLS guidelines				

Table 9.6: External Validation of Models in the Datasets with the Bach Model

The models improved in the NY Wynder dataset with no leading model being identified because they all reported an AUC of approximately 0.66. The prediction rules were similar for the models also with a Youden index of 0.12 at the optimal risk threshold. Evaluating the models using the UKLS guidelines a sensitivity around 18% was reported for all models. This is the baseline performance upon which the meta-model will attempt to improve upon.

There are concerns the poor calibration may hinder the aggregation methods based on this evidence, and the poor AUC across all the models may not provide sufficient information for the additional AUC weighting for BMA.

9.6 Summary

The chapter presented the methodology to review both the original models and the new models. A dataset analysis of the model updating and the NY Wynder datasets was conducted. The original models were then validated. The calibration was poor in the datasets due to the high lung cancer incidence rates observed. This was identified as a concern for single models updating, as the new models will be recalibrated to reflect the high lung cancer incidence rate. The proposed models then may be unrealistic when applied in a population with a more realistic lung cancer incidence rate. Additionally, the calibration information may be unreliable on which to base the model aggregation.

The $\text{PLCO}_{\text{M2014}}$ Model will be updated using single model updating techniques. In the external validation this model reported an AUC of 0.767, a Youden index of 0.4 at the optimal risk threshold, and a

sensitivity of 35% using the UKLS guidelines, for which the updated models will attempt to surpass. For model aggregation excluding the Bach Model the baseline performance in the external validation which the meta-models should aim to surpass was an AUC of 0.70, a Youden index of 0.25 at the optimal risk threshold, and a sensitivity of 25% for the UKLS guidelines. Finally, when considering the Bach Model, the proposed meta-models will attempt to eclipse an AUC of 0.66, a Youden index of 0.12 at the optimal risk threshold, and a sensitivity of 18% in the UKLS target population in the NY Wynder dataset.

The meta-model including the Bach Model may allow a better evaluation of BMA with an additional weighting based on the AUC as there was more disparity in the AUC results across the models.

The next stage of the project will apply the methods for the single model updating and the model aggregation to devise new models. These will then be validated and compared with these original model results, to assess if an improved lung cancer prediction model has been devised.

CHAPTER 10

Updating a Single Lung Cancer Prediction Model

10.1 Introduction

The literature review identified practical methods to update a single prediction model. These methods will be applied to the $PLCO_{M2014}$ Model, which was selected because the model demonstrated the leading performance in the external validations. The new updated models will be presented and externally validated to assess if the methods have created an improved model that could be considered as a clinical utility. This will also identify successful methods which could be used for prediction models of different diseases or prediction models that consider clinical markers once these models demonstrate an improvement over epidemiological prediction models.

10.2 Objectives

The chapter will apply updating methods to the $PLCO_{M2014}$ Model in an attempt to create an improved lung cancer prediction model. The chapter objectives are;

1. Apply the identified methods from the literature review to update a single prediction model to the $PLCO_{M2014}$ Model.
2. Present the updated model coefficients.
3. Externally validate the model in comparison to the original model to assess if an improved lung cancer prediction model was devised.
4. Present the Stata code and screening guidelines for any model shown to have an improved performance.
5. Review when methods are appropriate to update a prediction model.

10.3 $PLCO_{M2014}$ Model Formula

The $PLCO_{M2014}$ Model is a log-odds model [82]. The methods update the logistic regression equation in the model. To update the model the risk generated by the models needs to be converted so the logistic regression equation in the log-odds model is explicitly defined;

$$Risk = \frac{e^{LR}}{1+e^{LR}} \tag{10.1}$$

$$e^{LR} = \frac{Risk}{1-Risk} \tag{10.2}$$

$$LR = \log\left(\frac{Risk}{1-Risk}\right) \tag{10.3}$$

Where LR = Linear Regression Equation of the log odds model

Presented below is the original linear regression equation that will be recalibrated, re-estimated, and extended;

$$\begin{aligned} LogisticRegressionEquation = {} & -7.02198 + [0.079597 \times (Age - 62)] - [0.0879289 \times (Education - 4)] \\ & - [0.028948 \times (BMI - 27)] + (0.3457265 \times COPD) + (0.3211605 \times Black) \\ & - (0.8203332 \times Hispanic) - (0.5241286 \times Asian) - (1.364379 \times Islander) \\ & + (0.952699 \times Native) + (0.4845208 \times PMT) + (0.5856777 \times FH) \\ & + (2.542472 \times Former) + (2.799727 \times Current) \\ & - \left([0.1815486 \times \left(\frac{CPD^{-1}}{100} - 4.021541613\right)\right] + [0.0305566 \times (Duration - 27)] \\ & - [0.0321362 \times (QuitDuration - 8.593417626)] \end{aligned} \tag{10.4}$$

To apply this equation the following applies;

$$\begin{aligned} Condition/Race/SmokingStatus &= \begin{cases} 0 & \text{if not present} \\ 1 & \text{if present} \end{cases} \\ CPD/Duration/QuitDuration &= \begin{cases} 0 & \text{if not applicable} \\ \text{Participant Value} & \text{if present} \end{cases} \end{aligned} \tag{10.5}$$

10.4 Applying Single Updating Methods

The methods to update the single models are now applied to the $\text{PLCO}_{\text{M2014}}$ Model. The updated models are presented and their performance reviewed to assess if an improved model has been devised.

10.4.1 Updating the Intercept

Updating the intercept to reflect the observed incidence rate in the dataset was a straightforward method to apply. However, due to the high proportion of participants with lung cancer the intercept was drastically modified (Table 10.3). The new risks generated by the model would be inappropriate in a population with a more realistic lung cancer incidence rate.

The calibration improved in the internal and external validations, as both had high lung cancer incidence rates, shown by the Hosmer-Lemeshow chi-squared and the Brier score (Table 10.4). However, the model did not report a good calibration which would indicate the $\text{PLCO}_{\text{M2014}}$ Model may benefit from more extensive model updating.

This method could be very successful if the model was updated using a cohort that represented any future screening population. This is demonstrated by the improvement in calibration, and if the model is updated in a sample population that is reflective of the target population, which essentially the model updating and NY Wynder dataset were, then this method may produce a well calibrated model.

Overall, updating the intercept is a simple method to improve the calibration. This method is ideal for a successful model with a calibration deficiency in a new population. The results demonstrate how the method alone will not guarantee a good calibration for the updated model if the original model has a large calibration deficiency. In these instances, more extensive model updating techniques may yield better results. Since the discriminative ability remains unaltered, and the model has been recalibrated to estimate unreasonable risks, an improved prediction model has not been devised.

10.4.2 Recalibration of Intercept and Slope

The intercept and regression equation were recalibrated using the line of best fit between the predicted and log odds of the observed risk. Since the presence of lung cancer is a binary outcome, the participants in the dataset were grouped. Participants were grouped to their closest neighbours based on their predicted risks, in groups of 10. The observed incidence rate for the group was determined and plotted against the mean risk generated by the model for the participants per group. At this stage the intercept and slope modifiers could be simply calculated. The plot and line of best fit for the recalibration is presented.

Coefficient	Scaling Parameter
Intercept	0.5667
Slope	0.0259

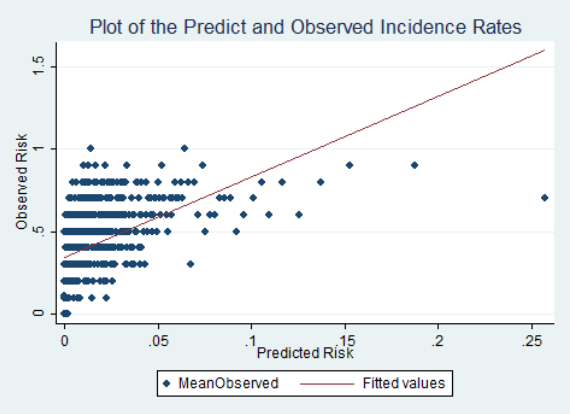

Table 10.2: Recalibrating the Intercept and Slope

The intercept from the line of best fit was added to the original model intercept term and the remaining parameters were scaled using the line of best fit gradient. This created the final recalibrated model as presented in Table 10.3.

The updated model reported a poor performance and updating only the intercept led to a greater improvement in the model calibration when externally validated (Table 10.4).

Overall, the method is easy to apply but was limited when applied to the $PLCO_{M2014}$ Model. The model underperformed probably due to the poor discriminative ability in the model updating dataset. As a result, when grouping the participants, individuals with lung cancer were not commonly grouped in higher predicted risk groups. An improved lung cancer prediction model was not created.

10.4.3 Recalibrated Null Model and Selective Re-estimation

The next method selectively re-estimates the variables of the model. Firstly the model was recalibrated using the previous method although this model was still not well calibrated. The variables included in the null model were identified through forward selection. Family history of lung cancer, CPD, and quit duration were removed from the null model as they did not improve the model goodness of fit. The final stage compared the goodness of fit between the recalibrated and null recalibrated model and the shrinkage

Method	Intercept	Age	Ethnicity					Education	BMI
		(Age - 62)	Black Yes = 1	Hispanic Yes = 1	Asian Yes = 1	Amer-Ind Yes = 1	Hawaiian Yes = 1	(Education - 4)	(BMI - 27)
No Updating	-7.02198	0.07960	0.32116	-0.82033	-0.52413	0.95270	-1.36438	-0.08793	-0.02895
Intercept	-3.13947	0.07960	0.32116	-0.82033	-0.52413	0.95270	-1.36438	-0.08793	-0.02895
Recalibration	-6.45146	0.00212	0.00855	-0.02184	-0.01395	0.02536	-0.03632	-0.00234	-0.00077
Re-estimation	-6.45146	0.00212	0.00855	-0.02184	-0.01395	0.02536	-0.03632	-0.00234	-0.00077
Mean	-0.33445	0.07955	0.32099	-0.81989	-0.52384	0.95218	-1.36363	-0.08788	-0.02893
Extension	-6.45146	0.00212	0.00855	-0.02184	-0.01395	0.02536	-0.03632	-0.00234	-0.00077
Sel. Extension	-6.45146	0.00212	0.00855	-0.02184	-0.01395	0.02536	-0.03632	-0.00234	-0.00077
Mean Extension	-0.33445	0.07955	0.32099	-0.81989	-0.52384	0.95218	-1.36363	-0.08788	-0.02893
	COPD	**PMT**	**Fam. Hist.**	**Smoking History**		**CPD**	**Smoke Dur.**	**Quit Duration**	**Gender**
				Former	Current				
	Yes = 1	Yes = 1	Yes = 1	Yes = 1	Yes = 1	(CPD^1/100) - 4.02154)	(Dur. - 27)	(Dur. - 8.593417)	Male = 1
No Updating	0.34573	0.48452	0.58568	2.54247	2.79973	-0.18155	0.03056	-0.03214	NA
Intercept	0.34573	0.48452	0.58568	2.54247	2.79973	-0.18155	0.03056	-0.03214	NA
Recalibration	0.00920	0.01290	0.01559	0.06768	0.07453	-0.00483	0.00081	-0.00086	NA
Re-estimation	0.00920	0.01290	0.00367	0.06768	0.07453	-0.00114	0.00081	-0.00020	NA
Mean	0.34554	0.48426	0.58536	2.54108	2.79820	-0.18145	0.03054	-0.03212	NA
Extension	0.00920	0.01290	0.00367	0.06768	0.07453	-0.00114	0.00081	-0.00020	6.81149
Sel. Extension	0.00920	0.01290	0.00367	0.06768	0.07453	-0.00114	0.00081	-0.00020	6.81149
Mean Extension	0.34554	0.48426	0.58536	2.54108	2.79820	-0.18145	0.03054	-0.03212	NA

Table 10.3: Single Model Updating Coefficients

Method	Validation	Internal/External	Hos.-Leme. (P-value (χ^2))	Brier Score	AUC [95% CI]	Threshold (%)	Sens. (%)	Spec. (%)	Youden	PLR
Intercept	1	Internal	0.0019 (804.34)	0.2481	0.6813 [0.669, 0.694]	13.08	78.57	50.10	0.2867	1.5744
		External	0 (1125.45)	0.2337	0.7667 [0.757, 0.776]	13.09	87.02	54.08	0.4109	1.8949
	2	Internal	0.1494 (731.71)	0.2504	0.6568 [0.642, 0.671]	61.57	19.60	90.00	0.0961	1.9609
		External	0.0002 (1085.13)	0.2134	0.7834 [0.773, 0.794]	59.63	35.02	89.99	0.2502	3.5002
Recalibration	1	Internal	0 (62292.87)	0.4157	0.6813 [0.669, 0.694]	0.16	78.57	50.10	0.2867	1.5744
		External	0 (13498.18)	0.5142	0.7667 [0.757, 0.776]	0.16	87.02	54.08	0.4109	1.8949
	2	Internal	0 (68377.97)	0.4330	0.6568 [0.642, 0.671]	0.17	19.60	90.00	0.0961	1.9609
		External	0 (93825.14)	0.5177	0.7834 [0.773, 0.794]	0.17	35.02	89.99	0.2502	3.5002
Re-estimation	1	Internal	0 (9642.2)	0.4157	0.6807 [0.668, 0.693]	0.17	74.12	54.19	0.2831	1.6181
		External	0.0102 (1027.66)	0.5142	0.7489 [0.739, 0.759]	0.17	80.74	57.15	0.3790	1.8845
	2	Internal	0 (11164.24)	0.4330	0.6569 [0.642, 0.671]	0.17	19.31	90.00	0.0931	1.9313
		External	0 (1138.76)	0.4728	0.7662 [0.755, 0.777]	0.17	31.81	89.99	0.2180	3.1790
Mean	1	Internal	0.001 (837.96)	0.3780	0.6840 [0.672, 0.697]	90.51	79.88	49.65	0.2953	1.5866
		External	0 (2029.36)	0.3177	0.768 [0.758, 0.778]	90.50	88.81	52.90	0.4171	1.8855
	2	Internal	0.002 (805.07)	0.4123	0.6591 [0.645, 0.674]	99.18	20.41	90.00	0.1041	2.0412
		External	0 (2361.17)	0.4238	0.7805 [0.770, 0.791]	99.18	34.48	89.99	0.2448	3.4463
Extension	1	Internal	0 (1429.68)	0.3418	0.5848 [0.571, 0.598]	15.58	45.83	55.77	0.0160	1.0362
		External	0 (2604.77)	0.3489	0.643 [0.632, 0.655]	0.18	60.95	42.85	0.0379	1.0664
	2	Internal	0 (1184.31)	0.3505	0.5657 [0.550, 0.581]	0.16	4.90	90.00	-0.0510	0.4902
		External	0 (2520.25)	0.3485	0.6513 [0.639, 0.664]	61.08	25.69	89.99	0.1568	2.5673
Sel. Extension	1	Internal	0 (1429.68)	0.3418	0.5848 [0.571, 0.598]	15.58	45.83	55.77	0.0160	1.0362
		External	0 (2604.77)	0.3489	0.643 [0.632, 0.655]	0.18	60.95	42.85	0.0379	1.0664
	2	Internal	0 (1184.31)	0.3505	0.5657 [0.550, 0.581]	0.16	4.90	90.00	-0.0510	0.4902
		External	0 (2520.25)	0.3485	0.6513 [0.639, 0.664]	61.08	25.69	89.99	0.1568	2.5673
Mean Extension	1	Internal	0.001 (837.96)	0.3780	0.6840 [0.672, 0.697]	90.51	79.88	49.65	0.2953	1.5866
		External	0(2029.36)	0.5142	0.768 [0.758, 0.778]	90.50	88.81	52.90	0.4171	1.8855
	2	Internal	0.002 (805.07)	0.4123	0.6591 [0.645, 0.674]	99.18	20.41	90.00	0.1041	2.0412
		External	0 (2361.17)	0.4238	0.7805 [0.770, 0.791]	99.18	34.48	89.99	0.2448	3.4463
Validation 1; Optimal Risk Threshold - Validation 2; UKLS Guidelines										

Table 10.4: Single Model Updating Validation Results

was applied. As shown in Table 10.3 these three variables were modified from the recalibrated model in method two.

The model calibration improved considerably in comparison to the recalibrated model in method two. This highlights how the three identified variables which did not explain lung cancer risk, and were minimised in the final model, successfully improved the model calibration. Although due to the calibration deficiencies of the original and recalibrated model, the new model was still not well calibrated. Additionally, the discrimination, and prediction rules were slightly inferior to the original model. The AUC was lower in both the internal and external dataset and the prediction rules were consistently lower (Table 10.4).

Overall the method underperformed although demonstrated how minimising weaker variables improved the model calibration. However, if the original and recalibrated models have calibration deficiencies then more extensive model updating may be required. Additionally, this method reported a lower discrimination and lower prediction rules indicating minimising variables may improve the model calibration but this does not guarantee a more robust selective screening tool has been devised.

10.4.4 Shrinkage and Re-estimation

The second re-estimation method centers the variables on the mean incidence rate observed in the dataset. These are then subtracted from the variable coefficients so that the risks are normalised around the mean. The intercept was modified to reflect the mean observed incidence rate.

The variables are not minimised if they do not improve the model goodness of fit, as with the previous re-estimation method, instead all are shrunk by the same shrinkage coefficient, which was easily determined. The final model is presented in Table 10.3, although the mean values are not presented in the table.

The updated model was still poorly calibrated. This could be a result of the calibration deficiency of the original model and the predictors in the model being limited in explaining lung cancer risk. This method can affect the model discrimination and prediction rules. Indeed, the model reported the highest AUC in both the internal and external validation, an improvement over the original model, although the improvement was not significant. The prediction rules had a slight improvement with an increased Youden index and PLR (Table 10.4). The risk thresholds for the optimal performance and to maintain a specificity of 90% was distorted when recalibrated to reflect the high incidence rate in the model updating dataset, and were 90.5% and 99.17% respectively. This would be inappropriate in a real screening programme and is a result of updating the models in a dataset with a high lung cancer incidence rate. Overall, this method had some encouraging results, offering a slight improvement in discrimination and prediction rules, the method may be successful when applied in cohort studies where the model would not be recalibrated such that inappropriate risks are generated.

10.4.5 Re-estimation and Extension

The next method updated the $\text{PLCO}_{\text{M2014}}$ Model by re-estimating the original parameters and extending the model to include any additional parameters. In the dataset, gender was the sole variable available for model extension. Unfortunately, only including one parameter will not allow a good comparison between the different model extension methods; to evaluate whether over-fitting is caused by forced extension as with this method.

It was expected that this method may under-perform as the new variables are added without considering whether they sufficiently explain lung cancer risk. The method was straightforward to implement and used the previous techniques to first re-estimate the model. Then the new parameters were included through standard model building techniques. The coefficients are presented in Table 10.3. It is noticeable that the new coefficient for gender is inappropriate and heavily distorts risks between males and females. This is due to the calibration deficiencies of the re-estimated model.

The inappropriate estimates for risk caused by the new gender variable can be seen in the external validation. The calibration improved upon the re-estimated model in the internal validation. However, in the external validation the calibration suffered and the model was poorly calibrated (Table 10.4). The

AUC was heavily reduced from the original models and the prediction rules performance were much poorer (Table 10.4).

This method underperformed and reported the weakest results of all the methods. This large reduction in calibration performance (in the external validation), AUC and prediction rules demonstrates model extension is inappropriate if the original model significantly under or over estimates the disease prevalence rate. The new variables are not added appropriately and while an improvement may be seen in an internal validation the model is not robust in external population.

10.4.6 Selective Re-estimation and Selective Extension with Recalibration

The next approach was a variation on the previous method by only including additional variables if they demonstrated a significant improvement in the model goodness of fit. Gender, the only variable available, was included at the 0.05 significance level (Figure 10.5) and the final model is identical to the previous method.

Variable	Coefficient	T-Value	P-Value
Gender	6.81	91.47	< 0.001

Table 10.5: Extending the Recalibrated Model

In summary, the final model corresponds to the previous method. The final model performance is suboptimal, resulting from the new variable for gender being inappropriate in new sample populations. This is due to the calibration deficiencies of the original model.

10.4.7 Re-estimation and Selective Extension without Recalibration

The final method extended the recalibrated mean model. Gender was considered as the additional variable for inclusion using forward selection with a critical p-value of 0.05. Gender was rejected as it did not significantly improve the goodness of fit. Therefore, the final model was the same as the recalibrated mean model in method 4 (Table 10.3).

Variable	Coefficient	T-Value	P-Value
Gender	-1.11	-1.93	0.053

Table 10.6: Extending the Recalibrated Mean Model

This method reported the leading performance with an improved AUC and prediction rules over the original model, although the improvement was marginal. Additionally, the high lung cancer incidence rate in the dataset recalibrated the model to produce unrealistically high lung cancer risk estimates, including by estimating 6-year risk exceeding 90% in individuals. Therefore, this model would not be appropriate in a population with more realistic lung cancer prevalence rates.

10.5 Summary

The chapter updated the $\text{PLCO}_{\text{M2014}}$ Model in an attempt to create an improved model. The validation of the new models revealed that the updated models generated unrealistic risks for an individual's 6-year risk of developing lung cancer, as a result of the model being updated in a dataset with an unrealistic lung cancer incidence rate. None of the models reported a good calibration but this can be attributed to the calibration deficiency of the original model in this dataset.

Additionally, none of the updated models showed a significantly improved discrimination or prediction rules in comparison to the $PLCO_{M2014}$ Model. The mean model had the strongest performance of the updated models and while offering some improvement, in comparison to the original model, there was not a large difference. Since the risk estimates generated by this model are unrealistic with participants having a 6-year risk of developing lung cancer exceeding 99%. This would create unnecessary panic and it would be unwise to promote this model to the public.

None of the methods were successful in devising an improved model, but this may be a limitation of the datasets collected for this project. The methods failed to improve the models discriminative ability and prediction rules performance. This included extending the model to include gender, which was inadequately utilised in the extended model to address the poor model goodness of fit. When extending a model to include additional variables the model needs to be reasonably well calibrated in the dataset to prevent inappropriate coefficients being generated.

The next chapter will evaluate methods to aggregate multiple prediction models to create a meta-model.

CHAPTER 11

Aggregating Multiple Prediction Models

11.1 Introduction

Model Averaging and BMA, with or without an additional weighting based on the models' discriminative ability, were identified as appropriate methods to aggregate the Bach, Hoggart, Pittsburgh, and $PLCO_{M2014}$ models. Two versions of model aggregation will be conducted, with or without the Bach Model, because excluding the Bach Model allowed for a larger IPD to base the model aggregation, and the final meta-model is applicable to a wider range of people. The new models will be presented and externally validated to assess if model aggregation created an improved lung cancer prediction model. This chapter will also review the ability of the methods to create a more robust model. The successful methods will be identified so they can then be implemented in future research to aggregate multiple prediction models; whether for newly developed lung cancer models or a completely different disease. Although, we will have to consider whether the high lung cancer incidence rate influences the methods' ability to aggregate the models.

11.2 Objectives

This chapter will apply the model aggregation methods and validate the meta-models to assess if an improved prediction model can be devised, and review the different methods. To achieve this the chapter will;

1. Apply Model Averaging and BMA methods to the Bach, Pittsburgh, Hoggart, and $PLCO_{M2014}$ Models.
2. Present the new meta-models.
3. Externally validate the new models in comparison to the original models to assess if an improved model is devised.
4. Present any improved model with a Stata code alongside recommendations on how to apply the model.
5. Review the practicality and performance of the model aggregation methods.

11.3 Method One - Model Averaging

Model averaging assigned a weight to each model based on their calibration results in the dataset. This method is advantageous as it can combine models with unique forms.

Firstly, the original models were recalibrated in the dataset. The four lung cancer models were re-calibrated by updating the intercept and modifying the slope using a single scaling factor. This method

was previously applied updating the $PLCO_{M2014}$ Model, and was a simple method although there are limitations to this method in a dataset with such a high lung cancer incidence rate.

The recalibrated models were then applied in the dataset to calculate each individual's risk. Then separately, for each model, the following equation was applied to assess the model calibration;

$$L_i = \sum_{j=1}^{N} (y_j \times ln(m_j)) + ((1 - y_j) \times ln(1 - m_j)) \tag{11.1}$$

$$\text{where } y_j = \begin{cases} 0, \text{ if participant is a control} \\ 1, \text{ if participant is a case} \end{cases} \tag{11.2}$$

Here, y_i is the observed outcome for the participants. This is a binary outcome which is either 0 if they are lung cancer free or 1 if they have lung cancer. After applying Formula 11.1 the BIC was determined as follows;

$$BIC_M = (-2 \times L_M) + (2 \times ln(N)) \tag{11.3}$$

Here, N is the number of participants in the dataset. The 2 was determined by the number of parameters estimated in the model recalibration. This is fixed for each model as they were all updated using the same method calculating the intercept modifier and the scaling factor for the slope. If more extensive recalibration was required then this would penalise the model. Next, the final weighting for each model was easily calculated as follows;

$$w_M = \frac{exp(-0.5 \times BIC_M)}{\sum_{m=1}^{4} exp(-0.5 \times BIC_m)} \tag{11.4}$$

Model Averaging was simple to apply with the final weights easily determined. The final model weightings are presented in Table 11.1.

Model	**Weightings**	
	With Bach	Without Bach
Bach	0.25271	NA
PLCO	0.25201	0.33766
Hoggart	0.24460	0.32651
Pittsburgh	0.25068	0.33582

Table 11.1: Model Weightings from Model Averaging

For both versions the final weights given to the separate models was very similar. This may be a result of all the models being poorly calibrated even after the recalibration. This is most likely a consequence of the high lung cancer incidence rate in the model aggregating dataset. Therefore, the method could not identify the better calibrated models to assign a higher weight.

The new models were validated to assess if model averaging created a more robust prediction model.

Bach	**Validation**	**Internal/External**	**Hos.-Leme.**	**Brier Score**	**AUC [95% CI]**	**Threshold (%)**	**Sens.**	**Spec.**	**Youden**	**PLR**
With	1	Internal	0.3617 (277.63)	0.4863	0.5337 [0.512, 0.556]	1.05	28.444	82.980	0.1142	1.6713
		External	0.4776 (416.96)	0.2859	0.6239 [0.605, 0.643]	0.76	99.967	0.000	-0.0003	0.9997
	2	Internal	0.3617 (277.63)	0.4863	0.5337 [0.512, 0.556]	1.06	19.556	89.993	0.0955	1.9541
		External	0.4776 (416.96)	0.2859	0.6239 [0.605, 0.643]	0.67	19.056	90.026	0.0908	1.9105
Without	1	Internal	0.0001 (648.84)	0.4753	0.5920 [0.577, 0.608]	1.41	84.755	31.648	0.1640	1.2400
		External	0.0003 (855.97)	0.3873	0.6808 [0.668, 0.694]	1.23	100.000	0.037	0.0004	1.0004
	2	Internal	0 (669.64)	0.4705	0.5801 [0.563, 0.597]	1.33	12.294	89.987	0.0228	1.2278
		External	0.0527 (779.39)	0.3885	0.6883 [0.674, 0.702]	1.08	25.399	89.982	0.1538	2.5353
Validation 1; Optimal Risk Threshold - Validation 2; UKLS Guidelines										

Table 11.2: Model Averaging Validation Results

The calibration results for the meta-model including the Bach Model was promising. The meta-model was well calibrated in both the internal and external datasets (Table 11.2), most likely a consequence of recalibrating the models before aggregating them. However, the promising calibration results are not replicated in the meta-model excluding the Bach Model, while the calibration improves, highlighted by the Hosmer-Lemeshow chi-squared and the Brier Score results, the model did not report a good calibration.

Unfortunately, the promising calibration improvement was not reflected in an improved discriminative ability or prediction rules performance. Indeed, the proposed aggregated model performs poorer than the original models.

In summary, Model Averaging is a simple method to apply and distinct models can be combined. However, when the original models are poorly calibrated, even after recalibration, then the weights assigned to each model in the final meta-model are very similar. While this method notably improved the model calibration, to the extent the meta-model with the Bach Model was well calibrated in both the internal and external validation, the discriminative ability of the new model was poorer than the original models. While Model Averaging can create a well calibrated model, if the model updating datasets has a high disease prevalence rate, as observed in this instance, then the proposed meta-model would be inappropriate in populations with more realistic incidence rates. Overall, Model Averaging failed to improve the prediction rules in comparison to the original models and a more robust lung cancer prediction model was not created.

11.4 Method Two - Bayesian Model Averaging

The next approach is Bayesian Model Averaging (BMA) with a non-informative prior. There are similarities between this method and Model Averaging, the models are applied separately and then weighted in the meta-model. The models are assigned a weight based on their calibration in the dataset. However, the weights are calculated slightly differently and the models are not initially recalibrated unlike Model Averaging. The original models were poorly calibrated in the dataset due to the high incidence rate. Not recalibrating the model originally could see some of the models being penalised in the final meta-model while actually being successful models.

For each model the likelihood function in the dataset was calculated in the same way as Model Averaging (Equation 11.5), as follows;

$$L_i = \sum_{j=1}^{N} \left(y_j \times ln\left(m_j\right)\right) + \left(\left(1 - y_j\right) \times ln\left(1 - m_j\right)\right) \tag{11.5}$$

$$\text{where } y_j = \begin{cases} 0, \text{ if participant is a control} \\ 1, \text{ if participant is a case} \end{cases} \tag{11.6}$$

Where, m_j is the predicted risk for participant j in the model.

After calculating the likelihood function, L_i, for each model the following calculation was performed, as presented in Equation 11.7. To perform this calculation the model dimension, d_j, is required, which is the number of distinct variables in the model;

$$M_i = e^{\left(L_i + \frac{d_i}{2} ln(N)\right)} \tag{11.7}$$

Here, N is the number of participants in the dataset which is fixed because they are applied in the same sample population for which all the models are applicable. For the lung cancer prediction models their model dimensions are as follows;

Model	Model Dimension
$PLCO_{M2014}$	12
Hoggart	12
Pittsburgh	4
Bach	7

Table 11.3: Lung Cancer Model Dimensions

Now, M_i can be determined and the final model weightings calculated;

$$w_i = \frac{M_i}{\sum_{j=1}^{n} M_j} \tag{11.8}$$

This was a straightforward method to combine multiple lung cancer prediction models. The weighting applied to each model to produce a final lung cancer risk are as follows, for both options including or excluding the Bach Model;

Model	Weightings	
	Model with Bach Model	Model without Bach Model
Bach	2.585E-09	NA
PLCO	1.852E-03	0.500002
Hoggart	0.99815	0.499998
Pittsburgh	7.292E-17	1.86E-18

Table 11.4: Model Weightings from BMA

The two different versions led to unusual model weights. When considering the Bach Model, the model weightings does not best synthesis the evidence from the distinct models by minimising the impact of some models in the meta-model. Indeed, the final model was effectively the Hoggart Model which was assigned a weight exceeding 0.998. The remaining models were assigned weights very close to zero. This demonstrates how BMA can assign excessive weights to some models because the approach assumes some models are correct and the remaining models should be penalised. When excluding the Bach Model from BMA the model weights were more evenly distributed. The meta-model is effectively a combination of the $PLCO_{M2014}$ and Hoggart models, both assigned a weight of effectively 0.5. The Pittsburgh Model was still minimised in the final model.

Bach	Validation	Internal/External	Hos.-Leme.	Brier Score	AUC [95% CI]	Threshold (%)	Sens. (%)	Spec. (%)	Youden	PLR
With	1	Internal	0 (494.26)	0.3726	0.5295 [0.508, 0.551]	25.05	29.63	76.48	0.0611	1.2597
		External	0.0021 (503.49)	0.5148	0.6103 [0.591, 0.629]	24.98	31.14	81.01	0.1215	1.6396
	2	Internal	0 (494.26)	0.3726	0.5295 [0.508, 0.551]	32.24	12.59	89.99	0.0259	1.2583
		External	0.0021 (503.49)	0.5148	0.6103 [0.591, 0.629]	32.50	18.66	90.03	0.0869	1.8708
Without	1	Internal	0 (726.68)	0.4220	0.6114 [0.596, 0.627]	1.50	82.86	36.21	0.1907	1.2990
		External	0 (998.86)	0.5307	0.6922 [0.680, 0.705]	1.50	84.06	42.15	0.2622	1.4532
	2	Internal	0 (735.91)	0.4152	0.5981 [0.581, 0.615]	15.97	12.67	89.99	0.0266	1.2654
		External	0 (876.01)	0.5227	0.7000 [0.686, 0.714]	14.05	25.91	89.98	0.1589	2.5865
Validation 1; Optimal Risk Threshold - Validation 2; UKLS Guidelines										

Table 11.5: BMA Validation Results

The meta-model including the Bach Model had a similar performance to the Hoggart Model, as expected, when validated across all tests 11.5. When excluding the Bach Model in the meta-model, there was an improvement in performance. However, this did not exceed the performance of the original $PLCO_{M2014}$ Model. Additionally, the Hoggart Model predicts unreasonable high estimates of developing lung cancer

over 6-years. Therefore, the meta-model to maintain a specificity of 90% would need to be applied at approximately the 15% risk threshold (Table 11.5). This may be inappropriate and suggests recalibration before applying the method may allow more appropriate risks to be developed. However, this may not be advisable in a dataset with an unrealistic lung cancer incidence rate because the recalibrated models may estiamte even further distorted risks.

In summary, there was a contrasting performance between the two meta-models created by BMA. When considering the Bach Model the method's limitations are highlighted as all except one model are negated in the final meta-model. When excluding the Bach Model the weighting is slightly less biased but the Pittsburgh Model is still minimised in the final model. The method may benefit from initially recalibrating the models to allow a fairer inclusion of the models. While the meta-model excluding the Bach Model had a reasonable performance this did not sufficiently improve upon the original models in the external validation so a more robust model was not created.

11.5 Method Three - Bayesian Model Averaging with an Additionally Weighting

The next method is an adaptation of BMA to include an additionally weighting in an attempt to create an improved selective screening tool. This is to allow more evidence to be included in determining the weights for the models. The additionally weighting is based upon a comparison of the discriminative ability of the models in the dataset. A model with an increased discrimination, measured by the AUC, is assigned a stronger prior. Although, unless there is a large difference in the model AUC the change in final weights will be minor. BMA with an additionally weighting is simple to apply, as with the original BMA method, once the AUC for each of the models has been determined.

The posterior, M_j of the model is calculated in exactly the same way as previously presented using Equations 11.5 & 11.7. The additionally weighting takes the form of the prior odds for each model and is determined as follows;

$$Pr(M_i) = \frac{AUC_i}{\sum_{j=1}^{n} AUC_j} \tag{11.9}$$

Then, for each model, j, we can assign a new value for m_j determined by the likelihood ratio. Then the final model weights can be calculated;

$$w_j = \frac{Pr(M_j) \times m_j}{\sum_{i=1}^{n} Pr(M_j) \times m_i} \tag{11.10}$$

This provides the finals weights for each model, which for the lung cancer prediction models were determined as follows;

Model	**Weightings**	
Bach	2.71E-09	NA
PLCO	0.00206	0.50742
Hoggart	0.99794	0.49258
Pittsburgh	7.24E-17	6.66E-16

Table 11.6: Model Weightings from BMA with an Informative Prior

As can be seen in Tables 11.4 & 11.6, the difference in weights between BMA and BMA with an additionally weighting was negligible because the AUC results were similar and the external data did not provide enough evidence to significantly alter the model weights. However, the $\text{PLCO}_{\text{M2014}}$ was slightly rewarded with a higher weight as it had the leading discriminative ability. Considering a prior based on

the AUC results would be a more appropriate method if there was a clear leading model which should be favourably weighted in the final model.

Bach	Validation	Internal/External	Hos.-Leme.	Brier Score	AUC [95% CI]	Threshold (%)	Sens. (%)	Spec. (%)	Youden	PLR
With	1	Internal	0 (491.05)	0.3726	0.5295 [0.508, 0.551]	25.04	29.63	76.48	0.0611	1.2597
		External	0.0024 (501.84)	0.5149	0.6103 [0.591, 0.629]	24.98	31.14	81.01	0.1215	1.6396
	2	Internal	0 (491.05)	0.3726	0.5295 [0.508, 0.551]	32.23	12.59	89.99	0.0259	1.2583
		External	0.0024 (501.84)	0.5149	0.6103 [0.591, 0.629]	31.20	18.79	90.03	0.0882	1.8840
Without	1	Internal	0 (744.14)	0.4226	0.6115 [0.596, 0.627]	1.50	82.70	36.44	0.1914	1.3011
		External	0 (1030.13)	0.5316	0.6924 [0.680, 0.705]	1.50	83.91	42.38	0.2628	1.4562
	2	Internal	0 (716.55)	0.4158	0.5983 [0.582, 0.615]	15.81	12.62	89.99	0.0261	1.2607
		External	0 (971.88)	0.5236	0.7003 [0.686, 0.714]	13.87	25.97	89.98	0.1595	2.5919
Validation 1; Optimal Risk Threshold - Validation 2; UKLS Guidelines										

Table 11.7: BMA Validation Results with an Informative Prior

Since the meta-model had a very similar weighting as the original BMA meta-models there were no major differences in the validation results. Neither version of the meta-model including or excluding the Bach Model reported a good calibration, although the version with the Bach Model nearly reported a good calibration in the external validation. This was an improvement upon the original models; further supported by the improved Brier Score. This suggests BMA can improve the accuracy of predictions.

The discriminative ability of the new models was similar to the initial models. The meta-model that excluded the Bach Model reported the strongest discriminative ability and prediction rules. This reported a small improvement over the BMA version without the informative prior. This demonstrates how considering the discriminative ability as prior knowledge can improve the potential of developing a robust selective screening tool. This had a comparable performance to $\text{PLCO}_{\text{M2014}}$ Model. Indeed, the model in the external validation had a higher Youden index and PLR than the leading original models. This meta-model, excluding the Bach Model, was optimal at the 1.5% risk threshold where the sensitivity was approximately 83% across the internal and external validation dataset. Here, the model reported a specificity between 36-42%. Using the UKLS guidelines the model should be applied at a risk threshold between 13.8-15.8%. This maintained a specificity of 90%, although the sensitivity was variable reporting a sensitivity of 12.6% in the internal validation which increased to 26% in the external validation. The results have only been confirmed in one external dataset; the model should be validated in additional sample populations to assess if it can consistently perform to a higher standard in comparison to the original models.

In summary, BMA with an additionally weighting led to a slightly improved model. The model version excluding the Bach Model had a strong performance. In the external validation the calibration improved upon the original models, although the meta-model still did not report a good calibration because of the calibration deficiencies of the original models. The discrimination and prediction slightly improved upon the leading original models. However, the results were variable between the internal and external validation. The model should be applied in new populations to ascertain if an improved prediction model has been created in comparison to the original models.

Overall, this method showed potential for creating a more robust model. Considering a prior for each model could be a better alternative to the original BMA method. The additionally weighting, based on the models discriminative ability, could allow a more robust selective screening tool to be devised. However, if there is not a large difference in the original models AUCs then the prior information does not significantly alter the model weights from the original BMA method.

11.6 Summary

The chapter applied Model Averaging and BMA methods to aggregate multiple lung cancer prediction models. Model Averaging dramatically improved the model calibration both in the internal and external validation. This was an advantage of recalibrating the original models before aggregating the models in the dataset. The models were recalibrated in a dataset with a high lung cancer incidence rate, so that when

they were applied in the external validation dataset, which also reported a high lung cancer incidence rate, they remained well calibrated. Unfortunately, the discrimination and the prediction rules were inferior to the original models. Additionally, the final weights were evenly split across all models considered in the meta-model. This may be a result of the original models still being poorly calibrated in the dataset so the method failed to identify more robust models.

BMA did not improve the calibration and minimised the impact of some of the original models by effectively assigning them a weight of zero in the final model. This was observed when considering the Bach Model in the model aggregating; the final model was effectively the Hoggart Model. When excluding the Bach Model, the final model was formed using an equal weighting between the $\text{PLCO}_{\text{M2014}}$ and Hoggart Model, however, the Pittsburgh Model was still minimised in the final meta-model. When incorporating in the models' discriminative ability as an additionally weighting the final weights differed only slightly because the AUC results were relatively similar. This method, however, created a meta-model with a slightly improved discrimination and prediction rules to the original models.

A lung cancer prediction model was devised with a strong performance in comparison to the original models. This model combined the evidence of the $\text{PLCO}_{\text{M2014}}$, Hoggart and Pittsburgh models. It can be applied to ever-smokers aged at least 35 years to predict six year risk of lung cancer incidence. The model performed optimally at the 1.5% risk threshold with a sensitivity of 83% and specificity between 36-42%. To maintain a specificity of 90% the model should be applied at a risk threshold between 13.8-15.8%, which is excessive because of the Hoggart Model estimated high risks. At this threshold the sensitivity varied between 12.6% and 26%. The model should be applied in new external validations to ascertain if an improved prediction model has been created in comparison to the original models. The final model is the weightings of the $\text{PLCO}_{\text{M2014}}$, Hoggart, and Pittsburgh models as presented in Table 11.6, and a Stata code of each of the original models are provided in 12.13.

CHAPTER 12

Discussion

12.1 Developing Project Objectives and Brief Summary

The primary objective of this project was to review how lung cancer diagnosis rates could be improved. We highlighted how the key to improving the diagnosis rate was to identify lung cancer early. This could be achieved by implementing a screening trial, where people who are high risk of developing lung cancer are periodically reviewed and screened to identify early lung cancer developments. Therefore, we aimed to review how lung cancer screening could be implemented and improved.

Initially, an overview of lung cancer and current screening programmes was conducted. This highlighted how research was still needed as to how to effectively select populations for lung cancer screening. Firstly, there was an indication the current proposed populations for screening trials were somewhat arbitrarily chosen without evidence to support why they had been selected. The two major current screening proposals either selected the LLP Model despite this model not demonstrating a leading prediction model performance, or selecting older ever-smokers without consideration into other key factors that could explain lung cancer risk. Despite these limitations, the trials, when evaluated, demonstrated some promising results, including improving early stage lung cancer diagnosis. This highlighted the potential benefit of a lung cancer screening programme. However, before a programme could be implemented there should be more justification as to why a specific criterion had been chosen with evidence demonstrating why this would be the most beneficial method available. Without this evidence available with the current screening programmes, a systematic review was proposed to quantify all information into available prediction models, which could become a selective screening programme, and any evidence into their expected performance as a clinical utility.

We also discussed the key measures that would be expected of a criterion before it could be implemented, with focus on the UK screening trial guidelines. Namely, a screening criterion would be required to demonstrate a high level of benefit in identifying lung cancer while reducing the potential for negative impacts caused by unnecessary screening. It also had to demonstrate the potential to be cost-effective, which could be challenging due to the expensive CT scanning commonly required to identify lung cancer. This was considered during all assessments of how prediction models could be implemented as a selective screening trial.

The next stage of the project aimed to identify any leading selective screening criteria formulated using either the existing lung cancer screening programmes or available prediction models. Therefore, a systematic review was conducted to identifying any leading screening guidelines currently published. However, our research found the current reporting into how lung cancer prediction models' could be utilised as a selective screening tool was limited. The results reported were inconsistent and, with a lack of compelling evidence demonstrating their potential benefits, this contributed to lung cancer screening not currently being considered on a large scale.

This formed the next objectives of this research to identify a leading selective screening criterion. We requested datasets from ILCCO to provide a range of sample populations to review existing lung

cancer models and currently considered selective screening criteria. The models and criteria were then analysed in the datasets and recommendations on how to define a high risk population for screening were presented. By the conclusion of the external validation we presented two alternatives, both using the successful $PLCO_{2014}$ Model. One criterion would identify a large proportion of lung cancer cases but have a substantial proportion of additional screening of controls; this would be beneficial if and when lung cancer screening becomes easier, cheaper, and less invasive. The second option, while still identifying a substantial proportion of lung cancer cases aimed to reduce screening of people who were lung cancer free, to allow the screening programme to be more economically viable and reduce harms. By the conclusion of external validation the objectives had been met and a leading selective screening tool had been presented.

Next, the project presented how models could be updated to combine the evidence and information available in the model building dataset with new information in an external dataset. This can be achieved by updating a single model or combining multiple models based on their performance in the external dataset. The literature review presented the different methods that were available which have differing merits depending on the success of the original model in the external dataset. Model aggregating techniques were also presented. These methods to combine models had not been utilised often, so we evaluated the methods and how they have been previously successful or can be successfully applied. We also reviewed whether the methods were appropriate and discussed any constraints when applying different methods, such as requiring the same model form. Once the methods had been presented and reviewed the final objective of the project assessed if an improved lung cancer screening tool could be created. Unfortunately, a significantly improved model could not be developed. This was mainly a consequence of the datasets obtained which reported a very high lung cancer incidence rate. Therefore, the models were poorly calibrated, which is the main assessment in the model updating, and therefore the methods were hindered. Additionally, the proposed models would be unrealistic to apply in the real world because they were recalibrated to predict exceptionally high levels of lung cancer risk based on the dataset.

Overall, the majority of the objectives were achieved. The systematic review synthesised the evidence of lung cancer prediction models and highlighted where they had been successful and where further research was required; namely how the models could be successfully applied as a selective screening tool. This became the project's objective and we conducted more extensive research into lung cancer prediction models. We provided clear evidence into the leading model and how this should be applied to maximise potential benefit. The different methods that are available for model updating were presented and analysed. These could then be used in subsequent research to update prediction models for lung cancer or other diseases. We also provided our own suggestion to model aggregation that considered the models' discriminative ability rather than solely aggregating models based off their calibration. This aimed to ensure models with a good discriminative ability were not too harshly penalised and nullified in the final meta-model. In addition the systematic review research has been published and there are plans to publish the external validation results.

Unfortunately, an improved model could not be created, and we decided not to publish a proposed model that did not improve on the existing literature. In addition we could not provide any conclusive evidence into the success of different model updating methods based off our own analysis. This was a consequence of the datasets where the model updating was conducted.

In summary, the work can provide a contribution to the field of lung cancer screening. Previously, multiple lung cancer models had been developed without further research into how the models should be applied. This has hindered models being implemented as a selective screening tool. In this research we demonstrated how all prediction models would benefit from thorough validations that allow constructive results. In addition we have provided clear recommendation into how to apply the $PLCO_{2014}$ Model as a selective screening tool, with recommendations for the next stage of research before hopefully this model being implemented as a screening tool at a regional, national or international level.

12.2 Detailed Results Review

12.2.1 How Lung Cancer Survival can be Improved

The first stage of the project presented the impact of lung cancer. Lung cancer is a global disease affecting millions of people worldwide with the second highest cancer incidence rate in both males and females [1]. In addition to this lung cancer is beset with a poor survival rate which results in a high mortality rate, indeed lung cancer is accountable to around 26% of all cancer related mortalities [1]. The poor long term survival for lung cancer can be attributed to late stage diagnosis. Like all cancers, early diagnosis is critical in improving available treatment options. Specifically for lung cancer the 5-year survival rate for Stage 1 diagnosis is 56%, however by Stage 3 or 4 this has been drastically reduced to 4% [27]. Unfortunately, by this stage the cancer has often metastasised and affected large areas of the lung, so the treatment options are limited to palliative care. It was clear from our initial research into lung cancer; the crux of improving lung cancer survival is to identify cancers at an early stage. However, this is not straightforward with 57% of lung cancer reported at Stage 3 or 4. Indeed, despite research into early diagnosis identification this has only improved from 12% to 18% of lung cancers between 1975 and 2011. This could be attributed to the 'hidden' symptoms of lung cancer as unfortunately the visual lung cancer symptoms, such as coughing blood, do not appear until later stages. Indeed, most early symptoms, including coughing, are masked by a prior smoking history which is prevalent in most lung cancer incidences.

Lung cancer diagnosis could be improved by taking the responsibility away from patients to report symptoms. Instead, people who would be considered likely to develop lung cancer could be placed on a priority list with regular check-ups and potentially periodic screening. This has been observed for other cancers, such as bowel and cervical cancer but has not been implemented for lung cancer. This would be infeasible on a large scale due to the costs associated with CT scanning. Therefore, before large level screening would be considered for lung cancer it needs to be shown that the programme can reliably identify lung cancer in the target population, limit false alarms for incorrect diagnoses, and be cost-effective. This has not been demonstrated at this stage, and in the presented lung cancer screening trials, namely NLST, UKLS, ELCAP and NELSON, they have not been conclusively effective. The main reporting in these trials has been on whether LDCT scanning can be effective over chest radiography. However, there were promising results from the previous screening trials. The results did demonstrate how the identified target population was more effective than blanket screening and therefore resources were correctly allocated. Crucially, lung cancer was diagnosed at early stages, with 66.7% diagnosed at Stage I and a further 19% diagnosed at Stage II in the UKLS trial [36]. Clearly, there is merit in screening a target group periodically. However, at this stage we felt the target group may not be correctly defined. The trials were not successful in restricting false alarms or being cost-effective. Too many people were unnecessarily screened and the trials reported up to $\$269,000$ per QALY [27]. The results indicated further research is required into identifying the correct target population for screening.

After the initial analysis, it became clear prediction models could play a key role in identifying this correct target population. They consider a range of different variables to determine an individual's risk of developing lung cancer over a defined time period. Then a risk can be selected such that anyone who reports a result higher than this threshold could be invited for screening periodically. Therefore, this project decided to review how models and arbitrary criteria could or should be utilised such that a target population is identified for screening that will balance having a positive impact of improving early stage lung cancer diagnosis with the negative effect of unnecessary screening and the negative costs associated with this.

12.2.2 Model Requirements

The next objective of the project was to review what can be expected for prediction models before they would be considered as a clinical utility. The review presented the different techniques available to review a model or screening guideline. These were separated into reviewing the model calibration, how accurately the

model estimated an individual's risk of developing lung cancer; the model discrimination, how successfully the model assigned a higher risk to people who have developed lung cancer; and the model prediction rules, which reviewed the effectiveness of the target population who would be selected for screening using the defined criteria. We presented all the available methods and concluded to review the calibration the Hosmer-Lemeshow and Brier Score should be reported as these have commonly been reported and were effective measures to review the model calibration. To calculate the Hosmer-Lemeshow result the participants need to be grouped, our research identified that these should be grouped into sets of 10 participants with neighbouring risks as generated from the model under consideration. Next, the AUC should be reported as this is an effective measure of the model discrimination. Finally, out of the wide range of methods available to review the model prediction rules the sensitivity, specificity, Youden Index and PLR were selected. The sensitivity and specificity were selected as these are commonly reported and are an easy measure to interpret the effectiveness of the screening criteria. The Youden Index is an extension of the sensitivity and specificity to review how effectively these separate measures work together. Finally, the PLR was selected as this measure will provide an indication into the probability of being considered high risk when you have lung cancer versus the probability of being considered high risk when lung cancer is not present. These methods were ultimately selected because of their analysis value and crucially they were not influenced by the prevalence rate in the dataset. It was anticipated that the datasets recruited for the planned validation would be case-control datasets with a high lung cancer incidence rate that would not be observed in the real world. Other methods such as the PPV could report elevated results suggesting the model had a better performance than would be realistic. Therefore, these methods were not preferred for our planned analysis. Once, all methods had been presented, and reviewed how the results could be interpreted, then the next stage was to review how lung cancer models had performed in previous validations.

12.2.3 The Systematic Review

The systematic review was conducted with the objective of identifying all lung cancer prediction models and their validation results. All models were identified so a critical analysis could be undertaken to identify their potential to be used as a selective screening tool. This was achieved by a review of how the models had performed when validated.

Once the review had concluded it was clear there was a large number of differing models had been developed. In total 29 different models were presented and these were classified into epidemiological, clinical, and TSCE models. Epidemiological models are free to run models so would be preferred if effective in comparison to clinical models which could become expensive when applied to a large volume of people in order to determine if an individual should go for regular screening. Therefore, for clinical models it was imperative to demonstrate an enhanced performance to justify the associated cost applying the model. This was not observed; indeed when a clinical model was produced that expanded on an existing epidemiological model the improvement was negligible. This was observed for the extended Spitz, LLP and Bach models. We concluded from the systematic review that at this stage more research is required into the genetic factors which should be included in clinical models and be able to demonstrate they significantly improve a model's predictive ability. The TSCE models were presented but disregarded as a potential screening tool as they commonly only consider one factor to review this factor's association with lung cancer risk. These would be impractical as a tool to identify populations for screening, particularly since the variable in question was often rare in individuals, such as asbestos or silicone exposure.

The epidemiological models did demonstrate some promising results, these included high AUC results up to 0.86 (PLCO Model). This highlighted these models could be successfully implemented into a screening criteria provided robust guidelines were identified and presented. However, this was not observed in the current reporting of epidemiological models. The validations varied in the results reported and the models were commonly considered in isolation. This made it difficult to infer a leading model from the large volume of models created. Additionally, the prediction rules have been inadequately reported, even when these have been reported in subsequent validations the models had not been tested at the same thresholds

to assess if the results were consistent. Therefore, there was no confidence into a leading model and how best to apply any identified leading model to maximise the impact in a selective screening programme. The results of the systematic review allowed the next stage of the project to develop. The poor reporting seemed to be a key contributor that models are not being implemented as a selective screening tool. Therefore we decided to address this limitation in our research. The models had been identified through the systematic review. The epidemiological models were selected for our research as we would be able to attain the basic patient information required to apply these models. In addition they had shown promise and if a leading model with consistent strong performance as a selective screening tool could be identified then it could be implemented. Since we would not be able to obtain datasets that would contain clinical factors and these models failed to demonstrate an improved performance to justify the cost required to estimate an individual's risk, these were discounted in our research. We recommend as a conclusion of the systematic review, that further research is required to assess whether the cost associated with applying clinical models is beneficial. The results presented in the systematic review currently suggested clinical models may not be more beneficial with the exception of the Korean Men Model which requires fasting glucose levels but reported an exceptional performance. Indeed, further validations would be required, but if this model can consistently demonstrate this level of performance it could be an exceptional tool to identify people for periodic screening.

The TSCE models were also disregarded for the next stage of the research in this project, as these were impractical to apply to identify a high risk population for screening.

By the conclusion of the systematic review we proposed to review all available epidemiological models that had been identified and address the inconsistent reporting and the lack of inter model comparisons. We decided to conduct our own series of external validations that would allow us to deduce which models were successful and how they should be applied to maximise benefit as a selective screening tool, and what would be the expected performance. This hoped to allow medical professionals and decision makers to make informed decisions whether prediction models should be considered as a screening tool.

12.2.4 Dataset Collection and Preparation

To evaluate the models, datasets were collected from ILCCO. This provided the sample populations to review the models and devise recommendations into how leading models should be applied as a screening tool. In total 60 datasets were requested for inclusion in our research. The only prerequisite was that they needed to contain the variables required by the prediction models. From the 60 requests sent there were 20 responses but only 10 datasets had enough information to review at least 2 models, which we felt was the minimum requirement so we could compare between them. Unfortunately, none of the datasets allowed all the models to be validated. An additional limitation was that the Spitz and African-American models could only be applied to one dataset. However, overall the datasets provided a great opportunity to test multiple models in a range of differing sample populations and evaluate their performance.

The datasets were reviewed and prepared. This assessed whether there were any peculiarities in the datasets that needed to be highlighted and in some instances addressed. The main finding was the datasets had a high lung cancer incidence rate that would not be observed in a real world population. In general the datasets had one-to-one matching between participants with or without lung cancer. For the most part this was not a major limitation as it would not influence the model discrimination or the selected prediction rules tests. However, the prevalence rate would influence the model calibration results because at an overall level the models predicted a much lower volume of lung cancer incidence than the 50% that was observed in the datasets. It was decided to still present the calibration results but acknowledge that these may not be a true reflection of the model calibration. In addition the majority of the datasets were case-control designs, this can slightly boost the models' discriminative ability as the cases are often at an advanced stage of lung cancer so may also be higher risk. Therefore the model finds it easier to assign a higher risk to individuals with lung cancer. However, this is negated somewhat by focusing on a comparison between the models; our primary objective was to identify a leading model or criteria that should be considered in preference to other models. This would not be influenced by the case-control dataset design. While we

provided recommendations where the model performed optimally based on our results, when these were presented we recommended assessing if the model continued to perform as strongly in a cohort study using our recommendations. While the model may have a boosted performance this should only be slight and this model would still be the leading criteria in other populations. The other peculiarity of the datasets was the population recruited for the study. The studies had different primary purposes when recruiting participants. In some instances this was to review participants where they could be considered very unlikely or highly likely to develop lung cancer. The ReSoLuCENT and New Zealand datasets recruited young people who may not have been expected to develop lung cancer whereas everyone in the CARET study were heavy smokers and a high proportion had a prior asbestos exposure. This may have made it difficult for the models to distinguish between cases and controls; however this was not a concern as it was still insightful to compare the models in these challenging populations. They still allowed an indication of how the models performed as it is desirable to capture all lung cancer cases, including those belonging to lower risk groups, while it is also beneficial to eliminate unnecessary screening in high risk groups like the CARET study. Overall, while the datasets were different and had some peculiarities, there were no major concerns with the sample populations recruited for our purpose of comparing between multiple models.

The datasets had randomly missing information for some participants across the variables. We reviewed how best to deal with the missing information with our primary objective to retain as much data as possible to thoroughly evaluate the models. However, we highlighted the importance when imputing the missing information this needed to be accurately estimated to fairly evaluate the models. There was caution that if the missing information was inadequately imputed then this would negatively influence the model's performance, particularly if there was missing information imputed for one variable that was required by one model. This would cause an unfair comparison between the models. A review of the available imputation methods was conducted and this allowed us to make informed decisions into the most appropriate approach to deal with the missing information. It was decided that there would always be some uncertainty with imputing missing information, therefore, for datasets with low levels of missing information (under 10%) then the participants with missing information were removed. However, for high levels of missing information this did not seem the most appropriate approach as we would lose too much valuable information. For the participants which reported with missing information they reported information for other variables which could form the imputation. After a review of the imputation methods, MICE was selected as this allowed missing information across multiple variables to be imputed in one imputation. Therefore, the imputation was able to consider whether any information was missing for the other variables used in the imputation and take the uncertainty around the value for these variables into consideration. It also iteratively applied the imputation, using all explanatory and imputed variables, until the imputed values stabilised, therefore giving the most appropriate estimation. We also undertook additional measures to allow us to have uppermost confidence in the imputed values. We conducted testing to demonstrate the missing information was MAR. We also restricted the imputed values once determined, to match to the closest observed value for that variable, such that inappropriate values like a negative CPD value would be corrected. Finally, the imputation was conducted on eleven independent occasions. Eleven was selected as this seemed a reasonably high number that in taking the average across these datasets, to allow for any imputation where the result was peculiar, this should not have a large impact on the model performance. Additionally, Rubin's Rule was applied when calculating the validation results for the AUC, Brier Score, sensitivity and specificity, which would then be used to calculate the PLR. This allowed the most reliable results possible from the imputation taking into account the variance in performance between the datasets. Therefore, we had confidence in the imputed values and the subsequent validation results obtained. The imputation allowed us to maximise the evidence available in the datasets to thoroughly evaluate the models.

12.2.5 External Validation

The next stage of the project used the prepared datasets to validate the models. To provide a thorough evaluation of the models and contextualise how the models should be applied it was decided two validations

should be conducted. The first validation aimed to give context into where the models' consistently reported a robust performance. This would be achieved by identifying a high proportion of lung cancer cases while restricting a reasonable proportion of participants without lung cancer from screening. In this validation the models were applied to their identified target population. This intended to demonstrate how the models could be maximised in a balance between sensitivity and specificity to have a large impact on lung cancer diagnosis. When lung cancer diagnosis options become more efficient with fewer false diagnoses, with less strenuous diagnosis options, and cheaper testing than the current CT screening method, then the results where the models performed optimally should be considered as this would be most beneficial. However, it can be argued that this is not currently appropriate as it would not be cost effective, by periodically screening up to approximately 30% of people who would not subsequently develop lung cancer. Therefore, a second validation was proposed which aimed to limit unnecessary screening to only 10% of people aged between 50-75 years old. This appeared to be a restrained and reasonable cost provided the model could still demonstrate a substantial impact in identifying lung cancer cases. There was no available evidence to assume a 90% specificity would be a feasible cost so this was chosen since in the process of identifying lung cancers there will be some people who are subjected to screening who won't develop lung cancer. Therefore, 10% did not seem an unreasonable number if the screening would capture a high number of lung cancer cases. This was supported by the UKLS trial which also would screen around 10% of lung cancer controls and the NLST criteria which reported a specificity of approximately 80% across the datasets. It was decided to compare the models to the two main screening trials, the NLST and UKLS, as these should still be considered as viable screening criteria in comparison to the models.

The first validation conclusively provided evidence that the $PLCO_{2014}$ Model was the strongest performing model and offered the most beneficial screening guidelines. The AUC result was consistently promising across the sample populations with results between 0.79-0.86. This was commonly the highest AUC result reported in the dataset. Furthermore the model performance was consistent and impressive whether in low risk datasets such as the ReSoLuCENT or New Zealand dataset and the higher risk CARET dataset. Based on the results we were also able to ascertain that to maximise the model performance this should be applied at the 0.5% risk threshold. Here the model should report in most populations a sensitivity of 70% for a specificity of 75%. This is an impressive performance with a Youden Index score above 0.4 uncommon across our validation and the validation results obtained in the systematic review. Clearly if periodically screening 25% of people without lung cancer is realistic then the $PLCO_{2014}$ Model could have a substantial impact in identifying people with lung cancer which would then lead to improved early stage diagnosis and ultimately can directly impact improved survival rates.

The second validation also demonstrated the $PLCO_{2014}$ was the leading performer. Applying this model at the 3% risk threshold would restrict screening of people who do not develop lung cancer to 10% amongst 50-75 year olds. The model would still manage to capture 30% of people who will develop cancer. This is still a high proportion of lung cancer cases and would have a positive impact on the current poor lung cancer prognosis. In addition the model eclipsed the UKLS trial which for no improvement in the specificity, would only identify 20-22% of lung cancer incidences. It should be noted the $PLCO_{2012}$ Model, which only considers ever-smokers, reported the same high standard of performance. This model should be applied at the 3.5% risk threshold.

The results clearly demonstrate the positive impact the $PLCO_{2014}$ Model can have as a selective screening tool. We were able to provide two potential options with a different balance between maximising lung cancer capture rates and restricting unnecessary screening, with clear instructions on how the model should be applied to achieve either objective. In addition, this could maximise the benefits in comparison to the other models and also the current screening criteria which have been considered for implementation. There are additional steps that need to be taken using the results that were obtained in these validations which will be discussed in more detail in the proposal for future work.

While the results obtained were conclusive, unfortunately the Spitz and African-American models could not be thoroughly evaluated as they were only applicable to one dataset. These should be evaluated in other sample populations, but in doing so, it would be our recommendation that these models were compared

to the $PLCO_{2014}$ Model for whichever of the two guidelines we recommended best match the validation objectives. Currently the $PLCO_{2014}$ Model could not be compared on the same dataset as these two models so a direct comparison could not be conducted.

12.2.6 Literature Review

The next stage of the project reviewed whether we could create an improved model that combined the ability of the leading $PLCO_{2014}$ Model or multiple models, and the additional evidence available in the external dataset. This required a literature review to be conducted to present the methods that are available and review how they have been applied previously to give an indication into how different methods are successful in different scenarios. The methods were also reviewed for their practicality in being applied to the identified lung cancer models. All single updating methods were applied to the $PLCO_{2014}$ Model as this was the leading original model so any improvements would create a new leading screening tool. Model aggregating was initially restricted to the $PLCO_{2012}$, $PLCO_{2014}$, Bach, LLP, Hoggart and Pittsburgh models as these could be applied in the same dataset which would serve as the external data to form the model updating.

The literature review identified multiple methods to update a single model and aggregate multiple models. The single updating methods were presented as model recalibration, re-estimation, and model extension methods. Model recalibration takes the existing model but recalibrates the parameters to reflect the observed incidence rate in the external dataset. The re-estimation methods re-estimate the model parameters, with focus on parameters that do not significantly improve the model goodness-of-fit. These parameter coefficients are reduced in the model such that an individual's risk cannot deviate as far as in the original model using this parameter. Finally, model extension methods were presented, which added new variables to the model. The review highlighted where the different methods could be successful. The general conclusions were that model recalibration should be applied to a successful model, particularly if the model discrimination is good, such that the model is slightly adjusted to predict accurate risks in the model dataset. Re-estimation can be successful for model with a calibration and discrimination deficiency, and is more successful when the dataset reviewing the model is small. Adding new parameters could result in over-fitting therefore readdressing weaker variables in the model would not allow these variables such as large an impact in the final model. Finally, model extension is more appropriate for a model with poor calibration and discrimination and a large external dataset is available. This will allow confidence in the variables that are added to the model are correct and do not cause over-fitting. For all methods, the model is updated based on the model calibration therefore, a well calibrated model will not require extensive updating.

The model aggregation methods were also presented in the review. After the literature review, Model Averaging and BMA were identified as appropriate methods to use on the lung cancer models. The other methods were discounted as they either did not use an external dataset to conduct the model aggregation or required the parameters in the models to be combined into one parameter. This required the models to have the majority of similar parameters and a same model type. This was not observed across the lung cancer models. Model Averaging combined the models by assigning them a weight based on their calibration and is an excellent method to allow distinct models to be combined. However, the method initially recalibrates the models; therefore if the external dataset has any peculiarities such as a high prevalence rate then this will be reflected in the final model. Additionally, this method heavily weights one model therefore, the evidence available from the remaining models is somewhat negated by being assigned a small weighting. The second method, BMA, is also a good method to combine distinct models. Like Model Averaging this method combines the models based on their calibration, which may result in a model that is poorly calibrated but a good screening tool being removed from the final model. Therefore, we proposed assigning an additional weighting to each model which considers their AUC performance to ensure any models that have a superior AUC result are not disregarded in the final model.

At the end of the literature review the model updating methods were presented and the methods that were appropriate for lung cancer models identified. The $PLCO_{2014}$ Model was selected to conduct

single model updating. The $PLCO_{2014}$, Pittsburgh, Hoggart and Bach models were available for model aggregation. After deliberation, two version of model aggregation would be conducted, with and without the Bach Model. This was chosen as including the Bach Model allowed us to use the evidence in the model rather than automatically disregarding this model. Excluding this model though, allowed a larger dataset to be used for model aggregation and the final model to be applicable to a larger range of people as per the Bach Model was more restrictive to heavy ever-smokers.

12.2.7 Model Updating

The single model updating methods did not create an improved lung cancer model. The main limitation was that the datasets used to conduct the model updating had a very high lung cancer incidence rate. As a consequence when the model was recalibrated the proposed model predicted very high risks for an individual's developing lung cancer within the next 6-years. In some instances the model predicted risks exceeding 99%. This could not be used by the public as inappropriate risks like this could be detrimental by creating panic. In addition to this, the methods did not substantially improve the model discrimination or prediction rules in comparison to the original model. Therefore, it would be unwise to promote any of the updated models as they would add to the large volume of published lung cancer models without improving upon the existing literature.

The model aggregation methods also underperformed due to the high lung cancer prevalence rates in the dataset. The models were poorly calibrated so the methods found it difficult to assign appropriate weights to each model. Despite this a model with a reasonable performance was presented, this was the aggregation of the Hoggart, $PLCO_{2014}$ and Pittsburgh models using BMA with an additional weighting based on the models' AUC results in the external dataset. This slightly improved upon the original models although would have to be further evaluated to assess if the performance can be replicated. The separate model weightings were presented in the thesis with recommendations to ascertain if the model can replicate the impressive sensitivity of 83% and specificity of 36-42% at the 1.5% risk threshold. In addition the model should be reviewed at a high risk threshold, between 13.8-15.8% to allow the specificity to remain at approximately 90% and then review the sensitivity. However, there is the possibility that the proposed model may be inappropriate in new environments with lower prevalence rates, in comparison to the original models.

12.3 Project Strengths

Overall the systematic review went well and achieved the objectives of the project. All available lung cancer prediction models were identified and critically reviewed. We highlighted how the current reporting of validations was inadequate but could be improved by comparing multiple models together and using consistent testing. The systematic review was not able to identify a leading model that could be considered further as a selective screening tool therefore, this became the new project objective. The systematic review could have an impact how models are reported and validated to facilitate models being considered as a screening tool. The results were also published in an article.

The external validation presented comprehensive results which provided a better indication into each model performance and demonstrated how the $PLCO_{2014}$ Model was the leading performer. It also provided clear recommendations how to apply this model to best achieve the objectives of a screening programme, whether this be to identify a high number of lung cancer cases or to reduce screening of people who would not develop lung cancer. Based off the success of the external validation there are plans to present these results in an article to be used to plan future research and potentially implement the model in a selective screening trial.

The literature review of model updating and model aggregation had some successes. This identified and presented all available methods to update a single model or aggregate multiple models as per the literature review objectives. The also presented where the different updating methods have been previously successful

for different original model deficiencies. The critiqued methods can then be applied in future research to improve the performance of prediction models. For instance model extension could be a key method as new markers that are predictive of lung cancer incidence are identified and added to existing successful models.

12.4 Limitations and Different Approaches

There were some aspects of this project that had limitations. This primarily stemmed from the datasets collected for our research which had a high lung cancer prevalence rate. This affected some aspects of the external validation and the model updating. The models were not able to predict the observed incidence rate. Therefore, it was difficult to truly evaluate the model calibration in the external validation. However, we acknowledged this during the validations and the other testing conducted was not influenced by the dataset prevalence rate so we could provide constructive results.

This also had an impact on the single model updating review using lung cancer prediction models. The model updating methods were based upon the model calibration. Therefore, the poor original model calibration meant the model was dramatically recalibrated. This was particularly noted when extending the model as the new parameters were extreme to try and predict a much higher lung cancer incidence in an attempt to improve the model goodness of fit. Some updated models predicted extremely high lung cancer incidence rates and would be inappropriate as a public tool as a consequence. In addition when conducting model aggregation the poor original models' calibration forced the final meta-model to assign extreme weights to some models while negating the other models. This made it difficult to critically appraisal the methods as they were heavily impacted by the datasets. However, while we could not demonstrate the successes and limitations of the methods using lung cancer models we were still provided critical analysis of the techniques in the literature review. This was based on their performance in published material when they had been applied to update models in more appropriate datasets.

Unfortunately the limitations with the datasets could not be rectified. We had to collect datasets to conduct the project within the time constraints rather than collating our own data. If more datasets had been received then these could have been more selectively included but this was not appropriate with the volume of data we received.

Overall, the datasets were the main limitation in this project however, it would have been impractical to correct this by collecting our data with the project time constraints. To address this concern the project focus adapted to compare models and review their potential as a selective screening tool which was not impacted by the high lung cancer prevalence rate. Tests were also selected when reviewing the prediction rules that were unaffected by the datasets.

12.5 Future Research

At the conclusion of our research we provided suggestions for future research that would be required before the $PLCO_{2014}$ Model would be considered as a selective screening tool.

The first focus should validate the $PLCO_{2014}$ Model to assess if the level of performance can be replicated at the risk thresholds presented. The validation could also review how the model performs in different environments, this can assess whether the model would be universally successful or if more research is required for specific populations. If the model demonstrates the same level of performance as observed in the external validation then the next stage of research needs to evaluate the model in a screening trial. This can review how many lung cancers would be identified and the volume of screening required to identify a lung cancer case. This can also determine how frequently someone should be screened if their risk is above the threshold. It should also review how periodically someone's risk should be calculated to assess if their risk varies across the risk threshold. Finally the screening trial should report at what stage the lung cancers were diagnosed and whether this would impact the current poor lung cancer survival. This research

will demonstrate whether the model would have a positive impact and be a viable screening programme. Based off the conclusion of the screening trial then the $PLCO_{2014}$ Model could be implemented.

Other research should review the level of performance of the Spitz and African-American models as these could only be considered in one dataset. These should be compared where possible to the $PLCO_{2014}$ Model to identify if they would offer any improvement.

12.6 Can Prediction Models be Successful Selective Screening Tools?

Before models would be considered they need to be thoroughly evaluated. There needs to be clear recommendations how to apply the model, with an expectation how the model would be perform. Specifically for lung cancer prediction models this had not been adequately reported and has hindered models being implemented as a screening tool.

The research in this project demonstrated how prediction models can have a positive impact. The models demonstrated the best option to identify a target population that would benefit from screening. Therefore, provided model reporting is improved and the focus shifts towards implementing the existing successful models rather than developing new models, then lung cancer prediction models can be a successful screening tool.

Part III

References

References

Bibliography

[1] Siegel RL, Miller KD, Jemal A. Cancer Statistics, 2016. Ca-a Cancer Journal for Clinicians 2016;66(1):7-30.

[2] National Cancer Institute. (2016). What Is Cancer? [online]Availableat:http://www.cancer.gov/about-cancer/understanding/what-is-cancer [Accessed 28 Jun. 2016].

[3] Nhs.uk. (2016). Lung cancer - NHS Choices. [online] Available at: http://www.nhs.uk/conditions/cancer-of-the-lung/Pages/Introduction.aspx [Accessed 28 Jun. 2016].

[4] Anon, (2016). Risk Factors for Lung Cancer: A Systematic Review. [online] Available at: https://canceraustralia.gov.au/sites/default/files/publications/risk-factors-lung-cancer-systematic-review/rtf/risk_factors_for_lung_cancer_-_a_systematic_review_-_wcag_version.docx [Accessed 27 Jun. 2016].

[5] Parkin DM. 2. Tobacco-attributable cancer burden in the UK in 2010. Br J Cancer 2011;105(S2):S6-S13.

[6] Cancer Research UK. (2015). Lung cancer risk factors. [online] Available at: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/risk-factors [Accessed 27 Jun. 2016].

[7] World Health Organization. (2016). Tobacco. [online] Available at: http://www.who.int/mediacentre/factsheets/fs339/en/ [Accessed 6 Jul. 2016].

[8] Parkin DM. 14. Cancers attributable to occupational exposures in the UK in 2010. Br J Cancer 2011;105(S2):S70-S72.

[9] Sin DD, Tammemagi CM, Lam S, et al. Pro-Surfactant Protein B As a Biomarker for Lung Cancer Prediction. Journal of Clinical Oncology 2013;31(36):4536-+.

[10] Sozzi G, Boeri M, Rossi M, et al. Clinical Utility of a Plasma-Based miRNA Signature Classifier Within Computed Tomography Lung Cancer Screening: A Correlative MILD Trial Study. Journal of Clinical Oncology 2014;32(8):768-+.

[11] Boeri M, Verri C, Conte D, et al. MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. Proceedings of the National Academy of Sciences of the United States of America 2011;108(9):3713-3718.

[12] Wozniak MB, Scelo G, Muller DC, et al. Circulating MicroRNAs as Non-Invasive Biomarkers for Early Detection of Non-Small-Cell Lung Cancer. Plos One 2015;10(5).

[13] Bediaga NG, Davies MPA, Acha-Sagredo A, et al. A microRNA-based prediction algorithm for diagnosis of non-small lung cell carcinoma in minimal biopsy material. British Journal of Cancer 2013;109(9):2404-2411.

[14] El-Zein RA, Lopez MS, D'Amelio AM, Jr., et al. The Cytokinesis-Blocked Micronucleus Assay as a Strong Predictor of Lung Cancer: Extension of a Lung Cancer Risk Prediction Model. Cancer Epidemiology Biomarkers & Prevention 2014;23(11):2462-2470.

[15] Brenner DR, McLaughlin JR, Hung RJ. Previous Lung Diseases and Lung Cancer Risk: A Systematic Review and Meta-Analysis. PLoS ONE 2011;6(3):e17479.

[16] Brenner DR, Boffetta P, Duell EJ, et al. Previous lung diseases and lung cancer risk: a pooled analysis from the International Lung Cancer Consortium. Am J Epidemiol 2012; 176(7):573-85.

[17] Merrill, R. (2000). Measuring the Projected Public Health Impact of Lung Cancer Through Lifetime and Age-Conditional Risk Estimates. Annals of Epidemiology, 10(2), pp.88-96.

[18] American Joint Committee on Cancer. (2016). Lung Cancer Staging. [online] Available at: http://cancerstaging.org/references-tools/quickreferences/documents/lungmedium.pdf [Accessed 27 Jun. 2016].

[19] Cancer Research UK. (2016). More about staging for lung cancer — Cancer Research UK. [online] Cancerresearchuk.org. Available at: http://www.cancerresearchuk.org/about-cancer/type/lung-cancer/treatment/more-about-lung-cancer-staging [Accessed 27 Jun. 2016].

[20] Cancer Research UK. (2016). Diagnosing lung cancer — Cancer Research UK. [online] Available at: http://www.cancerresearchuk.org/about-cancer/type/lung-cancer/diagnosis/ [Accessed 28 Jun. 2016].

[21] Cancer Research UK. (2016). Types of treatment for lung cancer — Cancer Research UK. [online] Available at: http://www.cancerresearchuk.org/about-cancer/type/lung-cancer/treatment/which-treatment-for-lung-cancer [Accessed 28 Jun. 2016].

[22] Nhs.uk. (2016). Lung cancer - NHS Choices. [online] Available at: http://www.nhs.uk/conditions/cancer-of-the-lung/Pages/Introduction.aspx [Accessed 28 Jun. 2016].

[23] Black WC (2007) Computed tomography screening for lung cancer: review of screening principles and update on current status. Cancer 110(11):2370–2384

[24] McWilliams A, Lam S (2005) Lung cancer screening. Current Opinion Pulmonary Medicine 11(4):272–277

[25] Jonnalagadda S, Bergamo C, Lin JJ, Lurslurchachai L, Diefenbach M, Smith C, Nelson JE, Wisnivesky JP (2012) Beliefs and attitudes about lung cancer screening among smokers. Lung Cancer 77:526–531

[26] Aberle DR, Adams AM, Berg CD, et al. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. New England Journal of Medicine 2011;365(5):395-409.

[27] Tota JE, Ramanakumar AV, Franco EL. Lung Cancer Screening: Review and Performance Comparison Under Different Risk Scenarios. Lung 2014;192(1):55-63.

[28] Kramer BS, Berg CD, Aberle DR, Prorok PC. Lung cancer screening with low-dose helical CT: results from the National Lung Screening Trial (NLST). J Med Screen 2011;18:109–11.

[29] Goulart, Bernardo. "Lung Cancer CT Screening Is Cost-Effective But Implementation Matters". Evidence Based Medicine 20.2 (2015): 78-78. Web.

[30] Neumann, Peter J., Joshua T. Cohen, and Milton C. Weinstein. "Updating Cost-Effectiveness — The Curious Resilience Of The $50,000-Per-QALY Threshold". New England Journal of Medicine 371.9 (2014): 796-797. Web.

[31] Cancer Research UK. (2015). Lung cancer incidence statistics. [online] Available at: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/lung-cancer/incidence#heading-Nine [Accessed 5 Jul. 2016].

[32] "Lung CT Screening Reporting And Data System". American College of Radiology. N.p., 2017. Web. 15 Feb. 2017.

[33] Doria-Rose VP, Szabo E. Screening and prevention of lung cancer. In: Kernstine KH, Reckamp KL, eds. Lung cancer: a multidisciplinary approach to diagnosis and management. New York: Demos Medical Publishing, 2010:53-72.

[34] McRonald FE, Yadegarfar G, Baldwin DR, et al. The UK Lung Screen (UKLS): Demographic Profile of First 88,897 Approaches Provides Recommendations for Population Screening. Cancer Prevention Research 2014;7(3):362-371.

[35] Field JK, Duffy SW. Lung cancer screening: the way forward. Br J Cancer 2008; 99: 557–62

[36] Field, J. Duffy, S. et al. (2015). UK Lung Cancer RCT Pilot Screening Trial: baseline findings from the screening arm provide evidence for the potential implementation of lung cancer screening. Thorax, 71(2), pp.161-170.

[37] Black, William. "Cost-Effectiveness Of CT Screening In The National Lung Screening Trial". New England Journal of Medicine 372.4 (2015): 387-388. Web.

[38] Ebell, Mark. "Poems: Lung Cancer Screening Is Cost-Effective, But Only If Done Correctly - American Family Physician". Aafp.org. N.p., 2017. Web. 15 Feb. 2017.

[39] Scotland, G S et al. "Cost-Effectiveness Of Implementing Automated Grading Within The National Screening Programme For Diabetic Retinopathy In Scotland". British Journal of Ophthalmology 91.11 (2007): 1518-1523. Web.

[40] Ru Zhao Y, Xie X, de Koning HJ, Mali WP, Vliegenthart R, Oudkerk M. NELSON lung cancer screening study. Cancer Imaging 2011;11 Spec No A:S79–84.

[41] Wisnivesky JP, Mushlin AI, Sicherman N, et al. The cost-effectiveness of low-dose CT screening for lung cancer - Preliminary results of baseline screening. Chest 2003;124(2):614-621.

[42] Henschke CI, McCauley DI, Yankelevitz DF, et al. Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 1999; 354:99-105

[43] Henschke CI, Naidich DP, Yankelevitz DF, et al. Early lung cancer action project. Cancer 2001; 92:153-159

[44] Moons KGM, Kengne AP, Woodward M, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012;98(9):683-690

[45] Moons KG, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:1317–20.

[46] Moons, Karel G.M. et al. "Transparent Reporting Of A Multivariable Prediction Model For Individual Prognosis Or Diagnosis (TRIPOD): Explanation And Elaboration". Annals of Internal Medicine 162.1 (2015): W1. Web.

[47] Cancer Research UK. (2015). Screening for cancer. [online] Available at: http://www.cancerresearchuk.org/about-cancer/screening [Accessed 23 Jun. 2016].

[48] www2.le.ac.uk. (2016). Groundbreaking lung cancer breath test in clinical trial — University of Leicester. [online] Available at: http://www2.le.ac.uk/offices/press/press-releases/2015/february/groundbreaking-lung-cancer-breath-test-in-clinical-trial [Accessed 23 Jun. 2016].

[49] Cancer.gov. (2016). Breast Cancer Risk Assessment Tool. [online] Available at: `http://www.cancer.gov/bcrisktool/` [Accessed 23 Jun. 2016].

[50] Vogel VG. Management of the high-risk patient. Surg Clin North Am. 2003;83:733–751.

[51] Steyerberg, E. (2009). Clinical prediction models. New York: Springer

[52] Bang, H. (2016). Building and Using Disease Prediction Models in the Real World. [online] `Hpr.weill.cornell.edu.` Available at: `http://hpr.weill.cornell.edu/divisions/biostatistics/ppt/Roundtable_JSM_20071.ppt` [Accessed 23 Jun. 2016].

[53] Steyerberg, E., Bleeker, S., Moll, H., Grobbee, D. and Moons, K. (2003). Internal and external validation of predictive models: A simulation study of bias and precision in small samples. Journal of Clinical Epidemiology, 56(5), pp.441-447.

[54] HARRELL, F., LEE, K. and MARK, D. (1996). MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. Statist. Med., 15(4), pp.361-387.

[55] Efron B, Gong G. A LEISURELY LOOK AT THE BOOTSTRAP, THE JACKKNIFE, AND CROSS-VALIDATION. American Statistician. 1983;37(1):36-48.

[56] Paul, P., Pennell, M. and Lemeshow, S. (2012). Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets. Statist. Med., 32(1), pp.67-80.

[57] Steyerberg, E., Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. and Kattan, M. (2010). Assessing the Performance of Prediction Models. Epidemiology, 21(1), pp.128-138.

[58] Toll, D., Janssen, K., Vergouwe, Y. and Moons, K.(2008).Validation, updating and impact of clinical prediction rules: A review. Journal of Clinical Epidemiology, 61(11), pp.1085-1094.

[59] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. British Medical Journal. 2009;338.

[60] Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. British Medical Journal. 2009;338

[61] Moons, K., Kengne, A., Grobbee, D., Royston, P., Vergouwe, Y., Altman, D. and Woodward, M. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. Heart, 98(9), pp.691-698.

[62] Youden, W. (1950). Index for rating diagnostic tests. Cancer, 3(1), pp.32-35.

[63] Bohning, D., Bohning, W. and Holling, H. (2008). Revisiting youden's index as a useful measure of the misclassification error in meta-analysis of diagnostic studies. Statistical Methods in Medical Research, 17(6), pp.543-554

[64] Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G. and Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol, 56(1), p.45.

[65] McGee, S. (2002). Simplifying likelihood ratios. J Gen Intern Med, 17(8), pp.647-650.

[66] Alberg, A., Park, J., Hager, B., Brock, M. and Diener-West, M. (2004). The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J Gen Intern Med, 19(5), pp.460-465.

[67] Gray, Eoin P. et al. "Risk Prediction Models For Lung Cancer: A Systematic Review". Clinical Lung Cancer 17.2 (2016): 95-106. Web.

[68] Medicine USNLo. Medical Subject Headings 1999. 1]. Available from:`https://www.nlm.nih.gov/mesh/`.

[69] D'Amelio AM, Jr., Cassidy A, Asomaning K, Raji OY, Duffy SW, Field JK, et al. Comparison of discriminatory power and accuracy of three lung cancer risk models. British Journal of Cancer. 2010;103(3):423-9.

[70] Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms. International Journal of Digitial Libraries. 2000;3(2):117 - 32.

[71] (NaCTeM) TNCfTM. TerMine 2013 [cited 2014 27/02/2014]. Available from: `http://www.nactem.ac.uk/` Available from: `http://www.nactem.ac.uk/software/termine/`.

[72] Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. British Journal of Cancer. 2008;98(2):270-6.

[73] Raji OY, Duffy SW, Agbaje OF, Baker SG, Christiani DC, Cassidy A, et al. Predictive Accuracy of the Liverpool Lung Project Risk Model for Stratifying Patients for Computed Tomography Screening for Lung Cancer A Case-Control and Cohort Validation Study. Annals of Internal Medicine. 2012;157(4):242-+.

[74] Raji OY, Agbaje OF, Duffy SW, Cassidy A, Field JK. Incorporation of a Genetic Factor into an Epidemiologic Model for Prediction of Individual Risk of Lung Cancer: The Liverpool Lung Project. Cancer Prevention Research. 2010;3(5):664-9.

[75] Spitz MR, Hong WK, Amos CI, Wu XF, Schabath MB, Dong Q, et al. A risk model for prediction of lung cancer. Journal of the National Cancer Institute. 2007;99(9):715-26.

[76] Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, et al. An Expanded Risk Prediction Model for Lung Cancer. Cancer Prevention Research. 2008;1(4):250-4.

[77] Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in lung cancer risk among smokers. Journal of the National Cancer Institute. 2003;95(6):470-8.

[78] Cronin, K., Gail, M., Zou, Z., Bach, P., Virtamo, J. and Albanes, D. (2006). Validation of a Model of Lung Cancer Risk Prediction Among Smokers. JNCI Journal of the National Cancer Institute, 98(9), pp.637-640.

[79] Hoggart C, Brennan P, Tjonneland A, Vogel U, Overvad K, Ostergaard JN, et al. A Risk Model for Lung Cancer Incidence. Cancer Prevention Research. 2012;5(6):834-46.

[80] Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung Cancer Risk Prediction: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Models and Validation. Jnci-Journal of the National Cancer Institute. 2011;103(13):1058-68.

[81] Selection Criteria for Lung-Cancer Screening. (2013). New England Journal of Medicine, 369(4), pp.394-394.

[82] Tammemagi MC, Church TR, Hocking WG, Silvestri GA, Kvale PA, Riley TL, et al. Evaluation of the Lung Cancer Risks at Which to Screen Ever- and Never-Smokers: Screening Rules Applied to the PLCO and NLST Cohorts. PLoS medicine. 2014;11(12):e1001764-e.

[83] Moyer VA, Force USPST. Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine. 2014;160(5):330-+.

[84] Etzel CJ, Kachroo S, Liu M, D'Amelio A, Dong Q, Cote ML, et al. Development and Validation of a Lung Cancer Risk Prediction Model for African-Americans. Cancer Prevention Research. 2008;1(4):255-65.

[85] Wilson, D. and Weissfeld, J. (2015). A simple model for predicting lung cancer occurrence in a lung cancer screening program: The Pittsburgh Predictor. Lung Cancer, 89(1), pp.31-37.

[86] Tammemagi MC, Lam SC, McWilliams AM, Sin DD. Incremental Value of Pulmonary Function and Sputum DNA Image Cytometry in Lung Cancer Risk Prediction. Cancer Prevention Research. 2011;4(4):552-61

[87] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996;15(4):361-87.

[88] Nam B-H, Shin H-R, Lee JS, Yang H-R, Lee JA, Han JT, et al. Individualized Risk Prediction Model for Lung Cancer in Korean Men. Figshare. 2013.

[89] Maisonneuve P, Bagnardi V, Bellomi M, Spaggiari L, Pelosi G, Rampinelli C, et al. Lung Cancer Risk Prediction to Select Smokers for Screening CT-a Model Based on the Italian COSMOS Trial. Cancer Prevention Research. 2011;4(11):1778-89.

[90] Veronesi, G., Maisonneuve, P., Rampinelli, C., Bertolotti, R., Petrella, F., Spaggiari, L. and Bellomi, M. (2013). Computed tomography screening for lung cancer: Results of ten years of annual screening and validation of cosmos prediction model. Lung Cancer, 82(3), pp.426-430.

[91] Young RP, Hopkins RJ, Hay BA, Epton MJ, Mills GD, Black PN, et al. A gene-based risk score for lung cancer susceptibility in smokers and ex-smokers. Postgraduate Medical Journal. 2009;85(1008):515-24.

[92] Li H, Yang L, Zhao X, Wang J, Qian J, Chen H, et al. Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model. Bmc Medical Genetics. 2012;13.

[93] Zeka A, Gore R, Kriebel D. The two-stage clonal expansion model in occupational cancer epidemiology: results from three cohort studies. Occupational and Environmental Medicine. 2011;68(8):618-24

[94] Heidenreich WF, Wellmann J, Jacob P, Wichmann HE. Mechanistic modelling in large case-control studies of lung cancer risk from smoking. Statistics in Medicine. 2002;21(20):3055-70.

[95] Meza R, Hazelton WD, Colditz GA, et al. Analysis of lung cancer incidence in the nurses' health and the health professionals' follow-up studies using a multistage carcinogenesis model. Cancer Causes & Control 2008;19(3):317-328.

[96] Foy M, Spitz MR, Kimmel M, et al. A smoking-based carcinogenesis model for lung cancer risk prediction. International Journal of Cancer 2011;129(8):1907-1913.

[97] Foy M, Deng L, Spitz M, et al. Rice-MD Anderson Lung Cancer Model. Risk Analysis 2012;32:S142-S150.

[98] Schultz FW, Boer R, de Koning HJ. Description of MISCAN-Lung, the Erasmus MC Lung Cancer Microsimulation Model for Evaluating Cancer Control Interventions. Risk Analysis 2012;32:S85-S98.

[99] Deng L, Kimmel M, Foy M, et al. Estimation of the effects of smoking and DNA repair capacity on coefficients of a carcinogenesis model for lung cancer. International Journal of Cancer 2009;124(9):2152-2158.

[100] Akushevich I, Veremeyeva G, Kravchenko J, et al. New stochastic carcinogenesis model with covariates: An approach involving intracellular barrier mechanisms. Mathematical Biosciences 2012;236(1):16-30.

[101] Hazelton WD, Jeon J, Meza R, et al. The FHCRC Lung Cancer Model. Risk Analysis 2012;32:S99-S116.

[102] Cancer Research UK (2015). Your cancer type — Cancer Research UK. [online] Cancerresearchuk.org. Available at: `http://www.cancerresearchuk.org/about-cancer/type/` [Accessed 6 Apr. 2015].

[103] Allison, P., 2001. Missing data — Quantitative applications in the social sciences. Thousand Oaks, CA: Sage. Vol. 136.

[104] Little, R. J. A. (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," Journal of the American Statistical Association, 88, 125–134.

[105] Molenberghs, G. and Kenward, M. G. (2007), Missing Data in Clinical Studies, New York: John Wiley & Sons.

[106] Nakai M, Weiming Ke. Review of Methods for Handling Missing Data in Longitudinal Data Analysis. Int. Journal of Math. Analysis. 2011;5(1):1-13.

[107] Farhangfar, A., Kurgan, L. and Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, 41(12), pp.3692-3705.

[108] Missing-data imputation, (2015). Missing-data imputation. [online] Available at: `http://www.stat.columbia.edu/~gelman/arm/missing.pdf` [Accessed 7 Apr. 2015].

[109] Missing Data, (2015). Missing Data & How to Deal: An overview of missing data. [online] Available at: `https://www.utexas.edu/cola/prc/_files/cs/Missing-Data.pdf` [Accessed 5 Jun. 2015].

[110] White, Ian R., Patrick Royston, and Angela M. Wood. "Multiple Imputation Using Chained Equations: Issues And Guidance For Practice". Statistics in Medicine 30.4 (2010): 377-399. Web.

[111] White, Ian R. and John B. Carlin. "Bias And Efficiency Of Multiple Imputation Compared With Complete-Case Analysis For Missing Covariate Values". Statistics in Medicine 29.28 (2010): 2920-2931. Web.

[112] Horton, Nicholas J and Ken P Kleinman. "Much Ado About Nothing". The American Statistician 61.1 (2007): 79-90. Web.

[113] White, I., Royston, P. and Wood, A. (2010). Multiple imputation using chained equations: Issues and guidance for practice. Statist. Med., 30(4), pp.377-399.

[114] Lee, K. (2015). Regression Modelling with Missing Data: Principles, Methods, Software, and Examples. In: 36th Annual Conference of the International Society for Clinical Biostatistics. Utrecht: ISCB 2015.

[115] Allison, P., 2001. Missing data — Quantitative applications in the social sciences. Thousand Oaks, CA: Sage. Vol. 136.

[116] Nakai M, Weiming Ke. Review of Methods for Handling Missing Data in Longitudinal Data Analysis. Int. Journal of Math. Analysis. 2011;5(1):1-13.

[117] D'Amelio, A., Cassidy, A., Asomaning, K., Raji, O., Duffy, S., Field, J., Spitz, M., Christiani, D. and Etzel, C. (2010). Comparison of discriminatory power and accuracy of three lung cancer risk models. Br J Cancer, 103(3), pp.423-429.

[118] Hanley, J. and McNeil, B. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148(3), pp.839-843.

[119] Comparing Two Diagnostic Tests. (2016). Design and Analysis of Clinical Trials. [online] Available at: `https://onlinecourses.science.psu.edu/stat509/node/152` [Accessed 1 Aug. 2016].

[120] Song, J. and Chung, K. (2010). Observational Studies: Cohort and Case-Control Studies. Plastic and Reconstructive Surgery, 126(6), pp.2234-2242.

[121] Evans, J. and Macdonald, T. (1997). Misclassification and selection bias in case-control studies using an automated database. Pharmacoepidemiology and Drug Safety, 6(5), pp.313-318.

[122] Cms.gov, (2015). ICD-9 Code Lookup. [online] Available at: http://www.cms.gov/medicare-coverage-database/staticpages/icd-9-code-lookup.aspx [Accessed 10 Apr. 2015].

[123] Cancer Research UK, (2015). A study to find out more about the causes of lung cancer (ReSoLuCENT). [online] Available at: http://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/a-study-to-find-out-more-about-the-causes-of-lung-cancer [Accessed 29 Sep. 2015].

[124] Resolucent.group.shef.ac.uk, (2015). ReSoLuCENT Resource for the Study of Lung Cancer Epidemiology in North Trent. [online] Available at: http://resolucent.group.shef.ac.uk/ [Accessed 29 Sep. 2015].

[125] Hashibe, M., Morgenstern, H., Cui, Y., Tashkin, D., Zhang, Z., Cozen, W., Mack, T. and Greenland, S. (2006). Marijuana Use and the Risk of Lung and Upper Aerodigestive Tract Cancers: Results of a Population-Based Case-Control Study. Cancer Epidemiology Biomarkers & Prevention, 15(10), pp.1829-1834.

[126] The β-carotene and retinol efficacy trial (CARET) for chemoprevention of lung cancer in high risk populations: Smokers and asbestos-exposed workers. (1994). Lung Cancer, 11(5-6), p.423.

[127] Muscat, J., Stellman, S. and Wynder, E. (1995). Insulation, asbestos, smoking habits, and lung cancer cell types. Am. J. Ind. Med., 27(2), pp.257-269.

[128] Epi.grants.cancer.gov, (2015). Singapore Chinese Health Study. [online] Available at: http://epi.grants.cancer.gov/Consortia/members/singapore.html [Accessed 29 Sep. 2015].

[129] Niehs.nih.gov, (2015). Singapore Chinese Health. [online] Available at: http://www.niehs.nih.gov/research/atniehs/labs/epi/studies/singapore/ [Accessed 29 Sep. 2015].

[130] Aldington, S., Harwood, M., Cox, B., Weatherall, M., Beckert, L., Hansell, A., Pritchard, A., Robinson, G. and Beasley, R. (2008). Cannabis use and risk of lung cancer: a case-control study. European Respiratory Journal, 31(2), pp.280-286.

[131] Donatella, U., Monica, N., Aldo, C., Cristina, C., Giuseppe, C., Paolo, I., Cecilia, L., Paola, M., Michela, P., Barbara, P., Paola, V., Riccardo, P. and Stefano, B. (2008). The CREST Biorepository: A Tool for Molecular Epidemiology and Translational Studies on Malignant Mesothelioma, Lung Cancer, and Other Respiratory Tract Diseases. Cancer Epidemiology Biomarkers & Prevention, 17(11), pp.3013-3019.

[132] Rennert, G., Kremer, R., Rennert, H., Wollner, M., Agbarya, A., Pinchev, M., Lejbkowicz, F., Spitz, M. and Muscat, J. (2015). Lower lung cancer rates in Jewish smokers in Israel and the USA. International Journal of Cancer, 137(9), pp.2155-2162.

[133] Wang, H., Rothenbacher, D., Löw, M., Stegmaier, C., Brenner, H. and Diepgen, T. (2006). Atopic diseases, immunoglobulin E and risk of cancer of the prostate, breast, lung and colorectum. International Journal of Cancer, 119(3), pp.695-701.

[134] Lunenfeld.ca, (2015). Dr. Rayjean Hung — The Lunenfeld-Tanenbaum Research Institute. [online] Available at: http://www.lunenfeld.ca/researchers/hung [Accessed 29 Sep. 2015].

[135] Zhang, L., Morgenstern, H., Greenland, S., Chang, S., Lazarus, P., Teare, M., Woll, P., Orlow, I., Cox, B., Brhane, Y., Liu, G. and Hung, R. (2014). Cannabis smoking and lung cancer risk: Pooled analysis in the International Lung Cancer Consortium. International Journal of Cancer, 136(4), pp.894-903.

[136] Centerwatch.com. (2016). A clinical research study of Selumetinib, Pemetrexed and Cisplatin for the treatment of Non-Small Cell Lung Cancer, Non-Squamous — Clinical Research Trial Listing in Toronto, Canada (Non-Small Cell Lung Cancer) (NCT02337530). [online] Available at: `https://www.centerwatch.com/clinical-trials/listings/external-studydetails.aspx?StudyID=NCT02337530&City=Toronto&Country=Canada` [Accessed 3 Aug. 2016].

[137] Kettering Cancer Center, M. (2016). Lung Cancer: Lung Cancer Screening Decision Tool — Memorial Sloan Kettering Cancer Center. [online] Mskcc.org. Available at: `https://www.mskcc.org/cancer-care/types/lung/screening/lung-screening-decision-tool` [Accessed 5 Oct. 2016].

[138] Janssen, K., Moons, K., Kalkman, C., Grobbee, D. and Vergouwe, Y. (2008). Updating methods improved the performance of a clinical prediction model in new patients. Journal of Clinical Epidemiology, 61(1), pp.76-86.

[139] Steyerberg, E., Borsboom, G., van Houwelingen, H., Eijkemans, M. and Habbema, J. (2004). Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. Statist. Med., 23(16), pp.2567-2586.

[140] Rauh, S., Heymans, M., Mehr, D., Kruse, R., Lane, P., Kowall, N., Volicer, L. and van der Steen, J. (2016). Predicting mortality in patients treated differently: updating and external validation of a prediction model for nursing home residents with dementia and lower respiratory infections. BMJ Open, 6(8), p.e011380.

[141] Su, T., Jaki, T., Hickey, G., Buchan, I. and Sperrin, M. (2016). A review of statistical updating methods for clinical prediction models. Statistical Methods in Medical Research.

[142] Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. British Medical Journal. 2009;338.

[143] Debray, T., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. and Moons, K. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of Clinical Epidemiology, 68(3), pp.279-289.

[144] Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine. 1996;15(4):361-87.

[145] Nieboer, D., Vergouwe, Y., Ankerst, D., Roobol, M. and Steyerberg, E. (2016). Improving prediction models with new markers: a comparison of updating strategies. BMC Medical Research Methodology, 16(1).

[146] Greenland S. METHODS FOR EPIDEMIOLOGIC ANALYSES OF MULTIPLE EXPOSURES - A REVIEW AND COMPARATIVE-STUDY OF MAXIMUM-LIKELIHOOD, PRELIMINARY-TESTING, AND EMPIRICAL-BAYES REGRESSION. Statistics in Medicine. 1993;12(8):717-36.

[147] Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology. 2004;159(9):882-90.

[148] Debray, T., Koffijberg, H., Nieboer, D., Vergouwe, Y., Steyerberg, E. and Moons, K. (2014). Meta-analysis and aggregation of multiple published prediction models. Statist. Med., 33(14), pp.2341-2362.

[149] Schorning, K., Bornkamp, B., Bretz, F. and Dette, H. (2016). Model selection versus model averaging in dose finding studies. Statist. Med., 35(22), pp.4021-4040.

[150] Debray, T. (2016). Meta-analysis of clinical prediction models.

[151] A Hoeting, J., Madigan, D. and E Raftery, A. (2016). Bayesian Model Averaging. [online] Technical Report 335. Available at: `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.451.4125&rep=rep1&type=pdf` [Accessed 12 Jan. 2016].

[152] Breiman L. Stacked regressions. Machine Learning. 1996;24(1):49-64.

[153] Bello, G., Gennings, C. and Dumancas, G. (2015). Development and Validation of a Clinical Risk-Assessment Tool Predictive of All-Cause Mortality. BBI, p.1

[154] Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. Journal of Mathematical Psychology, 44(1), pp.92-107.

[155] Raftery, A., Madigan, D. and Hoeting, J. (1997). Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association, 92(437), p.179.

[156] Jackson, C., Thompson, S. and Sharples, L. (2009). Accounting for uncertainty in health economic decision models by using model averaging. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172(2), pp.383-404.

[157] Clyde, M. (2016). Model Averging. [online] Available at: `https://www2.stat.duke.edu/courses/Spring05/sta244/Handouts/press.pdf` [Accessed 4 Oct. 2016].

[158] Debray, T., Koffijberg, H., Vergouwe, Y., Moons, K. and Steyerberg, E. (2012). Aggregating published prediction models with individual participant data: a comparison of different approaches. Statist. Med., 31(23), pp.2697-2712.

[159] Riley, R., Price, M., Jackson, D., Wardle, M., Gueyffier, F., Wang, J., Staessen, J. and White, I. (2014). Multivariate meta-analysis using individual participant data. Research Synthesis Methods, 6(2), pp.157-174.

Part IV

Appendix

Appendix

12.7 Ethical Permission for Datasets

Ethical permission was applied for to the University of Sheffield in the first year of the project to work with anonymised data. This was requested in preparation for the datasets that would be collected from the International Lung Cancer Consortium. This was granted for the duration of my project.

12.8 Systematic Review Search Terms

The search criteria include a range of searches using the terms in the table.

Lung terminology	Prediction model terms	Model titles
Lc	Model	Bach
Cancer of the Lung	Risk Model	Liverpool Lung Project
Lung Cancer	Risk	LLP
Neoplasms, Lung	Cancer risk	Spitz
Neoplasms, Pulmonary	Incidence rate	Two-Stage Clonal Expansion
Pulmonary Cancer	Risk prediction	TSCE
Pulmonary Neoplasms	Risk-prediction model	
Lung tumours	Absolute risk	
Carcinoma, Non Small Cell	Risk estimate	
Non-Small Cell Lung Cancer		
Non-Small Cell Lung Carcinoma		
Nonsmall Cell Lung Cancer		
Carcinoma Oat Cell		
Oat Cell Carcinoma		
Small Cell Carcinoma		

Table 12.1: Search Terms for Systematic Review

Using the above terms the following combinations will be utilised using a whole text search in the online databases with an **OR** separating all terms under the same heading and an **AND** between different headings.

1. Lung terminology **and** Prediction model terms
2. Model titles **and** Lung terminology
3. Model titles **and** Prediction model terms
4. Lung terminology **and** Prediction model terms **and** Model titles

12.9 Detailed Dataset Preparation

An overview of the changes made to each dataset was presented in Chapter 4. A more detailed analysis for each variable is required. Modifications were required to ensure every participant had accurate and reliable information before they were used to validate the lung cancer prediction models.

The dataset preparation focused on for each variable;

1. Removing participants with missing information unless there was the option to impute this information.
2. Assessing the inputted values for erroneous information; these participants would be removed or the information changed to missing in preparation for imputation.
3. Modifying how the variable is categorised/presented so that the information is applicable for the prediction models.

The first stage is to present which variables were assessed per dataset. Not all datasets provided information for every variable. The only variables modified were variables required by a model that was applicable to the dataset. These are presented in Table 12.2 where each variable and the models that required this variable are presented per dataset;

Variable \ Dataset	ReS	UCLA	CARET	NY	Sing	NZ	CREST	Israel	ESTHER	MSH-PMH
Case/Control Histology										
Case/Control Status (All)	X	X	X	X	X	X	X	X	X	X
Eligibility (All)	X	X	X	X	X	X	X	X	X	X
Personal Factors										
Age (All)	X	X	X	X	X	X	X	X	X	X
Gender (B,LLP,S,AA)	X	X	X	X		X	X			X
Ethnicity (PLCO)	X	X	X	X						X
BMI (PLCO)	X	X	X	X						X
Education (PLCO)	X	X	X	X						X
Prior Tumour (LLP,PLCO)	X	X	X	X				X		X
Smoking History										
Basic Smoking Info (All)	X	X	X	X	X	X	X	X	X	X
CPD & Pack Years (All)	X	X	X	X	X	X	X	X	X	X
ETS (S)							X			
Family History of Cancer										
Cases of Smoking Cancer (S)							X			
Cases of Lung Cancer (LLP,PLCO,S)	X	X	X	X			X			X
Age of Lung Cancer (LLP)	X		X				X			X
Exposures and Conditions										
Asbestos Exposure (B,LLP,S)	X		X	X		X	X			X
Dust (S,AA)							X			
Hay fever (S,AA)							X			
Emphysema (S)							X			
COPD (PLCO,AA)	X	X	X	X			X			X
Pneumonia (LLP,AA)	X		X				X			X
Model Key: B = Bach; AA= African-American; S = Spitz										
X = The variable is available in the dataset and will be prepared										

Table 12.2: Table Showing Important Variables per Dataset

For each variable per dataset any modifications are reported. Discrepancies or data preparation required will be listed under each variable and subdivided under each dataset.

12.9.1 Case Control Histology

The datasets require complete information on case control status which will be used in the model validation and updating. This information is essential and any participants with missing information will be removed.

12.9.1.1 All Studies

There are no concerns; every participant is classified so no alterations were required.

12.9.2 Eligibility

Records identified if the patients were eligible when the study was collected. Ineligibility could be a product of misdiagnosis or incorrectly selecting a participant for inclusion in the study. Ineligible participant should be removed as these could negatively affect the models in the external validation.

12.9.2.1 ReSoLuCENT Study

The ReSoLuCENT study had 29 participants that were ineligible with no further explanation provided. These were removed from the study alongside one additional participant who had not been classified as either eligible or ineligible.

12.9.2.2 Remaining Studies

This information was not reported in these studies so every participant was assumed eligible and remained in the study.

12.9.3 Age

Age is an essential variable required by every model. Testing of this variable needs to ensure there are no missing entries or any unreasonable responses that may be inputted through error.

12.9.3.1 Israel, MSH-PMH, and Singapore Studies

There are 1, 6, and 7 participants with missing information in the Israel, MSH-PMH, and Singapore studies respectively. This information cannot confidently be imputed so these individuals were removed. All other participants in these studies had reasonable age entries so no further modifications were required.

12.9.3.2 Remaining Studies

All the remaining studies had complete age information and the entries were reasonable so no formatting was conducted.

12.9.4 Gender

Gender is reported in all the studies. This information would be difficult to impute if missing so participants who did not provide this information with be removed from the study.

12.9.4.1 ReSoLuCENT

In the ReSoLuCENT dataset 3 participants had missing information and were removed. No further modifications were required.

12.9.4.2 Remaining Studies

The other ILCCO datasets provided complete information for this variable so no formatting was required.

12.9.5 Ethnicity

Ethnicity is required by the PLCO models and the reported ethnicities in the datasets need to be reclassified into one of the six specified ethnicities that are required by the PLCO models;

1. White
2. Black
3. Hispanic
4. Asian
5. American India or Alaskan

6. Hawaiian or Pacific Islander

Participants will be able to be reclassified into these categories and any participants who cannot through missing information will be removed from the studies.

12.9.5.1 ReSoLuCENT Study

In the ReSoLuCENT study there were 182 patients with missing information which were removed from the study. There were an additional three participants who were classified as "Other" who were also removed. The remaining participants could be correctly classified so were eligible for inclusion.

12.9.5.2 Remaining Studies

When collecting the other ILCCO studies they were all categorised before collection. These were recategorised to conform to the PLCO categories as follows;

PLCO Model Classification	**ILCCO Classification**
White	White/Caucasian
Black	Black/African-American
Hispanic	Hispanic Black Mexican Other Latino
Asian	Asian
American India or Alaskan Native	Native American
Native Hawaiian or Pacific Islander	Hawaiian

Table 12.3: Classifying Ethnicity for the ILCCO Prepared Datasets

All participants could be successfully reclassified. However there there were 46, 2, 28, and 253 participants from the UCLA, NY Wynder, New Zealand, and MSH-PMH studies respectively that had missing information and were removed.

12.9.6 Body Mass Index

BMI is required by the PLCO Models. This may be provided in the study or can be calculated provided information for height in metres (m) and weight in kilograms (kg) can be obtained. Then BMI is calculated as follows;

$$BMI = \frac{kg}{m^2} \tag{12.1}$$

12.9.6.1 ReSoLuCENT

In the ReSoLuCENT dataset information for each participant's weight (kg) and height (m) was provided. There were 193 participants with missing information for weight and 188 participants for height. This included 181 participants with missing information for both entries.

An assessment was made to assess if other information could be used to impute this information to avoid removing the participants. However, there were no variables collected in the study, such as diet and exercise; that may have been used to impute the missing information. As a result all participants with missing information for either of the variables were removed from the study.

Additionally there was 1 unreasonable measurement for weight (867.12kg) and 4 unexpected height entries (4.32m, 0m, 0.84m, and 3.51m). The correct values could not be obtained so these participants were also removed from the study.

12.9.6.2 NY Wynder Study

BMI was reported in the NY Wynder Study but there were 534 participants with missing information who were removed. There were three unreasonable entries for BMI (92, 108, and 192) which were deemed too high and so these participants were also removed from the study.

12.9.6.3 Remaining Studies

BMI was provided as a single measure rather than information for weight and height in the remaining ILCCO studies when reported. In the UCLA, CARET, and Canadian Studies there were 4, 15, and 273 missing entries which were removed from the study. An analysis of the entries showed data ranging between 10.5–64 which are reasonable values so no further data modifications were required.

12.9.7 Education

Education is required for the PLCO models. Participants are asked their highest completed education level based on six categories;

1. Less than high-school graduate (level 1)
2. High-school graduate (level 2)
3. Some training after high school (level 3)
4. Some college (level 4)
5. College graduate (level 5)
6. Postgraduate or professional degree (level 6).

The participants in the IPDs are required to be correctly classified into these categories as the information was collected with the study objectives in mind rather than the PLCO models.

12.9.7.1 ReSoLuCENT Study

The ReSoLuCENT dataset allowed entrants to provide any information rather than selecting a predetermined category. There were 382 different entries that had to be reclassified; the exhaustive list of the reclassifying the entries is presented in Section **??** Table **??**.

There were 188 participants with missing information and a further one participant who listed "No information available". These were removed from the study as this information could not be inferred from the information for other variables.

12.9.7.2 Remaining Studies

Education was collected in the UCLA, CARET, NY Wynder, and MSH-PMH datasets. The entries for the education in these datasets had been collected using the same categories. The participants required a slight reclassification so the information was formatted into the PLCO categories as follows;

PLCO Education Category	Classified ILCCo Grade
Less than high-school graduate	No education received
High-school graduate	Basic/elementary
Some training after high school	Apprentice/vocational
Some college	Secondary schools ending by graduation/High school with degree
College graduate	Post-secondary, but not university degree
Postgraduate or professional degree	University and techn. univ. educ. by receiving title

Table 12.4: Reclassifying Education for the ILCCo Datasets

In the UCLA, CARET, NY Wynder, and MSH-PMH studies there were 1, 261, 14, and 168 missing entries respectively. These were removed as the information could not be imputed.

12.9.8 Prior Malignant Tumour

A prior malignant tumour also defined as a "personal history of cancer" and was required as a dichotomous response for the models.

12.9.8.1 ReSoLuCENT and MSH-PMH Study

Participants only provide information when a prior tumour was present in these datasets. Participants were classified as "yes" when they reported a prior tumour and "no" for all blank entries. No participants were removed.

12.9.8.2 UCLA, CARET, and NY Wynder Studies

The information was already formatted as a dichotomous response. There was 1 participant removed in the CARET study with missing information but no further modifications were required.

12.9.9 Basic Smoking Information

Some form of basic smoking information (smoking status, start age, cessation age, duration, quit duration) is required in every prediction model. It is important that the information is both complete and correct. There were some concerns with the smoking start age as there are some very young entries. It was determined any start ages below 6 were probably inaccurate but some participants may have began smoking at a very young age younger than 10. The participants will be kept in the studies but the young values will be highlighted.

12.9.9.1 ReSoLuCENT Study

There were 183 individuals who did not provide information for any of the basic smoking variables. These could be never smokers who left the smoking section blank, however it would be careless to assume this and assign all participants as never-smokers. Unfortunately the participants had to be removed from the study as correct information could not be obtained.

Additionally there were 4 former smokers who had a missing start or cessation age and 6 current smokers without a start age. This information could not be accurately determined and these participants were also removed from the study.

12.9.9.2 UCLA Study

In the UCLA study there were 23 ever-smokers with a missing start age who were removed. There were no further concerns with the start age values. There were 12 former-smokers who did not list their quit duration, although these were part of the 23 without a start age who had already been removed. The 23 participant with missing information were removed and no further modifications were required in the UCLA study.

12.9.9.3 CARET Study

The heavy smoker CARET study had complete basic smoking information and the values provided were reasonable so no preparation was required.

12.9.9.4 NY Wynder Study

Analysis in the NY Wynder study identified 1 participant who was classified as an ever smoker despite quitting 27 years ago; this participant was reclassified as a former smoker.

There were 255 ever-smokers removed as they had not provided a start age. Further start age analysis identified some entries between 8-10 years who remained in the study. There were no further concerns and no additional formatting was conducted.

12.9.9.5 Singapore Study

In the Singapore dataset there was 1 participant removed as no smoking status was provided and another participant removed who was missing information for all basic smoking variables.

There were another 6 ever-smokers removed as they did not provide information for start age or smoking duration so neither variable could be inferred from complete information for the other variable and cessation age. The remaining participants had complete information.

On assessment of the complete information in the dataset there are no errors and all the listed start ages are reasonable (6-75) as is the cessation age (20-86). There is confidence that all the smoking information is correct and no additional participants are removed from the study.

12.9.9.6 New Zealand Study

In the New Zealand dataset there were 7 ever-smokers without a start age who also did not provide their duration or cessation age and were removed. The values reported for start age and cessation age are reasonable and no cessation age is listed that is younger than their start age. No further modifications to the dataset were required besides the removal of the 7 ever-smokers without information for start age.

12.9.9.7 CREST Study

The CREST dataset quit duration was not recorded however this can be calculated using the participant's age and cessation age provided this information is correct.

There were 2 participants classified only as 'ever-smokers' rather than former- or current-smokers. Analysis of the other variables found they had missing information for all the variables so were removed from the study. There were an additional 6 smokers with a missing start age who were removed. All former smokers listed a reasonable cessation age and no start age was older than their cessation age. Analysing the complete information, there are 3 start ages listed as '-9' so these participants were removed. An additional participant had a start age of 4 which was considered too young so they were removed. The remaining start ages (6-60) and cessation ages (15-89) were all deemed reasonable.

In total 12 participants were removed from the study and the quit duration was calculated using complete information for age and cessation age.

12.9.9.8 Israel Study

In the Israel study there were 8 participants removed from the study due to no information for smoking status. There was complete information and acceptable information for start ages (6-48) and cessation ages, except for one former-smoker who listed a cessation age of "5" which was younger than their start age. Correct information cannot be obtained so this individual was removed but there were no other concerns. The remaining cessation ages (20-78) were reliable and required no further formatting.

There were a few small discrepancies in the duration recorded for ever-smokers, in this instance the difference between the start and cessation ages were taken as this information was deemed reliable. There were no concerns with the quit durations that are recorded. In total 9 participants were removed and smoking duration was recalculated using the start and cessation age.

12.9.9.9 ESTHER Study

In the ESTHER study 6 participants had not listed a smoking status and a further 7 ever-smokers did not provide a start age and were removed from the study. The complete start age information (9-40) are acceptable values. There were no missing information for cessation ages and all ages listed were greater than the start age. There were no discrepancies for the smoking duration.

In total 13 participants were removed but no additional formatting was required.

12.9.9.10 MSH-PMH Study

In the MSH-PMH Study 189 participants did not record their smoking status and were subsequently removed. Another 56 participants listed their smoking status as "ever-smokers" which needed to be re-classified as a current- or former-smoker. On analysis of the 56 participants; 54 were removed as they had no other basic smoking information. The 2 other participants were reclassified as former smokers as the remaining information indicates accordingly with a cessation age and quit duration provided. A further 130 ever-smokers in this large dataset were removed as they were also missing additional key smoking information.

An ever-smoker listed an unreasonable start age of 1 but the remaining entries are reasonable (6-60) as are the cessation ages (12-87). For the remaining participants there were no other discrepancies and modifications required.

12.9.10 Cigarettes Per Day/Pack Years

Pack years and CPD were also required by the prediction. This information was reported in each study but there were differences in reporting between studies. If information was only provided for one of these variables then the other variable can be calculated as follows;

$$PackYears = \frac{CPD}{20} \times SmokeDuration \tag{12.2}$$

There can be a range of values submitted for CPD but any value below 100 would be considered reasonable.

12.9.10.1 ReSoLuCENT Study

The ReSoLuCENT data does not explicitly record the pack years or CPD quantities. Instead the study reported the average CPD smoked at 20, 30, 40, 50, and last year; these were reported when applicable to the participant. To calculate the pack year and CPD information the following assumptions are made;

1. The participants smoke at a constant rate from their start age until their first relevant entry.

(a) **Example:** A participant who started smoking at 22 will only enter their first CPD at 30. From 22-30 they are considered to have smoked at the same consistency as at 30.

2. Between two consecutive gaps (20-30, 40-50, 30 - Last Year for a 38 year old) there is a yearly linear increase or decrease between these values.

   (a) **Example:** A participant who smoked 40 CPD at 30 and 20 CPD at 40: gives the following;

Age	30	31	32	33	34	35	36	37	38	39	40
CPD	40	38	36	34	32	30	28	26	24	22	20

Table 12.5: Example CPD based on information provided by the ReSoLuCENT data

For all never-smokers a true zero value could be entered. Across the ever-smokers there were 20 participants who did not provide any smoking quantities and were removed. Next we evaluated the CPD values, here there were 5 unreasonable entries over 100 CPD (1015, 140, 140, 9999, 810). These entries were reclassified as omitted data at this stage.

The remaining ever-smokers either have complete information or semi-complete information across the variables. This data may be successfully imputed as all participants now have some level of information for these variables which may be used to infer the missing values rather than removing the participants. Attempts were made to impute this missing information. If the values can be successfully imputed then no further modifications are required in the dataset; otherwise these participants with partial missing information will be removed.

12.9.10.2 UCLA Study

The UCLA dataset provided the pack years and CPD information; in all instances these two values were consistent. 23 ever-smokers had missing information for both variables and were removed.

The complete CPD information was reasonable (0 - 80) which falls below the agreed 100 CPD ceiling for a chain smoker.

12.9.10.3 CARET Study

The CARET provided pack year and CPD information which for every participant matched using Equation 12.9.10. The CPD ranged was from 1 to 80 and with no missing information no formatting was required for the CARET dataset.

12.9.10.4 NY Wynder Study

Information for both CPD and pack years were provided in the NY Wynder study with no discrepancies between these values. In 6 instances, pack years were not provided but can be calculated using the CPD values and Equation 12.9.10. Analysing the CPD values all were below 100 CPD and with no missing information there was no further formatting required.

12.9.10.5 Singapore Study

The CPD and pack years were provided in the Singapore study and 3 ever-smokers were removed due to missing information for both variables. The remaining participants have complete information for pack years and CPD with no discrepancies between the values. The completed CPD values ranged from 1 to 80 which were reasonable responses so no further removals were required.

12.9.10.6 New Zealand Study

The New Zealand dataset records pack years and CPD although CPD was missing for several participants. Despite this only 2 ever-smokers had to be removed missing information for both variables. An additional 19 ever-smokers report pack years as 0 therefore they cannot be classified as ever-smokers despite complete information for start age and cessation age. As a result these were removed from the study. CPD was calculated when missing using Equation 1.3.10 and an analysis of the complete CPD information was conducted; the results were reasonable (1-75) so no participants were excluded. In total 21 ever-smokers were removed.

12.9.10.7 CREST Study

The CREST dataset had complete CPD and pack years information. For 1 participants the CPD and pack year results were '-9' so they were removed from the study. A further 3 reported '-9' for pack years but these participants have correct CPD values reported so their pack year information can be correctly calculated.

The CPD values ranged from 1-110 for ever-smokers; this included 1 participant who reported a CPD exceeding 100 CPD. This participant was removed from the study as the value was deemed unrealistic. The remaining results had no discrepancies between the pack years and the CPD values. In total 2 participants were removed and 3 pack year results were recalculated but no further formatting was required.

12.9.10.8 Israel Study

In the Israel study there were some minor discrepancies between the CPD and pack years. In these cases the pack years will be recalculated using the CPD values. 5 ever-smokers were removed because of missing information for both variables but no additional modifications were required with the remaining CPD values between $1 - 100$.

12.9.10.9 ESTHER Study

In the ESTHER study both CPD and pack years are reported with no discrepancies between the values. There were 15 ever-smokers with missing CPD and pack years values or with reported values of 0 so were excluded. The remaining CPD values (1-57) were reasonable.

12.9.10.10 MSH-PMH Trial

In the MSH-PMH study there are no discrepancies between the CPD and pack year values. 10 ever-smokers were removed as they stated they smoked 0 CPD but the remaining entries (1 - 75) are reasonable so no additional formatting is required.

12.9.11 Environmental Tobacco Smoke

Environmental tobacco smoke (ETS) is required as a dichotomous variable. The Spitz model article states ETS as having been exposed to someone else's cigarette smoke as home or at work on a regular basis i.e. daily or weekly, as well as on years of exposure to ETS [75]. In the datasets, exposure durations and intensity rates are not expected to be provided so the original values will be accepted.

12.9.11.1 CREST Study

ETS is reported in the CREST dataset as exposure at home or work. If a positive exposure is reported for either location then this will be re-stated as a positive ETS exposure. There were 16 participants with missing information for both variables that were removed from the study but the remaining participants were successfully reclassified as having a positive or negative exposure.

12.9.12 Family History of Cancer

When considering a family history of cancer in the LLP, Spitz, and PLCO models require a count of the volume of cancers occurring in first degree relatives in specific subgroups of cancers. Information needs to be reclassified into the following groups;

1. Any cancer
2. Any cancer except melanoma
3. Smoking related cancers
4. Lung cancer

All the groups are a nested subgroup of the groups above and the information will be obtained using the Cancer Research UK website and their causes of each type of cancer [**?**]. Cancers were only defined as melanomas if it is explicitly reported so skin cancer was not defined as a melanoma.

It is important the information is only reported for first degree relatives which are parents, siblings, and children. Additionally, the age of onset is required; this is reported for each subgroup and for participants with multiple family members with cancer in the same subgroup the youngest age of onset is reported.

12.9.12.1 ReSoLuCENT

In the ReSoLuCENT dataset all entries were first degree relatives to the participant. Cancers were reported in the study as an open entry which required reformatting into the four categories. The comprehensive list of results are listed in Section 12.11 in Table 12.7.

Analysing the age of onset 3 participants listed "999" who will be excluded. An additional 166 patients listed "0" or left the entry blank which could not be resolved and they were also removed.

12.9.12.2 Remaining Studies

The UCLA, CARET, NY Wynder, New Zealand, CREST, and MSH-PMH Study all include a family history of cancer. In all instances the cancer is only provided for first degree relatives. The type of cancer is reported differently between studies from blank entries for the participant to report any value to a series of ICD9 codes. The method of reclassifying the cancers is presented separately in the Appendix (Section 12.11) and the ICD9 codes were determined using an on-line database [122]. There were no concerns with the age of onset of the cancer in any instance so the only formatting required was to reclassify the cancers.

12.9.13 Asbestos Exposure

Asbestos exposure is defined as a prolonged period of time working in a profession that has a high risk of asbestos contact. In the studies the only information provided was a dichotomous response so these will be accepted as the exact exposure duration or intensity cannot be determined.

12.9.13.1 ReSoLuCENT Study

There were 46 counts of missing information in the ReSoLuCENT data. These participants will not be removed at this stage, as testing will be conducted to see if the missing information can be imputed. An assessment will be made whether to impute this information in Chapter 4 Section 4.10. If the information cannot be imputed then these participants with missing information will be removed.

12.9.13.2 CARET Study

Each participant of the CARET cohort is defined as either "Heavy smoker" or "Asbestos-exposed worker". There were no missing values in this information and a positive asbestos exposure was assigned to all participants classified as "Asbestos-exposed worker".

12.9.13.3 NY Wynder Study

Asbestosis was only reported in the NY Wynder study when it is present in a participant. When this was left blank they were confidently reclassified as not exposed. No participants were removed from the study.

12.9.13.4 New Zealand Study

In the New Zealand study asbestos exposure has complete information except for 2 participants who were removed from the study.

12.9.13.5 CREST Study

In the CREST study asbestos exposure was reported for their present job and two most recent previous jobs. If they were exposed in at least one of these job environments they were assigned a positive asbestos exposure. No participants were removed from the study.

12.9.13.6 MSH-PMH Study

There were 380 participants with missing information for asbestos exposure. Before removing the participants with missing values, assessments were made to impute the values. This is reported in Chapter 4 and if the missing values cannot be successfully imputed then the 380 participants will be removed from the study.

12.9.14 Dust Exposure

Dust exposure is a dichotomous response for work exposures to dust.

12.9.14.1 ReSoLuCENT Study

Dust exposure was reported in the ReSoLuCENT. There were 240 participants with missing information or stated as unknown in the dataset. This information will be evaluated to assess if it can be imputed. This was presented in Section 4. If the information was not successfully imputed then these participants were removed from the study

12.9.14.2 CREST Study

Dust exposure in the CREST dataset already has complete information so no formatting was required.

12.9.15 Hay Fever

Hay fever is required as a dichotomous response.

12.9.15.1 CREST Study

In the CREST dataset all previous lung diseases are reported in one entry using an ICD9 code. The ICD code's "477.0 - Allergic Rhinitis Due To Pollen" and "477.9 - Allergic Rhinitis Cause Unspecified" [122] will be classified as hay fever. If this is not reported then it is assumed they do not have hay fever so no participants were removed from the study.

12.9.16 Emphysema

Emphysema is required a dichotomous response.

12.9.16.1 CREST Study

In the CREST dataset this is reported under previous lung conditions using ICD9 codes. The following ICD9 codes' "492.0 - Emphysematous Bleb" and "492.8 - Other Emphysema" will be classified as a positive emphysema diagnosis. The remaining participants will be classified as a negative exposure.

12.9.16.2 MSH-PMH

There were 240 participants that had missing information for emphysema. Attempts will be made to impute this information rather than remove participants at this stage. This is explored in Chapter 4.

12.9.17 Chronic Obstructive Pulmonary Disease

COPD is required a dichotomous response by the models.

12.9.17.1 ReSoLuCENT Study

In the ReSoLuCENT data COPD has 1 participant with missing information. Attempts are made to impute the missing information rather than removing this participant which is reported in Chapter 4. If this information cannot be successfully imputed then they will be removed from the study.

12.9.17.2 UCLA Study

COPD is fully reported in the UCLA dataset with no missing values so no alterations were made.

12.9.17.3 CARET

COPD is fully reported in the CARET dataset so no formatting was required.

12.9.17.4 NY Wynder Study

In the NY Wynder study COPD is reported when the condition was present. When this was not reported the data was altered with confidence to not present.

12.9.17.5 CREST Study

COPD is reported in the CREST study as lung conditions using ICD9 codes. The following ICD9 codes "491.20 - Obstructive Chronic Bronchitis Without Exacerbation", "491.21 - Obstructive Chronic Bronchitis With (Acute) Exacerbation" and "491.22 - Obstructive Chronic Bronchitis With Acute Bronchitis" are de ned as a positive COPD diagnosis. If these ICD9 Codes were not reported then participants were defined as no COPD.

12.9.17.6 MSH-PMH Study

There were 249 participants with missing information in the MSH-PMH study. Before removing the participants with missing information, attempts will be made to impute this information. This will be reported in Chapter 4 and if the information can be imputed no participants would be removed from the study.

12.9.18 Pneumonia

Pneumonia is required as a dichotomous response and where possible missing information will be imputed.

12.9.18.1 ReSoLuCENT Study

There were no concerns with Pneumonia in the ReSoLuCENT dataset.

12.9.18.2 CARET Study

Pneumonia is reported in the CARET study but 6 participants with missing information were removed.

12.9.18.3 CREST Study

Pneumonia is reported in the CREST study using ICD9 codes. Any ICD9 code between 480 - 486.9 [122] will be reclassified as a positive pneumonia diagnosis.

12.9.18.4 MSH-PMH Study

In the MSH-PMH dataset there were 215 participants with missing information for this variable. This information may be imputed using additional information. At this stage no participants were removed from the study.

12.10 Categorising Education Levels and Types of Cancers

The datasets have now been prepared and the missing information will be imputed or removed. The table for classifying the cancers are presented;

12.10.1 ReSoLuCENT Education Levels

PLCO Education Type	Classified Entries
Less than high-school graduate	"CSE RSA 1 & 2 Word processing", "CSE State registered nurse", CSS, None, "None School leaving certificate", Health Care Certificate, H-G-V Licence, P.S.V Licence., "Level 1 Numeracy ITQ Level 2", NDRM RHS × 4 Certificate, NNEB class 1, Qualified HGV driver, Typing Certificate
High-school graduate	CSE, "CSE GCSE", "CSE GCSE G.C.E 'O' level", "CSE GCSE NVQ 2", "CSE GCSE NVQ 2 NVQ 3", "CSE GCSE NVQ 3", "CSE NVQ 1", "CSE NVQ 1 NVQ 2", "CSE NVQ 2", "CSE O level", "CSE O level BTEC Sports Management", "CSE O level GCSE", "CSE O level GCSE HND", "CSE O level GCSE NVQ 2 BTEC National Diploma", "CSE O level HND", "CSE O level NNEB", "CSE O level NVQ 1 NVQ 2", "CSE O level NVQ 2", "CSE O level Registered Nurse", "CSE O level State Registered Nurse", GCSE, "GCSE Army Certificate of Education", "GCSE City & Guilds NVQ 2", "GCSE HNC", "GCSE HND", "GCSE NVQ 1", "GCSE NVQ 1 NVQ 2", "GCSE NVQ 2", "GCSE Qualified Nail Technician", "GCSE RGN", "GCSE SEN RGN", NVQ 1, "NVQ 1 Northern counties english language/ History", "NVQ 1 NVQ 2", "NVQ 1 NVQ 2 School leaving certificate CSE O level", NVQ 2, "NVQ 2 NVQ 3 GCSE", O level, "O level BHSAI I", "O level City & Guilds", "O level City & Guilds 'Around drug abuse'", "O level City & Guilds HND", "O level CSE", "O level CSE City & Guilds NVQ 2", "O level GCSE", "O level HND", "O level MITD", "O level NVQ 2", "O level RGN", "O level RSA I,II and III typing RSA typing teacher", "O level RSA School Cert Pitman shorthand RSA typing Book keeping", "O level SRN", "O level Staff registered nurse", School leaving certificate, "School leaving certificate City & Guilds", RSA, "School leaving certificate CSE", "School leaving certificate CSE GCSE", "School leaving certificate CSE NVQ 3", "School leaving certificate CSE O level NVQ 3", "School leaving certificate GCSE", "School leaving certificate GCSE HND", "School leaving certificate GCSE NVQ 1 NVQ 2 NVQ 3", "School leaving certificate NVQ 1 NVQ 2", "School leaving certificate O level", "Technical college exams CSE", "Technical college exams GCSE", "Technical college exams NVQ 1 NVQ 2", South Africa std 8
Some training after high school	A level or Highers "A level or Highers GCSE" British Coal Deputy Course Completed apprenticeship "Completed apprenticeship Trade certificate" "CSE A level or Highers" "CSE Completed apprenticeship"

Some training after high school	"CSE Completed apprenticeship Trade certificate" "CSE CPC Transport management ADR hazardous goods handling" "CSE NVQ 1 NVQ 2 NVQ 3" "CSE NVQ 2 NVQ 3" "CSE NVQ 2 NVQ 3 Trade certificate" "CSE NVQ 2 NVQ 3 Trade certificate BTEC" "CSE O level A level or Highers" "CSE O level A level or Highers HND" "CSE O level A level or Highers RGN" "CSE O level A level or Highers Vocational Diploma" "CSE O level Completed apprenticeship" "CSE O level GCSE A level or Highers" "CSE O level GCSE Technical college exams NVQ 2 NVQ 3" "CSE O level NVQ 3" "CSE O level NVQ 3 BTEC Level 2 Door Supervision" "CSE O level NVQ 3 ILM Diploma in Massage" "CSE O level Technical college exams" "CSE O level Technical college exams Completed apprenticeship" "CSE O level Technical college exams Pre-nursing Nursing" "CSE Technical college exams" "CSE Technical college exams City & Guilds NVQ 3 Completed apprenticeship Trade certificate" "CSE Technical college exams Completed apprenticeship Trade certificate" "CSE Trade certificate" "CSE Trade certificate IOSH, RASWA Supervisor" "GCSE A level or Highers" "GCSE A level or Highers HND" "GCSE A level or Highers NVQ 1 NVQ 2" "GCSE A level or Highers NVQ 2" "GCSE A level or Highers NVQ 3" "GCSE A level or Highers Trade certificate" "GCSE City & Guilds NVQ 2 NVQ 3 Level 2 infection control Level 3 palliative care Level 2 health & safety" "GCSE City & Guilds NVQ 3 Completed apprenticeship Trade certificate" "GCSE NVQ 1 NVQ 2 NVQ 3" "GCSE NVQ 1 NVQ 2 NVQ 3 NVQ 4" "GCSE NVQ 3" "GCSE NVQ 3 Completed apprenticeship" "GCSE Completed apprenticeship" "GCSE Technical college exams" "GCSE Technical college exams City & Guilds Completed apprenticeship" "GCSE Technical college exams City & Guilds Completed apprenticeship Trade certificate" "GCSE Technical college exams Completed apprenticeship" "GCSE Technical college exams Completed apprenticeship Trade certificate" HND "HND Completed apprenticeship" "GCSE University degree BTEC and GCE (A level equivalent)" "GCSE City & Guilds NVQ 1 NVQ 2 NVQ 3" Memeber of the Institute of Chirpodists and Podiatrists "NVQ 1 NVQ 2 Completed apprenticeship" "NVQ 1 NVQ 2 NVQ 3" "NVQ 1 NVQ 2 NVQ 3 Trade certificate Football coach First aid, Health & Safety course Fire & Bomb awareness course" "NVQ 2 Completed apprenticeship" "NVQ 2 NVQ 3" NVQ 3 "NVQ 3 City & Guilds" NVQ 4 "O level A level or Highers" "O level A level or Highers City & Guilds" "O level A level or Highers City & Guilds HND Completed apprenticeship" "O level A level or Highers City & Guilds Trade certificate" "O level A level or Highers Civil Service Internal Qualifications" "O level A level or Highers Fellow of Chartered Insurance Institute" "O level A level or Highers HND" "O level A level or Highers HND HNC" "O level A level or Highers HND Qualified as a Chartered Accountant" "O level A level or Highers NIVEB (nursery nurse) RGN (nurse)"

Some training after high school	"O level A level or Highers NNEB Nursery Nurse RGN Nurse" "O level A level or Highers NVQ 4" "O level A level or Highers RGN/RSCN" "O level City & Guilds Completed apprenticeship" "O level City & Guilds Completed apprenticeship Diploma in Financial Services" "O level City & Guilds NVQ 4" "O level GCSE A level or Highers" "O level GCSE A level or Highers HND" "O level GCSE A level or Highers NVQ 2 HNC" "O level GCSE Nurse Training" "O level GCSE NVQ 2 NVQ 3 NVQ 4" "O level GCSE Technical college exams" "O level NVQ 1 NVQ 2 NVQ 3" "O level NVQ 3" "O level NVQ 4" "O level Technical college exams City & Guilds Completed apprenticeship Trade certificate" "O level Technical college exams City & Guilds NVQ 2 Managers Passport Managers Health and Safety" "O level Technical college exams City & Guilds NVQ 2 Trade certificate" "O level Technical college exams NVQ 1" OCN Horticultural Course "O level Technical college exams NVQ 3 Trade certificate NEBBS Cert Various H&S Qualifications" "O level Technical college exams Trade certificate" "School leaving certificate City & Guilds Completed apprenticeship" "School leaving certificate City & Guilds Completed apprenticeship Trade certificate" "School leaving certificate City & Guilds NVQ 3" "School leaving certificate City & Guilds Technical college exams" "School leaving certificate Completed apprenticeship" "School leaving certificate Completed apprenticeship Trade certificate" "School leaving certificate Completed apprenticeship Trade certificate Painter and decorating" "School leaving certificate CSE A level or Highers City & Guilds NVQ 1 Completed apprenticeship Trade certificate" "School leaving certificate CSE GCSE Technical college exams City & Guilds" "School leaving certificate CSE O level Technical college exams City & Guilds Completed apprenticeship" "School leaving certificate CSE Technical college exams City & Guilds NVQ 1" "School leaving certificate O level A level or Highers" "School leaving certificate O level Technical college exams RSA Typing/Shorthand Book keeping" "School leaving certificate Technical college exams City & Guilds NVQ 1 Completed apprenticeship Trade certificate Welder & Engineer" Technical college exams "Technical college exams City & Guilds" "Technical college exams City & Guilds Completed apprenticeship" "Technical college exams City & Guilds Completed apprenticeship Trade certificate" "Technical college exams City & Guilds HND" "Technical college exams City & Guilds HND Mechanical Engineering" "Technical college exams City & Guilds NVQ 1 NVQ 2 Computer studies" "Technical college exams City & Guilds ONC" "Technical college exams City & Guilds Trade certificate Paramedic (SYMAS)" "Technical college exams Completed apprenticeship Trade certificate" "Technical college exams Completed apprenticeship Trade certificate City & Guilds NVQ 2 NVQ 4", Work Training / Computers / SOS / Manager, "Technical college exams RSA - WP Typing, Audio – advanced Medical WP Doc Pres Mailmerge", "Technical college exams Trade certificate", TESOL (Teaching English), Trade certificate, "Trade certificate City & Guilds", "Trade certificate CSE O level"

Some college	3 Standard Grades, 1 HNC and various NCs "CSE A level or Highers Technical college exams" "CSE City & Guilds Completed apprenticeship Technical college exams" "City & Guilds Technical college exams Completed apprenticeship" "CSE GCSE Secretarial college" "CSE GCSE Technical college exams" "CSE GCSE Technical college exams NVQ 2" "CSE NVQ 2 NVQ 3 NVQ 4 HND" "CSE NVQ 3 NVQ 4" "CSE O level A level or Highers Technical college exams HND" "CSE O level City & Guilds NVQ 2 NVQ 4 Management Qualifications" "CSE O level NVQ 1 NVQ 2 NVQ 3 NVQ 4 Trade certificate" "CSE O level NVQ 3 Secretarial college" "CSE O level Secretarial college" "CSE O level Secretarial college Royal soc of arts" "GCSE A level or Highers Diploma in Nursing" "GCSE A level or Highers NVQ 3 NVQ 4 Certificate in Social Science" "GCSE A level or Highers Teaching diploma Diploma in Special Education" "GCSE A level or Highers Technical college exams Teaching diploma" "GCSE City & Guilds NVQ 3 Access to Health & Nursing Studies" "GCSE City & Guilds NVQ 3 HND Secretarial college" "GCSE City & Guilds NVQ 3 NVQ 4 Access into nursing" "GCSE Secretarial college" "GCSE NVQ 2 NVQ 3 Advanced Diploma in Nursing Studies" "GCSE NVQ 3Advanced Diploma in Nursing Studies" "GCSE NVQ 4 Association of Accounting Technicians" "NVQ 3 Secretarial college" "NVQ 4 Secretarial college Diploma" "O level A level or Highers HND Teaching diploma" "O level A level or Highers Secretarial college Teaching diploma" "O level A level or Highers Secretarial college University degree" "O level A level or Highers Teaching diploma" "O level A level or Highers Teaching diploma University degree" "O level A level or Highers Technical college exams City & Guilds GCSE" "O level A level or Highers Technical college exams City & Guilds HND" "O level A level or Highers Technical college exams HND" "O level Secretarial college" "O level Technical college exams" "O level Technical college exams City & Guilds" "O level Technical college exams City & Guilds Trade certificate Teaching diploma" "O level Technical college exams Completed apprenticeship Secretarial college RSA" "O level Technical college exams NVQ 3 Completed apprenticeship Secretarial college" Secretarial college "Secretarial college- RSA – Shorthand Typing – English Secretarial college" "Secretarial college O level" "Secretarial college School leaving certificate" "Technical college exams City & Guilds Secretarial college" "Technical college exams Secretarial college" Advanced Diploma in Nursing (Mental Health) City & Guilds "City & Guilds BTEC's" "City & Guilds Completed apprenticeship" "City & Guilds Completed apprenticeship Teaching diploma" "City & Guilds Completed apprenticeship Trade certificate" "City & Guilds CSE" "City & Guilds HND" "City & Guilds NVQ 1 NVQ 2" "City & Guilds NVQ 1 NVQ 2 NVQ 3" "City & Guilds NVQ 2" "City & Guilds NVQ 2 Trade certificate" "City & Guilds NVQ 3" "City & Guilds O level CSE" "City & Guilds Secretarial college" "City & Guilds Trade certificate" "City & Guilds Various Plant machinery / demolition" "City & Guilds Youth Leaders Certificate" "Completed apprenticeship Trade certificate City & Guilds" "CSE A level or Highers Nursing Diploma" "CSE City & Guilds" "CSE City & Guilds Completed apprenticeship" "CSE City & Guilds NVQ 1 NVQ 2" "CSE City & Guilds NVQ 2 Completed apprenticeship Trade certificate" "CSE City & Guilds NVQ 3 Teaching diploma Nursing Othe Diplomas"

Some College	"CSE City & Guilds Trade certificate" "CSE GCSE NVQ 4 University Certificate" "CSE GCSE O level Technical college exams City & Guilds Completed apprenticeship Trade certificate" "CSE O level A level or Highers Technical college exams City & Guilds Completed apprenticeship Trade certificate" "CSE O level City & Guilds" "CSE O level City & Guilds Completed apprenticeship" "CSE O level City & Guilds NVQ 1 NVQ 2 NVQ 3 NVQ 4 Completed apprenticeship Trade certificate" "CSE O level GCSE A level or Highers Teaching diploma" "CSE O level GCSE Matriculation Technical college exams City & Guilds HND Completed apprenticeship Trade certificate" "CSE O level NVQ 4 Registered mental nurse" "CSE O level Technical college exams City & Guilds" "CSE O level Technical college exams City & Guilds NVQ 3 NVQ 4 Completed apprenticeship" "CSE O level Technical college exams NVQ 1 NVQ 2 NVQ 3 NVQ 4" "CSE School leaving certificate City & Guilds NVQ 2 NVQ 3 NVQ 4 RMA IOSHH" "CSE Teaching diploma" "CSE Technical college exams City & Guilds Completed apprenticeship" "CSE Technical college exams City & Guilds Completed apprenticeship Trade certificate" "GCSE City & Guilds" "GCSE City & Guilds Completed apprenticeship Trade certificate" "GCSE City & Guilds National Certificate in Dental Nursing" "GCSE City & Guilds NVQ 1" "GCSE City & Guilds NVQ 1 NVQ 2" "GCSE City & Guilds NVQ 1 NVQ 2 Completed apprenticeship" "GCSE City & Guilds NVQ 1 NVQ 2 NVQ 3 NVQ 4" "GCSE City & Guilds NVQ 1 NVQ 2 NVQ 3 NVQ 4 Completed apprenticeship Trade certificate" "GCSE City & Guilds NVQ 2 NVQ 3 NVQ 4" "O level A level or Highers Diploma Interior Design Post grad diploma art therapy" "O level A level or Highers Teaching diploma Diploma in Educatiion" "O level A level or Highers Teaching diploma Post-graduate degree" "O level GCSE Art College Diploma" "O level Teaching diploma" "O level Teaching diploma SRN" RGN "School leaving certificate GCSE NVQ 2 Cache Diploma Child Care & Education" Teaching diploma "Teaching diploma Post-graduate degree A level or Highers CSE O level"
Postgraduate or professional degree	"A level or Highers O level University degree Post-graduate degree" "A level or Highers GCSE Teaching diploma University degree Diploma in French, France" "A level or Highers GCSE University degree" "CSE O level A level or Highers City & Guilds University degree Post-graduate degree" "CSE O level A level or Highers Nursing Diploma University degree" "CSE O level GCSE A level or Highers University degree Post-graduate degree" "CSE O level GCSE Registered Nurse RN Post Grad Diploma HE" "CSE O level Teaching diploma University degree Post-graduate degree" "CSE O level Technical college exams City & Guilds HND Completed apprenticeship Trade certificate Teaching diploma University degree Post-graduate degree" "CSE O level Technical college exams City & Guilds Trade certificate University degree" "CSE O level Technical college exams Completed apprenticeship University degree HNC Mechanical Engineering" "CSE O level Technical college exams HND University degree Post-graduate degree" "CSE O level University degree" "GCSE A level or Highers City & Guilds NVQ 1 NVQ 2 HND Teaching diploma University degree" "GCSE A level or Highers HND University degree"

Postgraduate or professional degree	"GCSE A level or Highers University degree" "GCSE A level or Highers University degree HNC" "GCSE A level or Highers University degree Post-graduate degree" "GCSE A level or Highers University degree Post-graduate degree Chartered Public Finance Accountant (CPFA)" "GCSE NVQ 3 Teaching diploma University degree Nursing Diploma in Critical Care" "GCSE Technical college exams City & Guilds NVQ 1 NVQ 2 NVQ 3 University degree" "GCSE University Diploma" "GCSE Technical college exams HND University degree" "O level A level or Highers City & Guilds HND University degree" "O level A level or Highers City & Guilds NVQ 3 University degree PGCE TDLB D32,33,34 Assessor Awards" "O level A level or Highers City & Guilds Teaching diploma University degree" "O level A level or Highers City & Guilds University degree" "O level A level or Highers NVQ 1 NVQ 2 NVQ 3 NVQ 4 HND University degree" "O level A level or Highers NVQ 4 University degree Post-graduate degree" "O level A level or Highers Post-graduate degree Fellowship of Royal College" "O level A level or Highers Post-graduate degree University degree" "O level A level or Highers Technical college exams HND Teaching diploma University degree Post-graduate degree" "O level A level or Highers Technical college exams NVQ 4 University degree ECDL" "O level A level or Highers University degree" "O level A level or Highers University degree Certificate in Psychiatric Social Work" "O level A level or Highers University degree Diplomas (2) Radiography Registered General Nurse Registered Health Visitor" "O level A level or Highers University degree Law Society, Solicitors Qualification" "O level A level or Highers University degree Post grad Diploma" "O level A level or Highers University degree Post-Grad diploma in town planning." "O level A level or Highers University degree Post-graduate degree" "O level A level or Highers University degree Post-graduate degree PHD" "O level A level or Highers University degree Professional" "O level GCSE A level or Highers University degree" "O level GCSE University degree" "O level Post-graduate degree" "O level University degree" Post-graduate degree "Post-graduate degree University degree A level or Highers O level" "School leaving certificate O level A level or Highers City & Guilds Completed apprenticeship University degree BSc" University degree "University degree Post-graduate degree"

Table 12.6: Reclassifying the ReSoLuCENT Education Levels

12.11 Classifying Cancers

12.11.1 ReSoLuCENT Data Cancer Types

Required Cancer Type	**Classified Entries**
Melanoma	Melanoma, ear (melanoma)
Any Type of Cancer	Beast, Blood, Bone, Bone Marrow, Bones (possibly secondary), Bowel, Bowel Cancer, Brain, Brain Tumor, Brain Tumor, Brain tumour, CLL, Caecum, Central Nervous System, Colon, Duodenum, Eye, Intestines, Leg, Lymph, Lymphatic, Lymphatic System, Lymphoma, Multiple, Multiple tumors., NHL, NK, Primary unknown, diagnosed with multiple, Prostate, Prostate Cancer, Prostate cancer, Prostrate, Right eye, Secondary bone cancer, Skin, Small Bowel, Spinal cord, Testicle, Testicular, Tumours Left, UK, Uk, Unknown, Unknown primary, Uterus, Womb, ampullary, blood, bone, bone marrow, bones, bowel, brain, colon, ear, ear (NHL), leg, lymph, lymphoma, n/k, nk, not known, not stated, prostate, prostate, bone, rectal, skin, skin (back), skin (face), skin on face, testicle, testicular, thyroid, uk, unknown, uterus, vagina, womb, whole body!
Smoking Related Cancer	?gynae, Bladder, Bladder & Bowel, Bowel and Liver, Breast, Breast, Breast cancer, Breast, Spine, Cancer of the breast, Cervical, Cervical Cancer, Cervix, Cheek bone, Chin - BCC, Eyebrow and lips, Face, Gullet, Gum Head Groin (Hodgkins), Hodgkins, Kidney, Kidneys, Larynx, Leukaemia, Liver, Liver ? secondary to breast, Liver and Spleen, Liver/pancreas, Mouth, Naso/pharyngeal, Nose, Nose/skin, Oesohagus, Oesophagus, Ovarian, Ovaries, Ovary, Ovary/Cervix, Pancreas, Pancreatic, Renal, Stoma - Intestinech, Stomach, Stomach and bowel cancer, Stomach, lymph glands, Throat, Throat and Bowel, Windpipe, bladder, breast, breat, cervical, cervix, chin, face, forehead, head, kidney, kidney or liver, kidney, liver & bowel, leukaemia, liver, liver & pancreas, mouth, mouth & throat, neck, nose, oesophagus, oral, ovarian, ovaries, ovary, ovrian, pancreas, spleen, stomach, throat
Lung Cancer	Lung, ?Lung, Breast and Lung, Breast, Lung, Lung, Lung, Lung - Panoast, Lung Cancer, Lung Cancer, Lung Caner, Lung and (Brain mets), Lung and Stomach, Lung cancer, Lung, Brain, Lung, Liver, Lung, Spine, Lung, lymph, Lung/Brain, Lung?, Lungs, Lungs, Lungs - Mesolithioma, Lungs and Brain, Lungs and Brian, Testicle and Lung, Throat/Lung, brain, lung & adrenal, lung, lung, lung (mets bones & liver), lungs, stomach/lung

Table 12.7: ReSoLuCENT: Classifying the Cancers

12.11.2 UCLA Data Cancer Types

Required Cancer Type	**Classified Entries**
Melanoma	Basal skin melanoma, melanoma, melanoma/skin, multiple melanoma-skin, skin - melanoma, skin : melanoma, skin melanoma, skin melanoma (nose), skin, melanoma, skin-basal squamous and melanoma, skin-melanoma, skin–melanoma
Any Type of Cancer	adreno carcenoma, Axillary Lymph Node, back, basal cell, basal cell carcinoma, basal cell skin, basal skin cell, Blood, Blood (not leukemia), bone, bone cancer, bowel/colon, brain, brain stem, brain tumor, brain/face, carcinoma, colon, colorectal, Endometrial, Fibroid, fibromyalgia, intestinal, intestinal (upper), lymph, lymph node, Lymph node, skin, lymphoma, lymphosarcoma, lympnoma on sacrum, milofibrosis, multiple myeloma, multiple myloma, myloma, myloma (bone), prostate, rectal, Recurring tumor on arm, renal. rhabdomyosarcoma, skin, skin - basal cell, skin - benign, skin - not melanoma, skin (basal cell), skin carcinoma, skin, basal, skin, carcinoma, skin, doesn't know if melanoma, skin, non melanoma, skin, non-melanoma, skin, not melanoma, skin, squamous cell, skin, various, skin-basal, skin-carcinoma, skin–non-melanoma carcinoma, skin-nose (minor), skin-squamous, skin–unknown if melanoma, spinal, spine, squamos cell carcinoma, squamous cell, squamous cell carcinoma, testicular, thyroid, tumor head/back region, unknown, uterus, uterus (doesn't know type), vaginal
Smoking Related Cancer	abdominal tumor, Acute leukemia, appendix and bowel, bilateral breast, bladder, breast, breast, brain, cervical, esophagus, Facial Tumor, gynecological, Hodgkins lymphoma, Hodgkins Nerve Disease, hysterectomy/cervix, jaw (oral), kidney, larynx, leukemia, lip, lip/throat, liver, liver then to lung, mandible jaw, Metastatic breast, neck, neck lymph nodes, nose, oral, ovarian, ovarian/ cervical, pancreas, pancreas/liver, pharynx, salivary gland, some type of stomach or abdominal, spleen, stomach, stomach OR colon, stomach–liver, throat, throat (Laryngectomy), tongue, vocal cords
Lung Cancer	breast, lungs, adrenal glands, kidney; Colorectal, lung; jaw and lung, lung, lung (small cell), lung adenocarcinoma, lung cavity, lung or esophagus, lung, adenocarcinoma, lung, small cell, lung/small cell, lymphoma/lung, misothelioma, oat-cell carcinoma (lung), oral and lung, Sarcoidosis,

Table 12.8: UCLA: Classifying the Cancers

12.11.3 CARET Data Cancer Types

Required Cancer Type	**Classified Entries**
Melanoma	18 = Skin-Melanoma
Any Type of Cancer	2 = Bone, 3 = Brain, 6 = Colon, 12 = Lymphoma, 16 = Prostate, 17 = Rectum, 19 = Skin-Not melanoma, 20 = Skin-not specified, 22 = Thyroid, 23 = Uterus, 24 = Other or unknown,
Smoking Related Cancer	1 = Bladder, 4 = Breast, 5 = Cervix, 7 = Oesophagus, 8 = Kidney, 9 = Liver, 10 = Leukaemia, 13= Mouth ,oral; 14 = Ovary, 15 = Pancreas, 21 = Stomach
Lung Cancer	11 = Lung, bronchus

Table 12.9: CARET: Classifying the Cancers

12.11.4 NY Wynder Data Cancer Types

Required Cancer Type	Classified Entries
Melanoma	1720 - MALIGNANT MELANOMA OF SKIN OF LIP, 1723 - MALIGNANT MELANOMA OF SKIN OF OTHER AND UNSPECIFIED PARTS OF FACE, 1725 - MALIGNANT MELANOMA OF SKIN OF TRUNK EXCEPT SCROTUM, 1726 - MALIGNANT MELANOMA OF SKIN OF UPPER LIMB INCLUDING SHOULDER, 1727 - MALIGNANT MELANOMA OF SKIN OF LOWER LIMB INCLUDING HIP, 1729 - MELANOMA OF SKIN SITE UNSPECIFIED
Any Type of Cancer	V – Not recorded, 179 - MALIGNANT NEOPLASM OF UTERUS-PART UNS, 185 - MALIGNANT NEOPLASM OF PROSTATE, 193 - MALIGNANT NEOPLASM OF THYROID GLAND, 199 - DISSEMINATED MALIGNANT NEOPLASM, 1162 - LOBOMYCOSIS, 1520 - MALIGNANT NEOPLASM OF DUODENUM, 1533 - MALIGNANT NEOPLASM OF SIGMOID COLON, 1534 - MALIGNANT NEOPLASM OF CECUM, 1539 - MALIGNANT NEOPLASM OF COLON UNSPECIFIED SITE, 1540 - MALIGNANT NEOPLASM OF RECTOSIGMOID JUNCTION, 1541 - MALIGNANT NEOPLASM OF RECTUM, 1561 - MALIGNANT NEOPLASM OF EXTRAHEPATIC BILE DUCTS, 1590 - MALIGNANT NEOPLASM OF INTESTINAL TRACT PART UNSPECIFIED, 1702 - MALIGNANT NEOPLASM OF VERTEBRAL COLUMN EXCLUDING SACRUM AND COCCYX, 1706 - MALIGNANT NEOPLASM OF PELVIC BONES SACRUM AND COCCYX, 1707 - MALIGNANT NEOPLASM OF LONG BONES OF LOWER LIMB, 1709 - MALIGNANT NEOPLASM OF BONE AND ARTICULAR CARTILAGE SITE UNSPECIFIED, 1712 - MALIGNANT NEOPLASM OF CONNECTIVE AND OTHER SOFT TISSUE OF UPPER LIMB INCLUDING SHOULDER, 1713 - MALIGNANT NEOPLASM OF CONNECTIVE AND OTHER SOFT TISSUE OF LOWER LIMB INCLUDING HIP, 1716 - MALIGNANT NEOPLASM OF CONNECTIVE AND OTHER SOFT TISSUE OF PELVIS, 1719 - MALIGNANT NEOPLASM OF CONNECTIVE AND OTHER SOFT TISSUE SITE UNSPECIFIED, 1733 - UNSPECIFIED MALIGNANT NEOPLASM OF SKIN OF OTHER AND UNSPECIFIED PARTS OF FACE, 1734 - UNSPECIFIED MALIGNANT NEOPLASM OF SCALP AND SKIN OF NECK, 1735 - UNSPECIFIED MALIGNANT NEOPLASM OF SKIN OF TRUNK, EXCEPT SCROTUM, 1736 - UNSPECIFIED MALIGNANT NEOPLASM OF SKIN OF UPPER LIMB, INCLUDING SHOULDER, 1737 - UNSPECIFIED MALIGNANT NEOPLASM OF SKIN OF LOWER LIMB, INCLUDING HIP, 1739 - UNSPECIFIED MALIGNANT NEOPLASM OF SKIN, SITE UNSPECIFIED, 1795 - MALIGNANT NEOPLASM OF UTERUS-PART UNS, 1840 - MALIGNANT NEOPLASM OF VAGINA, 1849 - MALIGNANT NEOPLASM OF FEMALE GENITAL ORGAN SITE UNSPECIFIED, 1855 - MALIGNANT NEOPLASM OF PROSTATE, 1857 - MALIGNANT NEOPLASM OF PROSTATE, 1869 - MALIGNANT NEOPLASM OF OTHER AND UNSPECIFIED TESTIS, 1909 - MALIGNANT NEOPLASM OF EYE PART UNSPECIFIED, 1919 - MALIGNANT NEOPLASM OF BRAIN UNSPECIFIED SITE, 1936 - MALIGNANT NEOPLASM OF THYROID GLAND, 1943 - MALIGNANT NEOPLASM OF PITUITARY GLAND AND CRANIOPHARYNGEAL DUCT, 1953 - MALIGNANT NEOPLASM OF PELVIS, 1955 - MALIGNANT NEOPLASM OF LOWER LIMB, 1958 - MALIGNANT NEOPLASM OF OTHER SPECIFIED SITES, 1990 - DISSEMINATED MALIGNANT NEOPLASM, 1991 - OTHER MALIGNANT NEOPLASM OF UNSPECIFIED SITE, 2021 - MYCOSIS FUNGOIDES UNSPECIFIED SITE, 2028 - OTHER MALIGNANT LYMPHOMAS UNSPECIFIED SITE, 2029 - OTHER AND UNSPECIFIED MALIGNANT NEOPLASMS OF LYMPHOID AND HISTIOCYTIC TISSUE UNSPECIFIED SITE, 2030 - MULTIPLE MYELOMA, WITHOUT MENTION OF HAVING ACHIEVED REMISSION,

Required Cancer Type	**Classified Entries**
Any Type of Cancer	2598 - OTHER SPECIFIED ENDOCRINE DISORDERS, 2899 - UNSPECIFIED DISEASES OF BLOOD AND BLOOD-FORMING ORGANS
Smoking Related Cancer	171 - MALIGNANT NEOPLASM OF CONNECTIVE AND OTHER SOFT TISSUE OF HEAD FACE AND NECK, 173 - UNSPECIFIED MALIGNANT NEOPLASM OF SKIN OF LIP, 175 - MALIGNANT NEOPLASM OF NIPPLE AND AREOLA OF MALE BREAST, 1409 - MALIGNANT NEOPLASM OF LIP UNSPECIFIED VERMILION BORDER, 1419 - MALIGNANT NEOPLASM OF TONGUE UNSPECIFIED, 1455 - MALIGNANT NEOPLASM OF PALATE UNSPECIFIED, 1459 - MALIGNANT NEOPLASM OF MOUTH UNSPECIFIED, 1460 - MALIGNANT NEOPLASM OF TONSIL, 1490 - MALIGNANT NEOPLASM OF PHARYNX UNSPECIFIED, 1509 - MALIGNANT NEOPLASM OF ESOPHAGUS UNSPECIFIED SITE, 1519 - MALIGNANT NEOPLASM OF STOMACH UNSPECIFIED SITE, 1550 - MALIGNANT NEOPLASM OF LIVER PRIMARY, 1552 - MALIGNANT NEOPLASM OF LIVER NOT SPECIFIED AS PRIMARY OR SECONDARY, 1560 - MALIGNANT NEOPLASM OF GALLBLADDER, 1579 - MALIGNANT NEOPLASM OF PANCREAS PART UNSPECIFIED, 1600 - MALIGNANT NEOPLASM OF NASAL CAVITIES, 1619 - MALIGNANT NEOPLASM OF LARYNX UNSPECIFIED, 1701 - MALIGNANT NEOPLASM OF MANDIBLE, 1749 - MALIGNANT NEOPLASM OF BREAST (FEMALE) UNSPECIFIED SITE, 1809 - MALIGNANT NEOPLASM OF CERVIX UTERI UNSPECIFIED SITE, 1830 - MALIGNANT NEOPLASM OF OVARY, 1889 - MALIGNANT NEOPLASM OF BLADDER PART UNSPECIFIED, 1890 - MALIGNANT NEOPLASM OF KIDNEY EXCEPT PELVIS, 1950 - MALIGNANT NEOPLASM OF HEAD FACE AND NECK, 1951 - MALIGNANT NEOPLASM OF THORAX, 1952 - MALIGNANT NEOPLASM OF ABDOMEN, 2001 - LYMPHOSARCOMA UNSPECIFIED SITE, 2019 - HODGKIN'S DISEASE UNSPECIFIED TYPE UNSPECIFIED SITE, 2089 - UNSPECIFIED LEUKEMIA, WITHOUT MENTION OF HAVING ACHIEVED REMISSION, 5759 - UNSPECIFIED DISORDER OF GALLBLADDER
Lung Cancer	1629 - MALIGNANT NEOPLASM OF BRONCHUS AND LUNG UNSPECIFIED 1639 - MALIGNANT NEOPLASM OF PLEURA UNSPECIFIED

Table 12.10: NY Wynder: Classifying the Cancers

12.11.5 CREST Data Cancer Types

Required Cancer Type	**Classified Entries**
Melanoma	172.9 - MELANOMA OF SKIN SITE UNSPECIFIED
Any Type of Cancer	153.9 - MALIGNANT NEOPLASM OF COLON UNSPECIFIED SITE 154.3 - MALIGNANT NEOPLASM OF ANUS UNSPECIFIED SITE 170.9 - MALIGNANT NEOPLASM OF BONE AND ARTICULAR CARTILAGE SITE UNSPECIFIED 173.9 - UNSPECIFIED MALIGNANT NEOPLASM OF SKIN, SITE UNSPECIFIED 184.9 - MALIGNANT NEOPLASM OF FEMALE GENITAL ORGAN SITE UNSPECIFIED 185.9 - MALIGNANT NEOPLASM OF PROSTATE 187.9 - MALIGNANT NEOPLASM OF MALE GENITAL ORGAN SITE UNSPECIFIED 191 - MALIGNANT NEOPLASM OF CEREBRUM EXCEPT LOBES AND VENTRICLES 194.9 - MALIGNANT NEOPLASM OF ENDOCRINE GLAND SITE UNSPECIFIED 195.4 - MALIGNANT NEOPLASM OF UPPER LIMB 199 - DISSEMINATED MALIGNANT NEOPLASM 200 - RETICULOSARCOMA UNSPECIFIED SITE
Smoking Related Cancer	146.9 - MALIGNANT NEOPLASM OF OROPHARYNX UNSPECIFIED SITE 149.9 - MALIGNANT NEOPLASM OF ILL-DEFINED SITES WITHIN THE LIP AND ORAL CAVITY 150 - MALIGNANT NEOPLASM OF CERVICAL ESOPHAGUS 151.9 - MALIGNANT NEOPLASM OF STOMACH UNSPECIFIED SITE 155.9 - MALIGNANT NEOPLASM OF LIVER PRIMARY 157.9 - MALIGNANT NEOPLASM OF HEAD OF PANCREAS 159.9 - MALIGNANT NEOPLASM OF ILL-DEFINED SITES WITHIN THE DIGESTIVE ORGANS AND PERITONEUM 161.9 - MALIGNANT NEOPLASM OF LARYNX UNSPECIFIED 174.9 - MALIGNANT NEOPLASM OF BREAST (FEMALE) UNSPECIFIED SITE 182.9 - MALIGNANT NEOPLASM OF CORPUS UTERI EXCEPT ISTHMUS 189 - MALIGNANT NEOPLASM OF KIDNEY EXCEPT PELVIS 193 - MALIGNANT NEOPLASM OF THYROID GLAND 195 - MALIGNANT NEOPLASM OF HEAD FACE AND NECK 195.2 - MALIGNANT NEOPLASM OF ABDOMEN 196.9 - SECONDARY AND UNSPECIFIED MALIGNANT NEOPLASM OF LYMPH NODES SITE UNSPECIFIED

Required Cancer Type	**Classified Entries**
Lung Cancer	162.9 - MALIGNANT NEOPLASM OF BRONCHUS AND LUNG UNSPECIFIED 163.9 - MALIGNANT NEOPLASM OF PLEURA UNSPECIFIED

Table 12.11: CREST: Classifying the Cancers

12.11.6 Canadian Study Data Cancer Types

Required Cancer Type	**Classified Entries**
Melanoma	Too many to list (NOT REQUIRED BY ANY MODEL)
Any Type of Cancer	Too many to list (NOT REQUIRED BY ANY MODEL)
Smoking Related Cancer	Too many to list (NOT REQUIRED BY ANY MODEL)
Lung Cancer	bone, lung, stomach; Adenocarcinoma of lung; Left lung cancer; LUNG; Lung (1st); lung (from asbestos); lung (non small cell); Lung (NSCLC); LUNG (PARTIAL); lung + brain; Lung + Colon; LUNG + OTHER; Lung and Bladder; LUNG AND BRAIN; Lung and Prostate; LUNG CANCER; Lung Cancer, Brain Cancer; Lung(Small Cell); lung, bone; lung, brain; lung, colon; lung, Hodgkin's lymphoma; lung, liver, leukemia; lung, possibly thymus; lung, prostate; lung, throat; lung.brain; lung/bone; LUNG/BRAIN; LUNG/BREAST/BRAIN; lung/kidney; lung/prostate; LUNG/SKIN; LUNG; ESOPHAGUS; Lung-spread to Bone; multiple carcinoid tumours of lungs; non-small cell lung cancer; non-small-cell carcinoma (lungs); pancreatic (spread to lungs or brain); Pancreatic/Lung; SKIN, LUNG, THROAT; PROSTATE, BONE, LUNG; prostate/ lung; PROSTATE/LUNG; prostate; spread to lungs & brain; RARE LUNG; Small cell lung cancer; SMALL LUNG; SMALL-CELL LUNG; liver, lung; Kidney,Lung,Bone; BOWEL LUNG; breast, bone, lung; BREAST/LUNG; breast/ lung; colon, lung; colon-lung; Breast + Lung; PROSTATE AND LUNG; stomach, lung; testicular, lung; stomach/lung/colon

Table 12.12: Canadian Study: Classifying the Cancers

12.12 Dataset Variable Codebook

Across the datasets and published Stata model codes the following variable codes were used;

Variable Name	Variable Code	Variable Classification
Unique ID	ID	Unique ID
Case or Control Status	CaseControl	Case=”Case”, Control = “Control”
Age	AgeRegistered	Age nearest year
Gender	Gender	Male = 1, Female = 2
Ethnicity	EthnicityScale	White = 1 Black = 2 Hispanic = 3 Asian = 4 American India or Alaskan Native = 5 Native Hawaiian or Pacific Islander = 6
BMI	BMI	American BMI calculation
Education	PLCOEducationScale	Less than high-school graduate = 1 High-school graduate = 2 Some training after high school = 3 Some college = 4 College graduate = 5 Postgraduate or professional degree = 6
Prior Tumour	PMT	No = 0, Yes = 1
Smoking Status	SmokingStatus	Never smoker = “No” Former smoker = “Ex” Current smoker = “Yes”
Start Age	StartAge	Age started smoking
Cessation Age	CessationAge	Age stopped smoking
Smoking Duration	Duration	Duration smoking
Cigarettes Per Day	CPD	Average CPD
Pack Years	PackYears	Total pack years
Quit Duration	QuitDuration	Duration since quitting
Environmental Smoke	ETS	No = 0, Yes = 1
Cases of Any Cancer	ACFHC	Family history of all cancers
Age of Onset for Any Cancer	AoOACFHC	Youngest recorded age of onset
Any Cancer w/o melanoma	ACFHCeM	All cancers except melanoma
Age of Onset for Any Cancer	AoOAeMC	Youngest recorded age of onset
Cases of Smoking Related Cancer	SRCC	Smoking related cancers
Age of Onset for Smoking Cancer	AoOSRC	Youngest recorded age of onset
Cases of Lung Cancer	LCC	Lung cancer
Age of Onset of Lung Cancer	AoOLCC	Youngest recorded age of onset
Asbestos Exposure	Asbestos	No = 0, Yes = 1
Dust	Dust	No = 0, Yes = 1
Hay fever	Hayfever	No = 0, Yes = 1
Asthma	Asthma	No = 0, Yes = 1
Emphysema	Emphysema	No = 0, Yes = 1
COPD	COPD	No = 0, Yes = 1
Pneumonia	Pneumonia	No = 0, Yes = 1

Table 12.13: Variable Code Book

12.13 Model Codes

PLCO(2012) Model Stata Code

E. Gray

2 Jun 2016

*PLCO 2012 Model Formula

*Import the Dataset

/*DATASET IS IMPORTED HERE* /

*Never Smokers Excluded

```
.  drop if SmokingStatus == "No"
(972 observations deleted)
```

*Coefficient Preperation

*Ethnicity Coefficient - As categorised in the original article.

```
.          foreach var of varlist ID {
.         gen EthnicityCoefficient = 0 if EthnicityScale == 1
.         replace EthnicityCoefficient = 0.3944778 if EthnicityScale == 2
.         replace EthnicityCoefficient = -0.7434744 if EthnicityScale == 3
.         replace EthnicityCoefficient = -0.466585 if EthnicityScale == 4
.         replace EthnicityCoefficient = 1.027152 if EthnicityScale == 5
.         replace EthnicityCoefficient = 0 if EthnicityScale == 6
.         }
(120 missing values generated)
(19 real changes made)
(8 real changes made)
(86 real changes made)
(7 real changes made)
(0 real changes made)
```

*Age Coefficient

```
.          foreach var of varlist ID {
.         gen AgeCoefficient = (AgeRegistered - 62) * 0.0778868
.         }
```

*Education Coefficient - As categorised in the original article

```
.          foreach var of varlist ID {
.         gen EducationCoefficient = -0.0812744 * (PLCOEducation - 4)
.         }
```

*BMI Coefficient

```
.          foreach var of varlist ID {
.         gen BMICoefficient = -0.0274194 * (BMI - 27)
.         }
```

*COPD Coefficient

```
.          foreach var of varlist ID {
.         gen COPDCoefficient = 0.3553063 * COPD
.         }
```

*Personal History of Cancer Coefficient

```
.          foreach var of varlist ID {
.         gen PHCCoefficient = 0.4589971 * PMT
.         }
```

*Family Hisoty of Lung Cancer Coefficient

```
.          foreach var of varlist ID {
.         gen FHoLCCoefficient = 0
.         replace FHoLCCoefficient = 0.587185 if LCC > 0
.         }
(216 real changes made)
```

*Smoking Status Coefficient

```
.          foreach var of varlist ID {
.         gen SSCoefficient = 0 if SmokingStatus == "Ex"
.         replace SSCoefficient = 0.2597431 if SmokingStatus == "Yes"
.         }
(552 missing values generated)
(552 real changes made)
```

*Smoking Intensity Coefficient

```
.          foreach var of varlist ID {
.         gen SICoefficient = -1.822606 * (((CPD/10)^-1)- 0.4021541613)
.         }
```

*Smoking Duration

```
.          foreach var of varlist ID {
.         gen SDCoefficient = 0.0317321 * (Duration - 27)
.         }
```

*Quit Duration Coefficient

```
.          foreach var of varlist ID {
.         gen QuitCoefficient = 0 if SmokingStatus == "Yes"
.         replace QuitCoefficient = -0.0308572 * (QuitDuration - 10) if SmokingStatus == "Ex"
.         }
(877 missing values generated)
(877 real changes made)
```

*Sum of the predictors

```
.                  foreach var of varlist ID {
.                 gen Logit = -4.532506 + AgeCoefficient + EthnicityCoefficient + EducationCoefficient ///
                + BMICoefficient + COPDCoefficient + PHCCoefficient + FHoLCCoefficient + SSCoefficient ///
                + SICoefficient + SDCoefficient + QuitCoefficient

.                 }
```

*Six Year Risk for PLCO_2012

```
.  foreach var of varlist ID {
. gen PLCO2012Risk = (exp(Logit)/(1+exp(Logit)))*100
. }
```

*Clean the dataset

```
.  drop (EthnicityCoefficient - Logit)
```

PLCO(2014) Model Stata Code

E. Gray

2 Jun 2016

*PLCO 2014 Model Formula

*Import the Dataset

/DATASET IS IMPORTED HERE /

*Variable Preperation

```
.          keep ID AgeRegistered EthnicityScale PLCOEducation BMI SmokingStatus ///
         CessationAge Duration QuitDuration CPD LCC PMT COPD
```

*Coefficient Preperation

*Ethnicity Coefficient - As categorised in the original article.

```
.                 foreach var of varlist ID {
.                gen EthnicityCoefficient = 0 if EthnicityScale == 1
.                replace EthnicityCoefficient = 0.3211605 if EthnicityScale == 2
.                replace EthnicityCoefficient = -0.8203332 if EthnicityScale == 3
.                replace EthnicityCoefficient = -0.5241286 if EthnicityScale == 4
.                replace EthnicityCoefficient = 0.952699 if EthnicityScale == 5
.                replace EthnicityCoefficient = -1.364379 if EthnicityScale == 6
.                }
(328 missing values generated)
(45 real changes made)
(25 real changes made)
(249 real changes made)
(9 real changes made)
(0 real changes made)
```

*Age Coefficient

```
.                 foreach var of varlist ID {
.                gen AgeCoefficient = (AgeRegistered - 62) * 0.079597
.                }
```

*Education Coefficient - As categorised in the original article

```
.                 foreach var of varlist ID {
.                gen EducationCoefficient = -0.0879289 * (PLCOEducation - 4)
.                }
```

*BMI Coefficient

```
.                 foreach var of varlist ID {
.                gen BMICoefficient = -0.028948 * (BMI - 27)
.                }
```

*COPD Coefficient

```
.                 foreach var of varlist ID {
.                gen COPDCoefficient = 0.3457265 * COPD
.                }
```

*Personal History of Cancer Coefficient

```
.                 foreach var of varlist ID {
.                gen PHCCoefficient = 0.4845208 * PMT
.                }
```

*Family Hisoty of Lung Cancer Coefficient

```
.                 foreach var of varlist ID {
.                gen FHoLCCoefficient = 0
.                replace FHoLCCoefficient = 0.5856777 if LCC > 0
.                }
(318 real changes made)
```

*Smoking Status Coefficient

```
.                  foreach var of varlist ID {
.                 gen SSCoefficient = 0
.                 replace SSCoefficient = 2.799727 if SmokingStatus == "Yes"
.                 replace SSCoefficient = 2.542472 if SmokingStatus == "Ex"
.                 }
(552 real changes made)
(877 real changes made)
```

*Smoking Intensity Coefficient

```
.                  foreach var of varlist ID {
.                 gen SICoefficient = 0
.                 replace SICoefficient = -0.1815486 * (((CPD/100)^-1)- 4.021541613) ///
                 if SmokingStatus == "Yes" | SmokingStatus == "Ex"

.                 }
(1429 real changes made)
```

*Smoking Duration

```
.                  foreach var of varlist ID {
.                 gen SDCoefficient = 0
.                 replace SDCoefficient = 0.0305566 * (Duration - 27) if SmokingStatus == "Yes" | ///
                 SmokingStatus == "Ex"

.                 }
(1400 real changes made)
```

*Quit Duration Coefficient

```
.                  foreach var of varlist ID {
.                 gen QuitCoefficient = 0
.                 replace QuitCoefficient = -0.0321362 * (QuitDuration - 8.593417626) ///
                 if SmokingStatus == "Ex"

.                 }
(877 real changes made)
```

*Sum of the predictors

```
.                       foreach var of varlist ID {
.                      gen Logit = -7.02198 + AgeCoefficient + EthnicityCoefficient + EducationCoefficient ///
                                  + BMICoefficient + COPDCoefficient + PHCCoefficient + FHoLCCoefficient ///
                                  + SSCoefficient + SICoefficient + SDCoefficient + QuitCoefficient

.                       }
```

*Six Year Risk for PLCO_2014

```
.  foreach var of varlist ID {
. gen PLCO2014Risk = (exp(Logit)/(1+exp(Logit)))*100
. }
```

*Clean the dataset

```
.  drop (EthnicityCoefficient - Logit)
```

African-American Model Stata Code

E. Gray

2 Jun 2016

*African American Five Year Risk

*Variable Preperation

```
.  keep ID CaseControl AgeRegistered Gender SmokingStatus CessationAge QuitDuration PackYears Dust Asbestos HayFever COPD
```

*Removed if under 20

```
.  drop if AgeRegistered < 20
(1 observation deleted)
```

*Pack Years, Quit Duration and COPD Coefficient

```
.          foreach var of varlist ID {
.         gen SSCOPD = 1
.         replace SSCOPD = 2.56 if PackYears >= 13.3 & PackYears < 26.4 & COPD == 0
.         replace SSCOPD = 2.69 if PackYears >= 26.4 & COPD == 0  & QuitDuration >= 3 & CessationAge < 56
.         replace SSCOPD = 6.06 if PackYears >= 26.4 & COPD == 0  & QuitDuration < 3 & CessationAge < 56
.         replace SSCOPD = 10.01 if PackYears >= 26.4 & COPD == 0  & CessationAge >= 56
.         replace SSCOPD = 20 if PackYears < 26.4 & COPD == 1
.         replace SSCOPD = 21.38 if PackYears >= 13.3 & PackYears < 26.4 & COPD == 0  & QuitDuration >= 23
.         replace SSCOPD = 34.29 if PackYears >= 26.4 & COPD == 1  & SmokingStatus == "Yes"
.         }
(100 real changes made)
(143 real changes made)
(2 real changes made)
(488 real changes made)
(5 real changes made)
(59 real changes made)
(31 real changes made)
```

*Dust Exposure Coefficient

```
.          foreach var of varlist ID {
.         gen DustCoefficient = 1
.         replace DustCoefficient = 1.46 if Dust == 1
.         }
(9 real changes made)
```

*Asbestos Exposure Coefficient

```
.          foreach var of varlist ID {
.         gen AsbestosCoefficient = 1.
.         replace AsbestosCoefficient = 1.38 if Asbestos == 1
.         }
(9 real changes made)
```

*Hay Fever Coefficient

```
.          foreach var of varlist ID {
.         gen HFCoefficient = 1
.         replace HFCoefficient = 0.66 if HayFever == 1
.         }
(0 real changes made)
```

*Attrituable Risk by Gender Coefficient

*Male - 1, Female - 2

```
.          foreach var of varlist ID {
.         gen AttRisk = 0.79 if Gender == 1
.         replace AttRisk = 0.59 if Gender == 2
.         }
(199 missing values generated)
(199 real changes made)
```

*Age and Gender LC Incidence Rates

```
.          foreach var of varlist ID {
.         gen AGLCIRates = 0
.         replace AGLCIRates = 0.2015/100000 if Gender == 1 & AgeRegistered >=20 & AgeRegistered <25
.         replace AGLCIRates = 0.2/100000 if Gender == 2 & AgeRegistered >=20 & AgeRegistered <25
.         replace AGLCIRates = 0.454/100000 if Gender == 1 & AgeRegistered >=25 & AgeRegistered <30
.         replace AGLCIRates = 0.4212/100000 if Gender == 2 & AgeRegistered >=25 & AgeRegistered <30
```

```
.         replace AGLCIRates = 1.1557/100000 if Gender == 1 & AgeRegistered >=30 & AgeRegistered <35
.         replace AGLCIRates = 0.6739/100000 if Gender == 2 & AgeRegistered >=30 & AgeRegistered <35
.         replace AGLCIRates = 3.988/100000 if Gender == 1 & AgeRegistered >=35 & AgeRegistered <40
.         replace AGLCIRates = 4.9258/100000 if Gender == 2 & AgeRegistered >=35 & AgeRegistered <40
.         replace AGLCIRates = 19.2688/100000 if Gender == 1 & AgeRegistered >=40 & AgeRegistered <45
.         replace AGLCIRates = 15.5789/100000 if Gender == 2 & AgeRegistered >=40 & AgeRegistered <45
.         replace AGLCIRates = 53.1998/100000 if Gender == 1 & AgeRegistered >=45 & AgeRegistered <50
.         replace AGLCIRates = 33.6673/100000 if Gender == 2 & AgeRegistered >=45 & AgeRegistered <50
.         replace AGLCIRates = 114.9708/100000 if Gender == 1 & AgeRegistered >=50 & AgeRegistered <55
.         replace AGLCIRates = 64.067/100000 if Gender == 2 & AgeRegistered >=50 & AgeRegistered <55
.         replace AGLCIRates = 230.228/100000 if Gender == 1 & AgeRegistered >=55 & AgeRegistered <60
.         replace AGLCIRates = 108.8192/100000 if Gender == 2 & AgeRegistered >=55 & AgeRegistered <60
.         replace AGLCIRates = 336.7714/100000 if Gender == 1 & AgeRegistered >=60 & AgeRegistered <65
.         replace AGLCIRates = 160.2701/100000 if Gender == 2 & AgeRegistered >=60 & AgeRegistered <65
.         replace AGLCIRates = 452.6758/100000 if Gender == 1 & AgeRegistered >=65 & AgeRegistered <70
.         replace AGLCIRates = 236.0585/100000 if Gender == 2 & AgeRegistered >=65 & AgeRegistered <70
.         replace AGLCIRates = 572.0534/100000 if Gender == 1 & AgeRegistered >=70 & AgeRegistered <75
.         replace AGLCIRates = 272.7912/100000 if Gender == 2 & AgeRegistered >=70 & AgeRegistered <75
.         replace AGLCIRates = 690.5254/100000 if Gender == 1 & AgeRegistered >=75 & AgeRegistered <80
.         replace AGLCIRates = 302.721/100000 if Gender == 2 & AgeRegistered >=75 & AgeRegistered <80
.         replace AGLCIRates = 647.7919/100000 if Gender == 1 & AgeRegistered >=80 & AgeRegistered <85
.         replace AGLCIRates = 273.0735/100000 if Gender == 2 & AgeRegistered >=80 & AgeRegistered <85
.         replace AGLCIRates = 650.5089/100000 if Gender == 1 & AgeRegistered >=85
.         replace AGLCIRates = 229.61176/100000 if Gender == 2 & AgeRegistered >=85
.         }
(6 real changes made)
(9 real changes made)
(15 real changes made)
(9 real changes made)
(16 real changes made)
(6 real changes made)
(15 real changes made)
(10 real changes made)
(19 real changes made)
(5 real changes made)
(47 real changes made)
(13 real changes made)
(71 real changes made)
(18 real changes made)
(108 real changes made)
(16 real changes made)
(92 real changes made)
(23 real changes made)
(112 real changes made)
(26 real changes made)
(97 real changes made)
(23 real changes made)
(67 real changes made)
(18 real changes made)
(26 real changes made)
(12 real changes made)
(17 real changes made)
(11 real changes made)
```

*Age and Gender Without LC Incidence Rates

```
.          foreach var of varlist ID {
.         gen AGWLCIRates = 0
.         replace AGWLCIRates = 224.5/100000 if Gender == 1 & AgeRegistered >=20 & AgeRegistered <25
.         replace AGWLCIRates = 70.6/100000 if Gender == 2 & AgeRegistered >=20 & AgeRegistered <25
.         replace AGWLCIRates = 249.6/100000 if Gender == 1 & AgeRegistered >=25 & AgeRegistered <30
.         replace AGWLCIRates = 93.6/100000 if Gender == 2 & AgeRegistered >=25 & AgeRegistered <30
.         replace AGWLCIRates = 262.5/100000 if Gender == 1 & AgeRegistered >=30 & AgeRegistered <35
.         replace AGWLCIRates = 132.5/100000 if Gender == 2 & AgeRegistered >=30 & AgeRegistered <35
.         replace AGWLCIRates = 334.8/100000 if Gender == 1 & AgeRegistered >=35 & AgeRegistered <40
.         replace AGWLCIRates = 213.1/100000 if Gender == 2 & AgeRegistered >=35 & AgeRegistered <40
.         replace AGWLCIRates = 507.1/100000 if Gender == 1 & AgeRegistered >=40 & AgeRegistered <45
.         replace AGWLCIRates = 325.1/100000 if Gender == 2 & AgeRegistered >=40 & AgeRegistered <45
.         replace AGWLCIRates = 798.1/100000 if Gender == 1 & AgeRegistered >=45 & AgeRegistered <50
.         replace AGWLCIRates = 486.7/100000 if Gender == 2 & AgeRegistered >=45 & AgeRegistered <50
.         replace AGWLCIRates = 1220.7/100000 if Gender == 1 & AgeRegistered >=50 & AgeRegistered <55
.         replace AGWLCIRates = 697.2/100000 if Gender == 2 & AgeRegistered >=50 & AgeRegistered <55
.         replace AGWLCIRates = 1678/100000 if Gender == 1 & AgeRegistered >=55 & AgeRegistered <60
.         replace AGWLCIRates = 671.3/100000 if Gender == 2 & AgeRegistered >=55 & AgeRegistered <60
.         replace AGWLCIRates = 2437/100000 if Gender == 1 & AgeRegistered >=60 & AgeRegistered <65
.         replace AGWLCIRates = 1445.8/100000 if Gender == 2 & AgeRegistered >=60 & AgeRegistered <65
.         replace AGWLCIRates = 3312/100000 if Gender == 1 & AgeRegistered >=65 & AgeRegistered <70
.         replace AGWLCIRates = 2047.3/100000 if Gender == 2 & AgeRegistered >=65 & AgeRegistered <70
.         replace AGWLCIRates = 4818.6/100000 if Gender == 1 & AgeRegistered >=70 & AgeRegistered <75
.         replace AGWLCIRates = 3020.5/100000 if Gender == 2 & AgeRegistered >=70 & AgeRegistered <75
.         replace AGWLCIRates = 6952.6/100000 if Gender == 1 & AgeRegistered >=75 & AgeRegistered <80
.         replace AGWLCIRates = 4495.8/100000 if Gender == 2 & AgeRegistered >=75 & AgeRegistered <80
.         replace AGWLCIRates = 9796.1/100000 if Gender == 1 & AgeRegistered >=80 & AgeRegistered <85
.         replace AGWLCIRates = 6654.1/100000 if Gender == 2 & AgeRegistered >=80 & AgeRegistered <85
.         replace AGWLCIRates = 14876.3/100000 if Gender == 1 & AgeRegistered >=85
.         replace AGWLCIRates = 13616.7/100000 if Gender == 2 & AgeRegistered >=85
.         }
(6 real changes made)
(9 real changes made)
(15 real changes made)
(9 real changes made)
(16 real changes made)
(6 real changes made)
(15 real changes made)
(10 real changes made)
(19 real changes made)
(5 real changes made)
(47 real changes made)
(13 real changes made)
(71 real changes made)
(18 real changes made)
(108 real changes made)
(16 real changes made)
(92 real changes made)
(23 real changes made)
```

```
(112 real changes made)
(26 real changes made)
(97 real changes made)
(23 real changes made)
(67 real changes made)
(18 real changes made)
(26 real changes made)
(12 real changes made)
(17 real changes made)
(11 real changes made)
```

*Hazard Rate

```
.                  foreach var of varlist ID {
.                 gen HazardRate = AGLCIRates * (1 - AttRisk)
.                 }
```

*Odds Rate

```
.                  gen OddsRate = SSCOPD * DustCoefficient * AsbestosCoefficient * HFCoefficient
```

*African-American Model Five Year Risk

```
.  gen AfricanAmericanRisk = ((HazardRate * OddsRate)/((HazardRate * OddsRate)+ AGWLCIRates))* ///
 (1 - exp((HazardRate * OddsRate* -5)-(5 * AGWLCIRates)))* 100
```

*Clean the data

```
.  drop (SSCOPD - OddsRate)
```

Bach Model Stata Code

*Bach Model Code

*Exclude ineligible participants

```
.  keep ID CaseControl AgeRegistered Gender CPD SmokingStatus Duration QuitDuration PackYears Asbestos

. keep if AgeRegistered >= 50 & AgeRegistered <= 75 & !(SmokingStatus == "No") & PackYears >= 30
(231 observations deleted)
```

*Redefine Gender for the linear predictor - Male assigned 0, Female 1.

```
.          foreach var of varlist ID {
.         gen GenderScore = 0 if Gender == 1
.         replace GenderScore = 1 if Gender == 2
.         }
(570 missing values generated)
(570 real changes made)
```

*Calculate the One Year Risk of Incidence of Lung Cancer

*One Year Risk of Incidence Linear Predictor

```
.                  foreach var of varlist ID {
.                 gen Incidence_LP= -9.7960571 + (0.060818386 * CPD) - (0.00014652216 * max(CPD - 15,0)^3) ///
                  + (0.00018486938 * max(CPD - 20.185718, 0)^3) - (3.8347226e-005 * max(CPD - 40, 0)^3) +  ///
                  (0.11425297 *Duration) - (8.0091477e-005 * max(Duration - 27.6577, 0)^3) + ///
                  (0.00017069483 * max(Duration - 40, 0)^3) - (9.0603358e-005 * max(Duration - 50.910335, 0)^3) - ///
                  (0.085684793 * QuitDuration) + (0.0065499693 * max(QuitDuration, 0)^3) - ///
                  (0.0068305845 * max(QuitDuration - 0.50513347, 0)^3) + (0.00028061519 * max(QuitDuration - 12.295688, 0)^3) ///
                  + (0.070322812 * AgeRegistered - (9.382122e-005) * max(AgeRegistered - 53.459001, 0)^3) + (0.00018282661 * ///
                  max(AgeRegistered - 61.954825, 0)^3) - (8.9005389e-005 * max(AgeRegistered - 70.910335, 0)^3) ///
                  + (0.2153936*Asbestos) - (0.05827261*GenderScore)

.                  }
```

*One Year Risk of Incidence Final Result

```
.                  foreach var of varlist ID {
.                 gen Bach_Incidence =(1- 0.99629^exp(Incidence_LP))
.                 }
```

*Calculate One Year Risk of Death in the Absence of Lung Cancer

*Competing Risk of Death in the Absence of Lung Cancer Linear Predictor

```
.                  foreach var of varlist ID {
.                 gen Death_LP = -7.2036219 + (0.015490665 * CPD) - (0.00001737645 * max(CPD - 15,0)^3) + (0.000021924149 * ///
                  max(CPD - 20.185718, 0)^3) - (0.0000045476985 * max(CPD - 40, 0)^3) + (0.020041889 *Duration) + ///
                  (0.0000065443781 * max(Duration - 27.6577, 0)^3) - (0.000013947696 * max(Duration - 40, 0)^3) + ///
                  (0.0000074033175 * max(Duration - 50.910335, 0)^3) - (0.023358962 * QuitDuration) + (0.0019208669 * ///
                  max(QuitDuration, 0)^3) - (0.0020031611 * max(QuitDuration - 0.50513347, 0)^3) + (0.000082294194 * ///
                  max(QuitDuration - 12.295688, 0)^3) + (0.099168033 * AgeRegistered) + (0.0000062174577 * ///
                  max(AgeRegistered - 53.459001, 0)^3) - (0.000012115774 * max(AgeRegistered - 61.954825, 0)^3) + ///
                  (0.0000058983164 * max(AgeRegistered - 70.910335, 0)^3) + (0.06084611*Asbestos) - (0.49042298*GenderScore)

.                  }
```

*One Year Risk of Death without Diagnosis Fina Result

```
.                  gen Bach_No_Diagnosis =(1- 0.99629^exp(Death_LP))

.                         foreach var of varlist ID {
.                        gen Bach_Absolute_Risk = Bach_Incidence * (1 - Bach_No_Diagnosis)
.                        }
```

*Final Percentages

```
.  foreach var of varlist ID {
. gen Bach_Incidence_Percentage = Bach_Incidence * 100
. gen Bach_Absolute_Risk_Percentage = Bach_Absolute_Risk * 100
. }
```

*Clean the dataset

```
.  drop (GenderScore - Bach_Absolute_Risk)
```

*To run recursively update age, smoking duration, and quit duration dependant on the part icipant

*Example for 10 years

```
.  forvalues i = 1/10 {
```

```
. gen AgeRegistered_`i' = AgeRegistered + `i'
. gen Duration`i' = Duration + `i' if SmokingStatus == "Yes"
. replace Duration`i' = Duration if SmokingStatus == "Ex"
. gen QuitDuration`i' = QuitDuration + `i' if SmokingStatus == "Ex"
. replace QuitDuration`i' = QuitDuration if SmokingStatus == "Yes"
. }
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
(353 missing values generated)
(353 real changes made)
(1491 missing values generated)
(1491 real changes made)
```

*Formula can be applied to calcualte risk and absolute risk in second year and so forth

*Then run recursively to calculate new risks

LLP Model Stata Code

*LLP Model Five Year Risk

*Pre-amble

```
. keep ID CaseControl AgeRegistered Gender Asbestos Pneumonia PMT Duration LCC AoOLCC
```

*Remove Participants who are ineligible for the model

```
. drop if AgeRegistered < 40 | AgeRegistered > 80
(3 observations deleted)
```

*Where Age of Onset for Youngest Family Member with Lung Cancer assign 0

```
. foreach var of varlist ID {
. replace AoOLCC = 0 if missing(AoOLCC)
. }
(21 real changes made)
```

*Assign participants an age of "Their Age = 0.5' (Age Midpoint)

```
.          foreach var of varlist ID {
.         replace AgeRegistered = floor(AgeRegistered) + 0.5
.         }
(28 real changes made)
```

*Determine with participant how long remaining in that age group and amount \

*of time for next 5 year in the next age group

```
.          foreach var of varlist ID {
.         gen LowerGroupTime = (floor(AgeRegistered/5)*5) + 5 - AgeRegistered
.         gen UpperGroupTime = 5 - LowerGroupTime
.         }
```

*Assign a category based on Gender and Age

```
.                  forvalues i = 1/16 {
.                 gen AgeGender_`i' = 0
.                 }

.                 foreach var of varlist ID {
.                 replace AgeGender_1 = 1 if Gender == 1 & AgeRegistered > 40 & AgeRegistered < 45
.                 replace AgeGender_2 = 1 if Gender == 1 & AgeRegistered > 45 & AgeRegistered < 50
.                 replace AgeGender_3 = 1 if Gender == 1 & AgeRegistered > 50 & AgeRegistered < 55
.                 replace AgeGender_4 = 1 if Gender == 1 & AgeRegistered > 55 & AgeRegistered < 60
.                 replace AgeGender_5 = 1 if Gender == 1 & AgeRegistered > 60 & AgeRegistered < 65
.                 replace AgeGender_6 = 1 if Gender == 1 & AgeRegistered > 65 & AgeRegistered < 70
.                 replace AgeGender_7 = 1 if Gender == 1 & AgeRegistered > 70 & AgeRegistered < 75
.                 replace AgeGender_8 = 1 if Gender == 1 & AgeRegistered > 75 & AgeRegistered < 80
.                 replace AgeGender_9 = 1 if Gender == 2 & AgeRegistered > 40 & AgeRegistered < 45
.                 replace AgeGender_10 = 1 if Gender == 2 & AgeRegistered > 45 & AgeRegistered < 50
.                 replace AgeGender_11 = 1 if Gender == 2 & AgeRegistered > 50 & AgeRegistered < 55
.                 replace AgeGender_12 = 1 if Gender == 2 & AgeRegistered > 55 & AgeRegistered < 60
.                 replace AgeGender_13 = 1 if Gender == 2 & AgeRegistered > 60 & AgeRegistered < 65
.                 replace AgeGender_14 = 1 if Gender == 2 & AgeRegistered > 65 & AgeRegistered < 70
.                 replace AgeGender_15 = 1 if Gender == 2 & AgeRegistered > 70 & AgeRegistered < 75
.                 replace AgeGender_16 = 1 if Gender == 2 & AgeRegistered > 75 & AgeRegistered < 80
.                 }
(1 real change made)
(1 real change made)
(4 real changes made)
(5 real changes made)
(2 real changes made)
(0 real changes made)
(1 real change made)
(0 real changes made)
(1 real change made)
(5 real changes made)
(4 real changes made)
(3 real changes made)
(0 real changes made)
(0 real changes made)
(1 real change made)
(0 real changes made)

.                 mkmat AgeGender_1 - AgeGender_16, matrix(Gender)
```

*Matrix of Incidence Coefficients

```
.                  matrix IncidenceRate = (-9.06\-8.16\-7.31\-6.63\-5.97\-5.56\-5.31\-4.83\-9.9\-8.06\-7.46\-6.5\-6.22\-5.99\-5.49\-5.23)
```

*Matrix of Incidence Coefficients in Next Group

```
.                  matrix NextIncidenceRate = (-8.16\-7.31\-6.63\-5.97\-5.56\-5.31\-4.83\-4.68\-8.06\-7.46\-6.5\-6.22\-5.99\-5.49\-5.23\-5.4
```

*Coefficients

```
.                  matrix A = Gender * IncidenceRate

.                 matrix B = Gender * NextIncidenceRate

.                 svmat double A, name(AgeIncidence)
```

```
.                svmat double B, name(NextAgeIncidence)
```

*Alpha Score based on incidences

```
.                 foreach var of varlist ID {
.                gen Alpha = ((LowerGroupTime * AgeIncidence) + (UpperGroupTime * NextAgeIncidence))/5
.                }
```

*Model Coefficients

*Smoking Duration

```
.          foreach var of varlist ID {
.         gen SmokingDurationCoefficient = 0 if Duration == 0
.         replace SmokingDurationCoefficient = 0.769 if Duration > 0 & Duration <= 20
.         replace SmokingDurationCoefficient = 1.452 if Duration > 20 & Duration <= 40
.         replace SmokingDurationCoefficient = 2.507 if Duration > 40 & Duration <= 60
.         replace SmokingDurationCoefficient = 2.724 if Duration > 60
.         }
(23 missing values generated)
(2 real changes made)
(14 real changes made)
(7 real changes made)
(0 real changes made)
```

*Pneumonia Coefficent - 1 if present, 0 otherwise

```
.          foreach var of varlist ID {
.         gen PneumoniaCoefficient = 0.602 * Pneumonia
.         }
```

*Asbestos Coefficent - 1 if present, 0 otherwise

```
.          foreach var of varlist ID {
.         gen AsbestosCoefficient = 0.634 * Asbestos
.         }
```

*Prior Malignant Tumour Coefficent - 1 if present, 0 otherwise

```
.          foreach var of varlist ID {
.         gen PMTCoefficient = 0.675 * PMT
.         }
```

*Family History of Lung Cancer with Early or Late Onset

```
.          foreach var of varlist ID {
.         gen FHCoefficient = 0
.         replace FHCoefficient = 0.703 if LCC > 0 & AoOLCC < 60
.         replace FHCoefficient = 0.168 if LCC > 0 & AoOLCC >= 60
.         }
(7 real changes made)
(0 real changes made)
```

*LLP Model Linear Predictor

```
.  foreach var of varlist ID {
. gen LLPPredictor = Alpha + SmokingDurationCoefficient + ///
 PneumoniaCoefficient + AsbestosCoefficient + PMTCoefficient + FHCoefficient

. }
```

*Final LLP 5 Year Risk

```
.  foreach var of varlist ID {
. gen LLP5YearRisk = (1/(1 + exp(LLPPredictor * - 1))) * 100
. }
```

*Clean the data

```
.  drop LowerGroupTime - LLPPredictor
```

Spitz Model Stata Code

*Spitz Model One Year Risk

*Pre-amble

*ETS = Environmental Tobacco Smoke, SRCC = Count of Relatives with Smoking Related Cancer , LCC = Count of Relatives with Lung Cancer

```
. keep ID CaseControl AgeRegistered Gender SmokingStatus CessationAge PackYears ETS SRCC LCC HayFever Emphysema Dust Asbestos
```

*Drop ineligible participants

```
. foreach var of varlist ID {
. drop if AgeRegistered < 40
. }
(8 observations deleted)
```

*Model Coefficients

*Former Smoker Coefficient

```
.          foreach var of varlist ID {
.         gen CessationCoefficient = 1 if !(SmokingStatus == "Ex") | (SmokingStatus == "Ex" & Cessation < 42)
.         replace CessationCoefficient = 1.24 if SmokingStatus == "Ex" & CessationAge >= 42 & CessationAge <= 53
.         replace CessationCoefficient = 1.50 if SmokingStatus == "Ex" & CessationAge > 53
.         }
(79 missing values generated)
(23 real changes made)
(56 real changes made)
```

*Current Smokers Pack Years Coefficient

```
.          foreach var of varlist ID {
.         gen PYCoefficient = 1 if !(SmokingStatus == "Yes") | (SmokingStatus == "Yes" & PackYears < 28)
.         replace PYCoefficient = 1.25 if SmokingStatus == "Yes" & PackYears >= 28 & PackYears < 42
.         replace PYCoefficient = 1.45 if SmokingStatus == "Yes" & PackYears >= 42 & PackYears < 57.5
.         replace PYCoefficient = 1.85 if SmokingStatus == "Yes" & PackYears >= 57.5
.         }
(66 missing values generated)
(6 real changes made)
(26 real changes made)
(34 real changes made)
```

*Emphysema Coefficient for Former and Current Smokers

```
.          foreach var of varlist ID {
.         gen EmphysemaCoefficient = 1 if SmokingStatus == "No" | Emphysema == 0
.         replace EmphysemaCoefficient = 2.65 if SmokingStatus == "Ex" & Emphysema == 1
.         replace EmphysemaCoefficient = 2.13 if SmokingStatus == "Yes" & Emphysema == 1
.         }
(3 missing values generated)
(2 real changes made)
(1 real change made)
```

*ETS in never-smokers

```
.          foreach var of varlist ID {
.         gen ETSCoefficient = 1 if !(SmokingStatus == "No") | ETS == 0
.         replace ETSCoefficient = 1.8 if SmokingStatus == "No" & ETS == 1
.         }
(26 missing values generated)
(26 real changes made)
```

*Dust Coefficient in Former and Current Smokers

```
.          foreach var of varlist ID {
.         gen DustCoefficient = 1 if SmokingStatus == "No" | Dust == 0
.         replace DustCoefficient = 1.59 if SmokingStatus == "Ex" & Dust == 1
.         replace DustCoefficient = 1.36 if SmokingStatus == "Yes" & Dust == 1
.         }
(3 missing values generated)
(2 real changes made)
(1 real change made)
```

*Asbestos Exposure in Current Smokers

```
.          foreach var of varlist ID {
.         gen AsbestosCoefficient = 1 if !(SmokingStatus == "Yes") | Asbestos == 0
.         replace AsbestosCoefficient = 1.51 if SmokingStatus == "Yes" & Asbestos == 1
.         }
(1 missing value generated)
(1 real change made)
```

*Family History of Cancers (Different based on smoking status)

```
.          foreach var of varlist ID {
.         gen FHCoefficient = 1
.         replace FHCoefficient = 2 if SmokingStatus == "No" & SRCC >= 2
.         replace FHCoefficient = 1.59 if SmokingStatus == "Ex" & SRCC >= 2
.         replace FHCoefficient = 1.47 if SmokingStatus == "Yes" & LCC >= 1
.         }
(3 real changes made)
```

```
(3 real changes made)
(9 real changes made)
```

*Hay Fever Coefficient in ever-smokers - as no exposure

```
.          foreach var of varlist ID {
.         gen HFCoefficient = 1 if SmokingStatus == "No" | HayFever == 1
.         replace HFCoefficient = 1.45 if SmokingStatus == "Ex" & HayFever == 0
.         replace HFCoefficient = 1.49 if SmokingStatus == "Yes" & HayFever == 0
.         }
(163 missing values generated)
(91 real changes made)
(72 real changes made)
```

*Linear Predictor of Coefficients

```
.          foreach var of varlist ID {
.         gen OddsRatio = CessationCoefficient * PYCoefficient * EmphysemaCoefficient * ETSCoefficient * ///
         DustCoefficient * AsbestosCoefficient * FHCoefficient * HFCoefficient

.         }
```

*Clean the dataset

```
.          drop (CessationCoefficient - HFCoefficient)
```

*Incidence Scaling Factors

*Attributable Risk by Smoking Status

```
.                  foreach var of varlist ID {
.                 gen AttRisk = 0.4751 if SmokingStatus == "No"
.                 replace AttRisk = 0.45352 if SmokingStatus == "Ex"
.                 replace AttRisk = 0.51404 if SmokingStatus == "Yes"
.                 }
(163 missing values generated)
(91 real changes made)
(72 real changes made)
```

*Gender and Smoking Status Risk Factor - Male = 1, Female = 2

```
.                  foreach var of varlist ID {
.                 gen GSSRisk = 0.21 if SmokingStatus == "No" & Gender == 1
.                 replace GSSRisk = 3.17 if SmokingStatus == "Ex" & Gender == 1
.                 replace GSSRisk = 3.88 if SmokingStatus == "Yes" & Gender == 1
.                 replace GSSRisk = 0.34 if SmokingStatus == "No" & Gender == 2
.                 replace GSSRisk = 3.76 if SmokingStatus == "Ex" & Gender == 2
.                 replace GSSRisk = 4.17 if SmokingStatus == "Yes" & Gender == 2
.                 }
(191 missing values generated)
(85 real changes made)
(60 real changes made)
(28 real changes made)
(6 real changes made)
(12 real changes made)
```

*Categorise by Age and Gender for lung cancer incidence and death SEER rates

*Year for SEER reference in the article

```
.                  forvalues i = 1/20 {
.                 gen AgeGender_`i' = 0
.                 }

.                 foreach var of varlist ID {
.                 replace AgeGender_1 = 1 if Gender == 1 & AgeRegistered >= 40 & AgeRegistered < 45
.                 replace AgeGender_2 = 1 if Gender == 1 & AgeRegistered >= 45 & AgeRegistered < 50
.                 replace AgeGender_3 = 1 if Gender == 1 & AgeRegistered >= 50 & AgeRegistered < 55
.                 replace AgeGender_4 = 1 if Gender == 1 & AgeRegistered >= 55 & AgeRegistered < 60
.                 replace AgeGender_5 = 1 if Gender == 1 & AgeRegistered >= 60 & AgeRegistered < 65
.                 replace AgeGender_6 = 1 if Gender == 1 & AgeRegistered >= 65 & AgeRegistered < 70
.                 replace AgeGender_7 = 1 if Gender == 1 & AgeRegistered >= 70 & AgeRegistered < 75
.                 replace AgeGender_8 = 1 if Gender == 1 & AgeRegistered >= 75 & AgeRegistered < 80
.                 replace AgeGender_9 = 1 if Gender == 1 & AgeRegistered >= 80 & AgeRegistered < 85
.                 replace AgeGender_10 = 1 if Gender == 1 & AgeRegistered >= 85
.                 replace AgeGender_11 = 1 if Gender == 2 & AgeRegistered >= 40 & AgeRegistered < 45
.                 replace AgeGender_12 = 1 if Gender == 2 & AgeRegistered >= 45 & AgeRegistered < 50
.                 replace AgeGender_13 = 1 if Gender == 2 & AgeRegistered >= 50 & AgeRegistered < 55
.                 replace AgeGender_14 = 1 if Gender == 2 & AgeRegistered >= 55 & AgeRegistered < 60
.                 replace AgeGender_15 = 1 if Gender == 2 & AgeRegistered >= 60 & AgeRegistered < 65
.                 replace AgeGender_16 = 1 if Gender == 2 & AgeRegistered >= 65 & AgeRegistered < 70
.                 replace AgeGender_17 = 1 if Gender == 2 & AgeRegistered >= 70 & AgeRegistered < 75
.                 replace AgeGender_18 = 1 if Gender == 2 & AgeRegistered >= 75 & AgeRegistered < 80
.                 replace AgeGender_19 = 1 if Gender == 2 & AgeRegistered >= 80 & AgeRegistered < 85
.                 replace AgeGender_20 = 1 if Gender == 2 & AgeRegistered >= 85
.                 }
(2 real changes made)
(6 real changes made)
(16 real changes made)
(15 real changes made)
(21 real changes made)
(35 real changes made)
(27 real changes made)
(20 real changes made)
(7 real changes made)
(3 real changes made)
(3 real changes made)
(2 real changes made)
(4 real changes made)
(4 real changes made)
(6 real changes made)
(11 real changes made)
(5 real changes made)
(7 real changes made)
(1 real change made)
(3 real changes made)
```

```
. mkmat AgeGender_1 - AgeGender_20, matrix(AgeGender)
```

*Matrix of Incidence Coefficients

```
. matrix SEERLCRate = (10.78/100000\25.49/100000\50.6/100000\116.58/100000\221.18/100000\346.77/100000\ ///
478.1/100000\564.36/100000\532.36/100000\498.44/100000\11.03/100000\23.19/100000\45.51/100000\93.93/100000\ ///
164.9/100000\246.85/100000\318.69/100000\344.67/100000\308.28/100000\266.72/100000)
```

*Matrix of Incidence Coefficients in Next Group

```
. matrix SEERDeathRate = (275.1/100000\400.7/100000\560/100000\786.9/100000\1210.2/100000\1855.1/100000\ ///
2947.4/100000\4836.4/100000\7980.7/100000\15559.4/100000\153.2/100000\218.8/100000\313.4/100000\479.1/100000\ ///
762.9/100000\1197/100000\1968.3/100000\3306.1/100000\5761.2/100000\14016.2/100000)
```

*Coefficients

```
. matrix A = AgeGender * SEERLCRate

. matrix B = AgeGender * SEERDeathRate

. svmat double A, name(SEERLCRate)

. svmat double B, name(SEERDeathRate)
```

*Clean the dataset drop (AgeGender_1 - AgeGender_20)

*Baseline Risk

```
. foreach var of varlist ID {
. gen BaseRisk = SEERLCRate * GSSRisk * (1-AttRisk)
. }
```

*Spitz Model One Year Risk - percentage

```
. foreach var of varlist ID {
. gen SpitzRisk = ((OddsRatio * BaseRisk) / ((OddsRatio *BaseRisk)+ SEERDeathRate))*(1 - exp(-((OddsRatio *BaseRisk)+ SEERDeathRate)))*100
. }
```

*Clean the dataset

```
. drop (OddsRatio - BaseRisk)
```

Pittsburgh Model Stata Code

*Pittsburgh Predictor for 6 Year Risk

*Pre-Amble

```
.  keep ID CaseControl AgeRegistered SmokingStatus CPD Duration
```

*Drop ineligible participants

```
.  foreach var of varlist ID {
. drop if SmokingStatus == "No" | AgeRegistered < 35
. }
(251 observations deleted)
```

*Model Coefficients

*Smoke Duration Coefficient

```
.          foreach var of varlist ID {
.         gen DurationCoefficient = -10 if Duration < 30
.         replace DurationCoefficient = 0 if Duration >= 30 & Duration < 40
.         replace DurationCoefficient = 8 if Duration >= 40 & Duration < 50
.         replace DurationCoefficient = 14 if Duration >= 50
.         }
(600 missing values generated)
(324 real changes made)
(261 real changes made)
(15 real changes made)
```

*Age Coefficient

```
.          foreach var of varlist ID {
.         gen AgeCoefficient = 0
.         replace AgeCoefficient = 4 if (AgeRegistered >= 57 & Duration < 30) | ///
         (AgeRegistered >= 59 & Duration >= 30 & Duration < 40) | ///
         (AgeRegistered >= 61 & Duration >= 40 & Duration < 50) | (AgeRegistered >= 68 & Duration >= 50)

.         }
(167 real changes made)
```

*Smoking Status Coefficient

```
.          foreach var of varlist ID {
.         gen SSCoefficient = 0 if SmokingStatus == "Yes"
.         replace SSCoefficient = -3 if SmokingStatus == "Ex"
.         }
(501 missing values generated)
(501 real changes made)
```

*Smoking Intensity Coefficient

```
.          foreach var of varlist ID {
.         gen SICoefficient = -4 if CPD < 20
.         replace SICoefficient = 0 if CPD >= 20 & CPD < 30
.         replace SICoefficient = 2 if CPD >= 30 & CPD < 40
.         replace SICoefficient = 5 if CPD >= 40
.         }
(365 missing values generated)
(283 real changes made)
(61 real changes made)
(21 real changes made)
```

*Final Pittsburgh Six Year Risk as percentage

```
.  foreach var of varlist ID {
. gen PittsburghRisk = 1/(1+(exp(4.2195 - 0.10*(DurationCoefficient + ///
 AgeCoefficient + SSCoefficient + SICoefficient))))*100
```

```
. }
```

*Clean the Dataset

```
.  drop (DurationCoefficient - SICoefficient)
```