ASSOCIATIONS BETWEEN PERSONAL CANCER HISTORY AND LUNG CANCER RISK

Karl Daniel Everett, BSc

Submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Applied Health Sciences

The Faculty of Applied Health Sciences

Brock University

St. Catharines, ON

ABSTRACT

The study aim was to investigate the relationship between factors related to personal cancer history and lung cancer risk as well as assess their predictive utility. Characteristics of interest included the number, anatomical site(s), and age of onset of previous cancer(s). Data from the Prostate, Lung, Colorectal and Ovarian Screening (PLCO) Cancer Screening Trial (N = 154,901) and National Lung Screening Trial (N = 53,452) were analyzed. Logistic regression models were used to assess the relationships between each variable of interest and 6-year lung cancer risk. Predictive utility was assessed through changes in area-under-the-curve (AUC) after substitution into the PLCOall2014 lung cancer risk prediction model. Previous lung, uterine and oral cancers were strongly and significantly associated with elevated 6-year lung cancer risk after controlling for confounders. None of these refined measures of personal cancer history offered more predictive utility than the simple (yes/no) measure already included in the PLCOall2014 model.

ACKNOWLEDGEMENTS

Half-way up a mountain, exasperation is the easy path to the bottom, determination is the tough trek to the top. Were it not for the support of many valued professors, friends and family, I might not have been able to enjoy the view at the summit. I feel privileged to have had such an experienced, knowledgeable and assured guide in Dr. Martin Tammemagi. If something is worth doing, it's worth doing well and I will always be thankful for the invaluable training you've given me, grounding me in the best possible principles and helping me develop as a researcher.

I am grateful to Dr. Jian Liu for his insights which have helped me grasp the concepts most fundamental to health research and for giving me the opportunity to mature as an educator as well as a student. I would also like to offer many thanks to Dr. Theos Tsakiridis for his participation and contributions to this project. I extend the sincerest gratitude to Dr. Carolynn Pietrangelli for instilling such confidence in me and to Dr. Phil Wilson for his continued reassurance and support when I needed it most. I appreciate the kindness and generosity of both of you more than you know.

Working alongside such fantastic fellow grad students, with whom I have shared as many laughs as frustrations, is what got me through the last few years. I could not have got here without you guys. I thank my wonderful friends, not least Lea and Teisha for always being there for me. I appreciate your compassion and backhanded compliments in equal measure. Finally, I am especially fortunate to have such devoted parents, without whom I could not have even begun this journey let alone finished it. Thanks for your love and support every step of the way.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACRIN - American College of Radiology Imaging Network

AUC - Area under the curve

BCC - Basal cell carcinoma

BMI - Body mass index

CARET - β-Carotene and Retinol Efficacy Trial

COPD - Chronic obstructive pulmonary disease

CT - Computed tomography

HR - Hazard ratio

IARC - International Agency for Research on Cancer

I^{2-} Percentage variation between studies in a meta-analysis attributable to heterogeneity

IMS - Information Management Systems, Inc.

LDCT - Low-dose computed tomography

LLP - Liverpool Lung Project

LR - Logistic regression

LSS - Lung Screening Study

NCI - National Cancer Institute

NLST - National Lung Screening Trial

OR - Odds ratio

PLCO - Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial

RISPC - Radiation-induced second primary cancer

ROC - Receiver operating characteristic

RR - Relative risk

SCC – Squamous cell carcinoma

SEER - Surveillance, Epidemiology, and End Results Program

SMR - Standardized mortality ratio

SPC - Subsequent primary cancer

SPLC - Subsequent primary lung cancer

SIR - Standardized incidence ratio

USPSTF - United States Preventive Services Task Force

VIF - Variance inflation factor

CHAPTER 1: INTRODUCTION

This thesis examines the associations between personal cancer history and lung cancer risk. The goal of this research is to develop a better understanding of which specific aspects of a personal history of cancer are associated with lung cancer risk and to use this information to attempt to refine existing lung cancer risk prediction models. Prediction models are used to identify those at highest risk who are most likely to benefit from screening as well as smoking cessation programs for current smokers.

In the United States, data from 2009-2011 indicates that the cumulative risk of developing cancer of the lung or bronchus over their lifetime is 1 in 13 for men and 1 in 16 for women. The cumulative mortality figures are 1 in 15 for men 1 in 20 for women (American Cancer Society, 2013b). In Canada, estimates based on data from 2010 show that the lifetime risk of developing lung cancer is 1 in 12 for men and 1 in 15 for women. The cumulative mortality risk is 1 in 13 for men and 1 in 17 for women (Canadian Cancer Society's Advisory Committee on Cancer Statistics, 2014).

When considering males and females separately, lung cancer has the second highest incidence rate of all cancers in the United States after prostate cancer and breast cancer, respectively, with estimates of 70.1 cases per 100,000 persons per year in men and 50.2 per 100,000 per persons per year in women, age standardized to the 2000 U.S. standard population (Census P25-1130). In terms of cancer deaths, lung cancer has the highest age-adjusted mortality rate in both men and women; 59.8 per 100,000 and 37.8 per 100,000 persons per year, respectively (U.S. Cancer Statistics Working Group, 2015). Among Canadians, the incidence rate of lung cancer in men has declined by 1.8% per

year since 1998 but increased in women by 1.1% per year between 1998 and 2007. Despite this increase, the annual incidence rate is still higher in males than females (57.6 vs. 47.5 per 100,000 respectively, age-standardized to the 1991 Canadian population (Canadian Cancer Society's Advisory Committee on Cancer Statistics, 2014)). These statistics are summarized in Table 1.1 but it may not be appropriate to draw comparisons from them due to the disagreement between the age-standardized incidence and mortality rates, which are consistently higher in the United States, and the lifetime risks of developing and dying from lung cancer which appear to reflect less favorably on the situation in Canada. It is likely that this anomaly is artifactual as a result of different methodological approaches taken to calculate the estimates.

Table 1.1 Lung Cancer Incidence and Mortality in the United States and Canada

	United States		**Canada**	
	Males	Females	Males	Females
ASIR* (per 100,000 people per year)	70.1	50.2	57.6	47.5
ASMR* (per 100,100 people per year)	59.8	37.8	46.3	35.6
Lifetime Risk of Incidence	1 in 13	1 in 16	1 in 12	1 in 15
Lifetime Risk of Mortality	1 in 15	1 in 20	1 in 13	1 in 17

Abbreviations: ASIR: Age-standardized incidence rate; ASMR: Age-standardized mortality rate.
*Age-standardized to the 2000 U.S. population and 1991 Canadian Standard Population, respectively.

There is sufficient evidence linking tobacco smoking to the following cancer sites: oral cavity, nasal cavity, pharynx, larynx, esophagus, lung, stomach, pancreas, liver, kidney (body and pelvis), ureter, urinary bladder, cervix, ovary (mucinous), colon and rectum, bone marrow (myeloid leukaemia) (Secretan et al., 2009). Smoking is by far the most influential determinant for the development of lung cancer; a greater intensity and longer duration of smoking as well as a shorter quit-time all increase the risk (Tammemagi et al., 2011). The predominance of smoking can eclipse other, more subtle but nevertheless important risk factors. Thus, this research was undertaken to investigate some of these

factors in detail, focusing on personal cancer history. The study participants were taken from two multi-centre screening trials in the United States, the National Lung Screening Trial (NLST) and the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. Data regarding the number, anatomical site and age of onset of previous cancers were collected as part of the screening trials.

1.1 Screening and Lung Cancer-Specific Mortality

Lung cancer is the leading cause of cancer-related mortality worldwide, responsible for more deaths per year than breast, prostate and colorectal cancers combined (American Cancer Society, 2013c). This occurs because the majority of lung cancer cases are diagnosed at locally/regionally advanced or metastatic stages and surgery is less likely to be beneficial or even feasible (American Cancer Society, 2013a). However, treatment is possible if the cancer is detected early enough. 49% of patients with lung cancer diagnosed at stage IA survive for over 5 years compared with only 5% of those diagnosed at stage IIIB (American Cancer Society, 2015)**.** Lung cancer screening is being increasingly adopted in many areas and risk prediction models are an effective tool with which to determine who would most likely benefit from screening. Thus, improving risk prediction is important.

Survival duration, however, is not an ideal measure of whether or not screening saves lives; it is inherently prone to lead-time bias, that is, when screening wrongly appears to improve survival length because detection occurred earlier than it otherwise would have, but the actual course of the disease was unaffected (Reynolds, 2012). It is also prone to length time bias (excessive sampling of less aggressive cases, leading to

overestimation of survival duration among screen-detected cases) and over-diagnosis bias (underestimating mortality due to the inclusion of subclinical disease that would never have become symptomatic were it not screened for) (Welch, Birkmeyer, Schwartz, Black, & Woloshin, 1999). Measuring cause-specific mortality is more appropriate. The NLST found a 20% reduction in lung-cancer specific mortality among those screened with low-dose computed tomography (LDCT) compared to those screened with chest X-rays as well as a 6.7% reduction in all-cause mortality (Aberle et al., 2011).

1.2 Gaps in Current Knowledge

Studies focusing on lung cancer as a secondary or tertiary malignancy are relatively scarce, particularly those where the primary tumour is not of pulmonary origin. Most of the research into previous cancer as a risk factor for lung cancer has been concerned with a previous history of cancer within the family of the individual rather than the person themselves, and most often has focused on previous history of lung cancer. Most studies of subsequent primary cancer (SPC) have used population-based cancer registries to assess many subsequent cancer types in one comprehensive analysis rather than focusing on subsequent primary lung cancer (SPLC) specifically. These broad analyses usually lack adjustment for confounding by smoking-related factors that could be particularly important with respect to associations with SPLC.

Two current prediction models ($PLCO_{M2012}$ & $PLCO_{all2014}$), developed using data from the PLCO trial, take personal history of cancer into account (Tammemagi et al., 2013, 2014). However this predictor is included as a binary covariate in both models (yes vs. no), that is, having ever had or not had any previous cancer at the baseline of the

study (Tammemagi et al., 2013, 2014) and captures no information about the number of previous cancers, cancer type or age of onset. Although a personal history of cancer was a significant predictor of lung cancer diagnosis, more specific information may lead to superior prediction.

1.3 Response to Gaps

In order to address the research questions of this project, data from the NLST and PLCO trials were analyzed. The datasets from the NLST and PLCO trials contain comprehensive, detailed information with respect to many participant characteristics, including data regarding personal cancer history. Details regarding the number of previous cancers each participant recorded, as well as the anatomical site, and age at diagnosis were collected as part of the trials.

Logistic regression models (Hosmer, Lemeshow, & Sturdivant, 2013) were used to obtain risk estimates for these characteristics as predictors of future lung cancer in the individual. Effect sizes were estimated using odds ratios (ORs). Data on smoking status, intensity, duration and quit-time for participants facilitated controlling for smoking exposures. The same modeling technique was used to build prediction models that incorporated new factors in addition to predictors established in previous models. The utility of the new variables was determined based on the predictive performance of models that included them; discrimination (area under the receiver-operating-characteristic curve [AUC]) and calibration (Spiegelhalter's statistic, a component of the Brier score) were assessed.

1.4 Study Aims and Research Questions

In general terms this study aimed to develop an understanding of the relationship between personal cancer history and lung cancer risk. More specifically, the goal was to evaluate the independent risk of lung cancer associated with having been previously diagnosed with commonly occurring cancers, number of previous cancers and age at diagnosis, and to use this information to refine risk prediction modeling for lung cancer, where possible. In order to achieve these aims, the following research questions will be addressed:

1. Type of cancer

Are any specific types of previous cancers strong independent predictors of subsequent lung cancer?

2. Number of cancers

Is the number of previous cancers an important predictor of future lung cancer?

3. Age at diagnosis

Is being diagnosed with cancer before age 60 years associated with future lung cancer risk more strongly than diagnosis at 60 years or older?

4. Predictive performance

Do any of these associations lead to improved lung cancer risk prediction models?

1.5 Conclusion

There are a number of reasons why this research is important. Firstly, lung cancer represents a major global health concern and the NLST has demonstrated the ability of screening with computed tomography to significantly reduce lung cancer mortality, but continual improvements in targeting high-risk individuals is crucial since screening individuals at low risk has been shown to be ineffective, inefficient, and potentially harmful to the patient (Kovalchik et al., 2013). Secondly, the focus on a history of

previous cancer in the current study is justified; Bae et al. (Bae et al., 2011) found that 93% of lung cancers as a SPC and 61% of lung cancers as recurrences were detected by computed tomography and treatment can be both feasible and effective for those with lung cancer as a second cancer, if the cancer is detected early enough, with a 5-year survival rate of 77% (Bae et al., 2011).

This research builds on lung cancer prediction models established using data from the NLST and PLCO trials that have demonstrated high discrimination and calibration but have the potential to be further refined and improved. In this study, detailed prospective data, lengthy follow-up and a large sample size consisting of both smokers and non-smokers facilitated the investigation of multiple risk factors that are relatively uncommon while maintaining a reasonable degree of statistical power. This project was able to explore the effect of a personal history of cancer with respect to lung cancer risk in greater detail than in previous studies while being able to adjust for other known risk factors in a way that was superior to past studies due to quality and detail of information. The findings from these analyses may facilitate improvements to currently existing lung cancer risk prediction models as well as provide valuable insights into lung cancer carcinogenesis.

CHAPTER 2: REVIEW OF THE LITERATURE

2.1 Overview

This chapter provides a background and context for the current study as well as offering the rationale for the research; the main focus is placed on the factors most pertinent to this study. The association between previous cancer and subsequent primary cancer (SPC) in general is commented on but the main focus of the review is on associations with subsequent primary lung cancer (SPLC). The standardized incidence ratios (SIRs) from the studies discussed in this section are summarized along with their corresponding sample sizes and adjustment criteria in Table 2.1. An SIR expresses the incidence within a defined study population relative to what would be expected if the incidence seen among a comparison group (most often the general population) were applied to the population being studied. If the number of observed cases equals the number expected, the SIR is 1. An SIR greater than 1 indicates a greater than expected number of cases and conversely, fewer than expected would give an SIR less than 1 (Aschengrau & Seage, 2008).

An overview of the current state of screening for lung cancer is provided and finally, the process of risk prediction model development is outlined as well as an overview of notable risk prediction models for lung cancer.

Table 2.1 - Summary of standardized incidence ratios (SIRs) from studies investigating associations between previous cancer and SPLC by initial cancer type

Initial cancer type	Study	N	SIR (95% CI)	Adjustment criteria
Breast	(Hemminki et al., 2005)	3,409 males	1.26 (0.96 – 1.61)	Age, calendar year, registry
	(Andersson, Jensen, Engholm, & Henrik Storm, 2008)	53,418 females	1.0 (0.9 – 1.1)	Sex, age, calendar period
	(Utada, Ohno, Hori, & Soda, 2014)	174,477	1.41 (1.13 – 1.73)	Sex, age group, calendar year
Digestive				
Colorectal	(Hemminki, Li, & Dong, 2001)	68, 104	1.90 (0.36 – 4.66)	Age, sex
	(Phipps, Chan, & Ogino, 2013)	170,159	1.14 (1.10 – 1.18)	Sex, age, calendar year, race
	(Jégu et al., 2014)	289,967	1.16	Age, sex, year of index cancer
Esophageal	(Chuang, Hashibe, et al., 2008)	52,589	1.55 (1.28 – 1.87)* 0.91 (0.60 – 1.32)†	Age, sex, calendar year
	(Zhu et al., 2012)	24,557	3.19 (2.12 – 4.61)	Sex, age, race and calendar year
	(Jégu et al., 2014)	160, 807 males	4.25	Age, sex, year of index cancer
	(Utada et al., 2014)	174,477	1.39 (1.02 – 1.85)	Sex, age group, calendar year
Endocrine				
Thyroid	(Rubino et al., 2003)	6,841	1.0 (0.6 – 1.4)	Sex, calendar year, age
	(Canchola, Horn-Ross, & Purdie, 2006)	10,932 females	1.0 (0.6 – 1.4)	Sex, age and calendar period
	(Verkooijen, Smit, Romijn, & Stokkel, 2006)	282	0.46 (0.08 – 1.15)	Age, sex
	(A. P. Brown et al., 2008)	30,278	0.85 (0.75 – 0.96)	Age, sex, calendar year
	(C.-H. Lu et al., 2013)	19,068	1.57 (1.25 – 1.95)	Age, sex, calendar year
	(Utada et al., 2014)	174,477	1.83 (1.31 – 2.49)	Sex, age group, calendar year
Genitourinary				
Kidney	(Czene & Hemminki, 2002)	9,344	1.36 (1.12 – 1.47)	Sex, calendar year, age
	(Utada et al., 2014)	174,477	1.29 (1.13 – 1.73)	Sex, age group, calendar year
Prostate	(Thellenberg, Malmer, Tavelin, & Grönberg, 2003)	135,713	1.05 (0.99 – 1.12)	Calendar period, age
	(Hinnen et al., 2011)	1,888	0.48 (0.32 – 0.71)	Age, calendar year
	(Utada et al., 2014)	174,477	0.85 (0.71 – 0.99)	Age group, calendar year
Testicular	(Travis et al., 2005)	40,576	1.5 (1.2 – 1.7)	Area, age, calendar year
Gynecologic				
Cervical	(Hemminki, Dong, & Vaittinen, 2000)	117, 830	2.81 (2.43 – 3.21)§ 2.17 (2.00 – 2.34)‖	Age, calendar period
	(Jégu et al., 2014)	289,967	4.44	Age, year of index cancer
Uterine	(Koivisto-Korander et al., 2012)	8,606	1.73 (1.04 – 2.70)	Calendar year, age at diagnosis
	(Jégu et al., 2014)	289,967	1.62	Age, year of index cancer
	(Utada et al., 2014)	174,477	1.91 (1.50 – 2.38)	Age group, calendar year
Head and Neck	(Chuang, Scelo, et al., 2008)	99,257	3.30 (3.19 – 3.41)	Sex, calendar year, registry
	(Jégu et al., 2014)	160, 807 males	8.71	Age, sex, year of index cancer
		129,160 females	18.81	Age, sex, year of index cancer
Laryngeal	(Morris, Sikora, Hayes, Patel, & Ganly, 2011)	75,087	4.07 (3.92 – 4.22)	Age, sex, race, calendar period
	(Chen et al., 2011)	63,720	2.60 (2.04 – 3.25)	Age, calendar year
	(Utada et al., 2014)	174,477	1.41 (1.13 – 1.73)	Sex, age group, calendar year
Oropharyngeal	(Morris et al., 2011)	75,087	4.86 (4.54 – 5.20)	Age, sex, race, calendar period
	(Chen et al., 2011)	63,720	1.56 (1.34 – 1.80)	Age, calendar year
Nasopharyngeal	(Scélo et al., 2007)	8,947	1.35 (0.89 – 1.96)	Age, sex, calendar period
	(Chen et al., 2011)	63,720	1.12 (0.90 – 1.38)	Age, calendar year
Hematologic				
Hodgkin lymphoma	(Jégu et al., 2014)	160, 807 males	4.02	Age, sex, year of index cancer
		129,160 females	4.14	Age, sex, year of index cancer
Non-Hodgkin lymphoma	(Mudie et al., 2006)	2,456	1.6 (1.1 – 2.3)	Age, sex, calendar year
	(Morton et al., 2010)	43,145	1.19 (1.09 – 1.30)	Age, sex, ethnicity, calendar year
	(Jégu et al., 2014)	289,967	1.62	Age, sex, year of index cancer
Skin				
Non-melanoma	(Nugent, Demers, Wiseman, Mihalcioiu, & Kliewer, 2005)	43,275	1.09 (1.00 – 1.19)¶	Age, sex
	(Cantwell et al., 2009)	14,442 6,401	1.15 (0.97 – 1.33)¶ 1.05 (0.81 – 1.30)*	Age

Abbreviations: Adc: Adenocarcinoma; BCC: Basal cell carcinoma; CI: confidence interval; N: Number; SPLC (subsequent primary lung cancer); SSC: Squamous cell carcinoma.

* SCC; † Adc; § *in situ*; ‖ Invasive; ¶ BCC.

2.2 Major Risk Factors for Lung Cancer

Numerous factors that are associated with the development of lung cancer have been identified. Cigarette smoking (both intensity and duration) is by far the most important risk factor and exposure to radon gas is believed to be the next leading cause of lung cancer in North America and Europe (American Cancer Society, 2013a). Around 85% of lung cancer cases are attributable to cigarette smoking (Canadian Cancer Society's Advisory Committee on Cancer Statistics, 2014).

Smoking cessation can decrease the risk substantially but not to that of never-smokers (Peto et al., 2000). Other risk factors include secondhand smoke, asbestos exposure, air pollution as well as having a history of lung cancer in the family (American Cancer Society, 2013c). Age, race/ethnicity, education level, body-mass index (BMI), and previous diagnosis of chronic obstructive pulmonary disease (COPD) have all be found to be associated with lung cancer risk among smokers (Tammemagi et al., 2013). A summary of several other key risk factors for lung cancer that might serve as useful predictors is presented in Supplemental Table S1 of reference Tammemagi & Lam (2013) (Tammemagi & Lam, 2013).

2.2.1 Previous cancer

It has been established that developing cancer can be associated with an increased risk of SPC. For all cancers combined, both male (SIR = 1.22; 95% CI: 1.20 – 1.24) and female (SIR = 1.36; 95% CI: 1.33 – 1.39) cancer survivors have been found to have a significant excess risk of developing a second cancer relative to the general population (Youlden & Baade, 2011). This association is particularly apparent among smokers.

Smoking is an independent risk factor for several SPC even after controlling for confounders, including a smoking-related first cancer (Tabuchi et al., 2013). Tabuchi and colleagues estimated that smokers had a 59% increased risk SPC in general and 102% increased risk of smoking-related SPC compared to never-smokers (Tabuchi et al., 2013). Compared with never-smokers, current smokers of ≥ 20 cigarettes per day have been found to be associated with increased second smoking-associated cancer risk among survivors of stage I lung cancer (HR = 3.26; 95% CI: 0.92 to 11.6), bladder (HR = 3.67; 95% CI: 2.25 to 5.99), head/neck (HR = 4.45; 95% CI: 2.56 to 7.73), and kidney (HR = 5.33; 95% CI: 2.55 to 11.1) cancers (Shiels et al., 2014).

For SPLC specifically, ever smokers who have survived cancer have a significantly elevated risk of compared to never-smokers regardless of the site of the first cancer (Tabuchi et al., 2013). The increased risk is likely attributable to factors such as similar disease etiologies, genetic susceptibility and the effects of treatment procedures undergone for the initial cancer, particularly if exposed to radiation (Youlden & Baade, 2011). In terms of the prognostic significance, lung cancer patients who have had a previous primary cancer have comparable prognoses and respond similarly to treatment as those whose lung cancer is their first malignancy (Koppe et al., 2001). Quitting smoking immediately after cancer diagnosis can significantly reduce the risk of developing SPC, after adjusting for confounding factors (Tabuchi et al., 2013).

Finally, it should be noted that second primary malignancies can either be classed as synchronous (cancers detected or treated at the same time as the original (index) cancer) or metachronous (cancers detected sometime after the initial primary cancer that are physically distinct and separate) (Martini et al., 1995). Naturally, this is of most

concern when discussing SPLC among lung cancer survivors but the following discussion will focus on previous cancers at sites other than the lung.

2.2.1.1 Previous Cancer Sites associated with SPLC

Breast

The risk of developing a SPC following breast cancer is particularly important since it is the most commonly diagnosed cancer among women in Canada, most of whom (88%) survive their diagnosis (Canadian Cancer Society's Advisory Committee on Cancer Statistics, 2014), consequently increasing the opportunity of future cancer. Lung cancers account for approximately 5% of SPC among breast cancer survivors (Mariotto, Rowland, Ries, Scoppa, & Feuer, 2007). No evidence of an increased SPLC risk, relative to the general population, has been observed for male breast cancer survivors (Hemminki et al., 2005).

Several population-based studies have evaluated the risk of SPLC after breast cancer in women. Utada and colleagues reported a 41% increased risk of lung cancer following previous breast cancer (SIR 95% CI: 1.13- 1.73) among Japanese survivors, after controlling for age group and calendar year (Utada et al., 2014). A summary of relative risks (RRs) found in population cancer registry studies of SPC among breast cancer survivors is presented by Evans et al.; RRs for SPLC relative to the general population ranged from 1.4 in Denmark to 1.7 in Connecticut, and the USA (in those diagnosed with breast cancer before age 45) (Evans et al., 2001). The increased risk of SPLC associated with prior breast cancer were all statistically significant at an alpha level of 0.05 (Ewertz & Mouridsen, 1985; Harvey & Brinton, 1985; Teppo, Pukkala, & Saxén,

1985; Volk & Pompe-Kirn, 1997). Among a cohort of 145,677 women diagnosed with breast cancer between 1961 and 1996, Evans and colleagues reported an increased SPLC risk in those initially diagnosed before age 50 (SIR = 1.49; 95% CI: 1.26 – 1.78) but a protective effect was seen among those diagnosed between 50-84 (SIR = 0.68; 95% CI: 0.62 – 0.74) (Evans et al., 2001). These estimates were adjusted for age and calendar year, but not for smoking. Raymond and Hogue (2006) also present SIRs for SPLC stratified by age at diagnosis in their analysis of 335,191 females with breast cancer and found that SPLC risk was positively associated with younger age at diagnosis (SIRs of 6.7, 1.31 and 1.24 for age groups of 30-39, 40-49 and 50-59, respectively) (Raymond & Hogue, 2006).

In a further analysis of 525,527 women using data taken from 13 population-based cancer registries in Europe, Canada, Australia and Singapore, Mellemkær and colleagues reported a statistically significantly increased risk of SPLC, after adjustment for 5-year age group and calendar period (Mellemkjær et al., 2006). SPLC risk was highest for women with premenopausal breast cancer diagnosis (SIR = 2.12; 95% CI: 1.92 – 2.33) compared to perimenopausal (SIR = 1.53; 95% CI: 1.42 – 1.64) or postmenopausal (SIR = 1.05; 95% CI: 1.00 – 1.10) (Mellemkjær et al., 2006). A lack of an association with SPLC was reported among a Danish cohort of 53,418 breast cancer patients (SIR = 1.0; 95% CI: 0.9 – 1.1) after adjustment for sex, 5-year age group and calendar period. However, SIRs varied by follow-up interval since breast cancer diagnosis, from 0.8 (95% CI: 0.7 – 0.9) for 1-9 years, to 2.9 (1.6 – 4.1) for 20+ years; the trend was statistically significant (Andersson et al., 2008). SPLC risk was highest for

individuals diagnosed before age 50 years (SIR = 1.5; 95% CI: 1.2 – 1.9) (Andersson et al., 2008).

Women who received radiation treatment for breast cancer have been shown to be at a 1.5- to 3-fold increased risk for subsequently developing lung cancer compared with women who did not receive radiation (Neugut et al., 1993; Zablotska & Neugut, 2003). The finding that lung cancers after breast cancer therapy are most frequently found in the ipsilateral lung supports the contributing role of radiation to the risk (Zablotska & Neugut, 2003). A systematic review and meta-analysis of second non-breast cancer in those who previously underwent radiotherapy for breast cancer reported a pooled RR for SPLC of 1.23 (95% CI: 1.07 – 1.43), adjusted for smoking status and adjuvant therapy, compared to those who did not receive radiotherapy (Grantzau & Overgaard, 2014).

There is growing evidence from recent studies that women who smoke have an increased risk of breast cancer, particularly if they start smoking before they have their first child. Both a higher intensity and a longer duration are associated with elevations in risk (Dossus et al., 2014; Gaudet et al., 2013). Controlling for smoking status alone could be an inadequate way of adjusting for the effects of smoking.

Digestive

Colorectal

An elevated risk of SPLC has been found among those with previous colorectal cancer relative to the general population (SIR = 1.14; 95% CI: 1.10 – 1.18) after adjustment for age (5 year grouping), sex, and race (black, white, other) (Phipps et al., 2013). This finding was replicated by Jégu and colleagues (SIR = 1.16) after adjustment

for age, year of first cancer diagnosis (Jégu et al., 2014). However, Hemminki and colleagues (2001) found no statistically significant relative increase in risk SPLC among 68,104 cases of colorectal cancer, relative to the general population, regardless of follow-up length (< 1 year, 1 – 10 years or > 10 years) (Hemminki, Li, & Dong, 2001).

An increased risk of SPLC after radiotherapy treatment for rectal cancer was seen among participants of the Swedish Rectal Cancer Trial, compared to those treated with surgery alone, but it was not statistically significant (RR = 4.36; 95% CI: 0.51 – 37.33) (Birgisson, Påhlman, Gunnarsson, & Glimelius, 2005).

Esophageal

Among survivors of esophageal cancer, subsequent lung cancer risk was found to be significantly elevated (SIR = 3.19; 95% CI: 2.12–4.61, adjusted for sex, race, age and calendar year) compared with the general population (Zhu et al., 2012). A similar increase in SPLC risk among male esophageal cancer survivors, relative to the general population, was published by Jégu and colleagues (SIR = 4.25) after adjustment for age, sex, year of first cancer diagnosis (Jégu et al., 2014). SIRs for SPLC by histological type of esophageal cancer were reported in a pooled analysis of 13 cancer registries; SPLC risk was elevated for squamous cell carcinoma (SCC) of the esophagus (SIR = 1.55; 95% CI: 1.28 – 1.87), but not for adenocarcinoma (AdC) (SIR = 0.91; 95% CI: 0.60 – 1.32) and only persisted for SPLC diagnosed at least one year after esophageal cancer diagnosis (Chuang, Hashibe, et al., 2008). Utada and colleagues reported a 39% increased risk of lung cancer following previous esophageal cancer (SIR 95% CI: 1.02- 1.85) among Japanese survivors, after controlling for sex, age group and calendar year (Utada et al., 2014).

Gastric

Most studies investigating SPCs have not reported the risk of SPLC specifically among survivors of gastric cancer, but the frequency of SPC types following gastric cancer was reported for 2,250 patients in Japan by Ikeda and colleagues (2002). The lung was the most common site of SPC (46.9%) but no comparison relative to a control group was presented (Ikeda et al., 2002).

Endocrine

Thyroid

The risk of SPC among thyroid cancer survivors was evaluated by pooling data from three major Swedish, French and Italian cohorts; among 6,841 patients, 32 developed SPLC but no statistically significant increase was seen compared to those with no previous thyroid cancer after adjustment for sex, age and calendar year (SIR = 1.0; 95% CI: 0.6 – 1.4) (Rubino et al., 2003). A lack of association with SPLC was reported for both male (n = 146) or female (n = 729) thyroid cancer patients in France after adjustment for sex, age and 5-year calendar period (Berthe et al., 2004). An analysis specific to women with papillary thyroid cancer (N=10,932) also failed to find any association with SPLC, relative to the general population (SIR = 1.0; 95% CI: 0.6 – 1.4) (Canchola et al., 2006).

A protective effect for SPLC relative to the general population was reported by Verkooijen and colleagues (2006) but their sample size was relatively small (n=282) (Verkooijen et al., 2006). However, this was also observed in a much larger study (N=30, 278) of SPC up to 30 years after treatment of thyroid cancer (SIR = 0.85; 95% CI: 0.75 –

0.96) relative to the general population, controlling for age, sex and calendar year (A. P. Brown et al., 2008). Only one study of SPLC among 19,068 Asian survivors of thyroid cancer reported an elevated risk (SIR = 1.57; 95% CI: 1.25 – 1.95, adjusted for age, sex and calendar year) (C.-H. Lu et al., 2013).

Sawka and colleagues (2009) assessed the effect of treatment with radioactive iodine with respect to SPC risk after thyroid cancer. The magnitude of the association for SPLC relative to those not treated with radioactive iodine was suggestive of a clinically relevant increase but the association was not statistically significant (RR = 1.50; 95% CI: 0.86 – 2.60) (Sawka et al., 2009).

Genitourinary

Kidney

Studies of SPC after cancer of the kidney do not typically include estimates of SPLC but an increased risk of SPLC was seen among cases of kidney cancer reported to the Swedish Family Cancer Database, relative to general population after controlling for sex, calendar year and age (SIR = 1.36; 95% CI: 1.13 – 1.61) (Czene & Hemminki, 2002). Utada and colleagues reported a 29% increased risk of lung cancer following previous cancer of the kidney, bladder or urinary tract (SIR 95% CI: 1.12 – 1.47) among Japanese survivors, after controlling for sex, age group and calendar year (Utada et al., 2014).

Prostate

The risk of SPC is particularly relevant for prostate cancer survivors as men are being diagnosed at earlier stages and younger ages than in the past (Brawley, 2012). Thus,

patients may receive treatment earlier and are surviving longer, increasing the period in which subsequent cancer can develop.

In a large population-based study of all cases of prostate cancer reported to the Swedish Cancer Registry between 1958-1996 (N=135,713), Thellenberg and colleagues reported no evidence of an association with SPLC relative to general population overall (SIR = 1.05; 95% CI: 0.99 – 1.12), but an elevated risk was seen among those diagnosed with SPLC up to five months after their prostate cancer diagnosis (SIR = 3.49; 95% CI: 3.14 – 3.86), most likely due to heightened surveillance during that time period (Thellenberg et al., 2003). In contrast, evidence of a statistically significant protective effect on SPLC risk among a cohort of 1,888 prostate cancer patients, relative to the general population (SIR = 0.48; 95% CI: 0.32 – 0.71) after controlling for age and calendar year (Hinnen et al., 2011). This result is supported by the study of Japanese cancer survivors, in which Utada and colleagues reported an SIR for SPLC of 0.85 (95% CI: 0.72 – 0.99) among prostate cancer survivors (Utada et al., 2014). It has been suggested that the 'protective' effects reported in several other studies (Kleinerman, Liebermann, & Li, 1985; Liskow et al., 1987; Pawlish, Schottenfeld, Severson, & Montie, 1997) could be explained by underreporting in survivors who live to an elderly age (Pickles & Phillips, 2002) as well as a lower than average frequency of smokers among prostate cancer survivors compared to the general population, since curative treatments for prostate cancer could be contraindicated by smoking-related comorbidities (Hinnen et al., 2011).

A study by Brenner and colleagues found that patients who received radiation treatment for prostate cancer had a significantly increased risk for subsequent lung cancer

compared with those treated who received surgery (RR = 1.11; 95% CI: 1.01 – 1.21) after adjustment for age at diagnosis (Brenner, Curtis, Hall, & Ron, 2000), a finding later replicated by Moon and colleagues (OR = 1.25; 95% CI: 1.13 – 1.37) (Moon, Stukenborg, Keim, & Theodorescu, 2006).

Testicular

Among a population of 40,576 survivors of testicular cancer, 256 developed SPLC. The SIR compared with the general population was 1.5 (95% CI: 1.2 – 1.7) after adjustment for registration area, 5-year age group and calendar year (Travis et al., 2005). However, among cases of testicular cancer in the Swedish Family Cancer Database (N= 4,650) risk of SPLC was not associated with previous testicular cancer of any histological type, regardless of follow-up length, after adjustment age, period, residence and occupation (Dong, Lönnstedt, & Hemminki, 2001).

The extent to which any association with SPLC may be confounded by smoking is uncertain, but several studies have failed to demonstrate an association between smoking and testicular cancer risk (L. M. Brown, Pottern, & Hoover, 1987; Henderson, Benton, Jing, Yu, & Pike, 1979; Oliver, 1994).

Gynecologic

Cervical

A study by Arnold and colleagues found that the majority of SPCs diagnosed following cervical cancer were smoking-related (SIR = 1.8; 95% CI: 1.4 – 2.2); among the 12,048 patients in their study, 676 developed a SPC during the study period, of which 147 (21.7%) were lung cancers (SIR = 4.7; 95% CI: 4.0 – 5.5, adjusted for 5-year age group

and calendar year). The median follow-up for smoking-related cancers was 5 years (Arnold et al., 2014). In the population study by Jégu and colleagues (2014), the risk of SPLC among females with cervical cancer was highly elevated (SIR = 4.44) relative to the general population, after adjustment for age and year of first cancer diagnosis.

A large population-based analysis of the Swedish Family-Cancer Database assessed SPC after *in situ* (N=117, 830) and invasive cervical cancers (N=17,556). SIRs for SPLC, relative to the general population, were elevated for both types but was highest for invasive (2.81; 95% CI: 2.43 – 3.21) versus in situ (2.17; 95% CI: 2.00 – 2.34). Estimates were adjusted for age and calendar period (Hemminki, Dong, & Vaittinen, 2000).

Cervical cancer survivors treated with radiotherapy were at higher risk for a second tumour compared to those who were not, especially at smoking-related sites (incidence rate ratio (IRR) = 1.6 (95% CI:1.2 – 2.3)) (Arnold et al., 2014). An increased lung cancer risk was also observed among a cohort of cervical cancer survivors not treated with radiation suggests that common etiological factors other than radiation exposure such as tobacco use are likely to be attributable (Chaturvedi et al., 2007).

Uterine

As part of an international study of SPC using data from 13 cancer registries, an increased risk of SPLC (SIR = 1.73; 95% CI: 1.04 – 2.70, adjusted for calendar period and age at diagnosis) was seen among 8,606 cases of uterine sarcomas, compared to the general population (Koivisto-Korander et al., 2012).

Jégu and colleagues reported an elevated risk of SPLC among females with uterine cancer (SIR = 1.62) relative to the general population, after adjustment for age and year of first cancer diagnosis (Jégu et al., 2014). An even greater, 91% increased risk for SPLC (SIR 95% CI: 1.50 – 2.38) was reported among Japanese females, after controlling for age group and calendar year (Utada et al., 2014).

Head and Neck

Patients with head and neck SCC are at highly elevated risk of SPC most commonly within the head and neck, lung, and esophagus (León et al., 1999; Sturgis & Miller, 1995; Yamamoto, Shibuya, Yoshimura, & Miura, 2002). Carcinoma of the lung has also been reported as the most common SPC among laryngeal cancer survivors (Licciardello, Spitz, & Hong, 1989; Yamamoto et al., 2002).

SIRs of 4.86 (95% CI: 4.54 – 5.20) and 4.07 (3.92–4.22) for SPLC have been reported for survivors of oropharyngeal and laryngeal cancers respectively, after adjustment for age, sex, race and calendar period (Morris et al., 2011). Jégu and colleagues (2014) also reported the risk of SPLC, relative to the general population, in those who were diagnosed with head and neck cancers; the SIRs were 8.71 for males and 18.81 for females after adjustment for age, sex, year of first cancer diagnosis (Jégu et al., 2014). In their study of 63,720 Taiwanese cancer survivors, Chen and colleagues (2011) reported an increased SPLC risk following laryngeal (SIR = 2.05; 95% CI: 1.69 – 2.45) or oropharyngeal cancer (SIR = 1.56; 95% CI: 1.34 – 1.80) but not nasopharyngeal carcinoma (SIR = 1.12; 95% CI: 0.90 – 1.38) (Chen et al., 2011). The majority of nasopharyngeal carcinoma cases in Asia are strongly associated with prior Epstein-Barr infections (Abdel-Hamid, Chen, Constantine, Massoud, & Raab-traub, 1992) so the

location may have influenced the result reported by Chen et al.. However, in an international multicenter study by Scélo and colleagues (2007), no statistically significant increase in SPLC risk was seen for survivors of nasopharyngeal cancer from either Singapore (SIR = 0.65; 95% CI: 0.13 – 1.90) or Australia, Canada, Europe (combined) (SIR = 1.18; 95% CI: 0.43 -2.56) (Scélo et al., 2007).

In their pooled analysis of 13 cancer registries (N=99,257) by Chuang and colleagues (2008), the overall SIR for SPLC among those with head and neck cancer was 3.30 (95% CI: 3.19 – 3.41) after adjustment for sex, year and registry, relative to the general population (Chuang, Scelo, et al., 2008). The elevation in risk was especially pronounced if the person's head and neck cancer was diagnosed before age 56 (SIR = 4.46; 95% CI: 4.17 – 4.77). The SIR declined successively at each older age group ($p <$ .0001 for each age group compared to the youngest category using Poisson regression) (Chuang, Scelo, et al., 2008).

Chen and colleagues also found that SPLC risk was highest in patients whose initial cancer was diagnosed before age 50 years among survivors of oropharyngeal cancer. SIRs for SPLC were 5.41, 1.91 and 0.98 for those whose initial age of onset was < 50 years, 50-59 and 60-69, respectively. Corresponding estimates for survivors of laryngeal cancer were 4.88, 3.93 and 1.90 (Chen et al., 2011).

Hematologic

Hodgkin Lymphoma (HL)

Compared to the general population, survivors of HL have a have a 2 to 3-fold risk of developing a second cancer, including lung cancer. This elevated risk is likely

attributable to radiation treatments in the chest region as well as chemotherapy with alkylating agents (American Cancer Society, 2012).

Jégu and colleagues (2014) reported the risk of SPC, relative to the general population using data from ten French population-based cancer registries (N = 289,967). In those who were diagnosed with Hodgkin lymphoma, the SIR for SPC of the lung, bronchus and trachea was 4.02 for males and 4.14 for females after adjustment for age, sex, year of first cancer diagnosis (Jégu et al., 2014).

A meta-analysis of 21 studies investigating SPLC risk among HL survivors calculated a pooled RR for SPLC of 4.62 (95% CI: 3.18 – 6.70, I^2 = 98%). This estimate was positively associated with sample size, sex distribution, institutional vs. population-based datasets, and the use of radiotherapy or combined modality therapy. No association with risk was seen for age at diagnosis (Ibrahim et al., 2013). I^2 expresses the degree of inconsistency across studies used in a meta-analysis – i.e. how much of the variation between them is attributable to heterogeneity as opposed to chance (Higgins, Thompson, Deeks, & Altman, 2003). In this case, 98% indicates very high disagreement between the studies, thus pooling their results may not have been appropriate. The heterogeneity could be due to diversity between participants or methodological factors such as varying degrees of measurement error.

Travis et al. conducted a case–control study of 222 cases of lung cancer and 444 matched controls in individuals previously treated for HL; lung cancer risk increased with radiation dose ($p < .001$) for doses of 30 Gy or higher, relative to those who received <5 Gy of radiation (Travis et al., 2002). For solid tumours, typical doses are 60 – 80 Gy for curative radiotherapy, and 45-60 for adjuvant therapy, delivered in 1.8-2 Gy fractions

(Chung, 2008; Nieder & Baumann, 2011; Sonett et al., 2004; UK National Collaborating Centre for Cancer, 2009).

A case–control study by Travis and colleagues (2002) in which the reference group comprised HL patients who had minimal radiation exposure, those treated with only alkylating agent chemotherapy experienced a fourfold increased lung cancer risk. This increased to sevenfold in those treated with 5 Gy or more of radiation (Travis et al., 2002). Relative risks of 16.8 and 20.2, respectively, were seen for in those patients who also smoked at least one pack of cigarettes per day. For cigarette smokers (at least one pack per day) who had also received both alkylating chemotherapy *and* 5 Gy or more of radiation to the area in question, the relative risk for subsequent lung cancer was 49.1 (Travis et al., 2002).

Non-Hodgkin Lymphoma (NHL)

A British cohort study of 2,456 patients with NHL found a significantly elevated risk of SPLC (SIR = 1.6; 95% CI: 1.1 – 2.3, adjusted for age, sex and calendar year) compared to the general population (Mudie et al., 2006). Morton and colleagues evaluated the risk of second cancers among 43,145 survivors of NHL using SEER registries and reported statistically significant increased risk of lung cancer (SIR = 1.19, adjusted for 5-year age group, sex, ethnicity and calendar year) compared to the general population (Morton et al., 2010).

In their analysis of ten French population-based cancer registries (all patients diagnosed with cancer between 1989 and 2004), Jégu and colleagues also reported an elevated risk of SPLC relative to the general population (SIR = 1.62), among male

patients with NHL after adjustment for age, sex, year of first cancer diagnosis (Jégu et al., 2014).

Skin

Among 21 studies of SPC risk following non-melanoma skin cancer, pooled relative risks for SPLC specifically were 1.23 (95% CI: 1.13 – 1.33), 1.13 (95% CI: 1.01 – 1.27) and 1.34 (95% CI: 1.22 – 1.47) in survivors of non-melanoma skin cancer of any kind, basal cell carcinomas (BCC) and SCC, respectively (Wheless, Black, & Alberg, 2010). However, a 2014 meta-analysis of SPC among non-melanoma skin cancer survivors, comprising over 350,000 patients from population-based cancer registries across several continents, found no evidence to suggest an elevated SPLC risk after pooling study-specific estimates (pooled RR = 0.86; 95% CI: 0.63 – 1.18, I^2 *71%*) (Caini et al., 2014).

In two US cohorts, the Health Professionals Follow-up Study (HPFS) and the Nurses' Health Study (NHS) (N = 46,237), Song and colleagues (2013) found an increased SPLC risk following non-melanoma skin cancer in women (RR = 1.32; 95% CI: 1.14 – 1.52) but not in men (RR = 0.99; 95% CI: 0.79 – 1.24); estimates were adjusted for age, BMI, physical activity, smoking status, smoking intensity, multi-vitamin use, UV-index in place of residences, and menopausal status, hormone replacement therapy use in women (Song et al., 2013). A study of patients with basal cell carcinoma of the skin in Northern Ireland (N=14,442) found no evidence of elevated SPLC risk relative to the general population in either women or men, after controlling for age (Cantwell et al., 2009). A small increase in SPLC risk following basal cell carcnioma (BCC) was observed by Nugent and colleagues in their analysis of all 43,275 non-

melanoma skin cancer cases reported between 1956 and 2000 in Manitoba (SIR = 1.09; 95% CI: 1.00 – 1.19) but not for SCC (Nugent et al., 2005).

2.2.1.2 Cancer Treatment

A well-established potential cause of cancer is exposure to radiation, including the radiation commonly used to treat cancer itself (American Cancer Society, 2012). The risk of developing a radiation-induced second primary cancer (RISPC) is dependent upon factors such as the dose of radiation received, the anatomical site treated and the age of the patient at the time of treatment (American Cancer Society, 2012). Increased risks have been associated with several tumour sites, including the lungs (Travis, 2006), which usually occur 10 or more years after treatment (Boice Jr, Land, & Preston, 1996). A ranked summary of the carcinogenic effects of radiation exposure on various tissues is presented by Travis (Travis, 2006). Solid tumours induced by radiotherapy include lung cancers. Some chemotherapy treatments have also been linked to the development of solid tumours (American Cancer Society, 2012).

Lung cancer risk increases in an approximately linear fashion with radiation dose (Preston et al., 2009; Thompson et al., 1994). Dose-response relationships with SPLC have been reported among patients given thoracic radiotherapy for Hodgkin lymphoma (Gilbert et al., 2009; Travis et al., 2002; van Leeuwen et al., 1995) and breast cancer (Inskip, Stovall, & Flannery, 1994). Among patients diagnosed with a second primary lung cancer five or more years after breast cancer treatment, the risk of lung cancer increased linearly with radiation dose, 8.5% per Gy (95% CI: 3.1–23.3%; $p < 0.001$) (Grantzau, Thomsen, Væth, & Overgaard, 2014); the association was exaggerated for ever-smokers with an excess rate of 17.3% per Gray (95% CI: 4.5 – 54%; $p < 0.005$). A

dose-response relationship with lung cancer risk has also been reported for alkylating agents used in chemotherapy for Hodgkin lymphoma, after controlling for radiation exposure and tobacco smoking (Travis et al., 2002).

Smoking can exacerbate the SPC risks associated with radiotherapy. In a nested case-control study of future non-breast cancer among 23,627 early breast cancer patients treated with radiotherapy, 91% of the cases (those who developed SPLC) were ever-smokers versus only 40% of the controls (those who did not develop SPLC) (Grantzau et al., 2014). Increased lung cancer risk in those treated with radiation may be confounded by smoking status since it is related to both lung cancer risk and exposure to radiotherapy. Smoking-related comorbidities can make surgery a non-viable option, thus smokers may be more likely to be treated with radiation therapy alone than with potentially curative surgery.

There is also evidence that radiotherapy can modify the effect of cigarette smoking on subsequent lung cancer risk. Kaufman and colleagues conducted a population-based nested case-control study among women with breast cancer and found that, compared to non-smokers who were not exposed to radiotherapy, odds ratio for SPLC were 5.9 (95% CI: 2.7 – 12.8) for ever-smokers who did not receive radiotherapy and 18.9 (95% CI: 7.9 to 45.4) for ever-smokers who received it. Radiotherapy did not increase the risk of SPLC in non-smoking women (OR = 0.6; CI: 0.2 – 2.4) (Kaufman, Jacobson, Hershman, Desai, & Neugut, 2008).

2.2.2 Age

There is a marked increase in lung cancer incidence after age 60. The percentage of men in the United States who develop lung cancer in the following 10 years is 2.05 at age 60, compared to 0.71 at age 50 and the figures among women are 1.56 and 0.58, respectively, based on data from 2009 (Howlander et al., n.d.). Although not directly analogous, the most comparable figures for 2009 in Canada also show a considerable increase as follows: 2.3% of men aged 60-69 compared to 0.7% of men aged 50-59. The respective figures for Canadian women are 1.8% and 0.7%.

Table 2.2 Probability (%) of developing lung cancer in next 10 years by age, United States and Canada, 2009

United States			Canada		
Age	Males	Females	Age	Males	Females
50	0.71	0.58	50-59	0.7	0.7
60	2.05	1.56	60-69	2.3	1.8

Similarly, Tabuchi et al. reported marked increases in absolute risk of subsequent primary cancer (SPC) of any type among Japanese cancer survivors aged 60-69 compared to 50-59; this was the case for maximum follow-up lengths of both 5 (4.1% vs 2.7%) and 10 years (8.6% vs 5.8%) (Tabuchi, Ito, Ioka, Miyashiro, & Tsukuma, 2012). A cut-point of 60 years of age was also used to distinguish between early vs. late onset of a family history of lung cancer in the LLP risk prediction model (Cassidy et al., 2008).

2.3 Lung Cancer Screening

Ideally, screening facilitates the detection of lung cancer at an early stage when treatment is most likely to be effective; currently, lung cancers are typically diagnosed at an advanced stage. The hazards associated with lung cancer screening include futile detection of small, aggressive tumours which have already spread or indolent disease which would never have led to advanced disease if ignored, complications from

diagnostic workup, the consequences of both false-positive and false-negative results, exposure to radiation and cost. However, the potential of decreased lung cancer mortality and improvement of quality of life can make screening worthwhile, provided that it is cost-effective enough to invest in (Wood et al., 2012). A systematic review of the harms and benefits of low-dosage computed tomography (LDCT) screening for lung cancer concluded that screening may be beneficial to those at high-risk but the potential harms and generalizability of results requires further investigation (Bach et al., 2012).

Precise guidelines for lung cancer screening in the United States are provided by the United States Preventive Services Task Force (USPSTF). Annual lung screening with LDCT is recommended for those aged 55-80 years with at least 30 pack-years history of smoking, who are either currently smoking or have quit within the last 15 years. A person should not be screened if they have a substantially shortened life expectancy or are either unable or unwilling to have curative lung surgery and it should be discontinued if a person has not smoked for 15 years (Moyer, 2014). Medicare also has its recommendations for lung cancer screening since it reimburses those over age 65. The Center for Medicare and Medicaid Services considers those aged 55-77 with $\geq$ 30 pack year smoking histories and who have quit within 15 years to be eligible for lung cancer screening (Centers for Medicare & Medicaid Services, 2015).

It has been shown that both the harms and benefits of LDCT screening are more favorable among those who have a higher risk of lung cancer. Among the participants of the NLST, the number of lung-cancer related deaths per person-years that were prevented by LDCT in comparison to chest radiography increased with lung cancer risk (Kovalchik et al., 2013).

2.4 Risk Prediction Models for Lung Cancer

Risk prediction models are mathematical functions which use predictor variables (covariates) to estimate the absolute probability that a particular outcome will occur within a specific time period for an individual, given their predictor profile (Moons et al., 2012). Predictors might include subject characteristics, examination results, or imaging results and so forth. Both independent (predictors) and dependent (outcome) variables may be continuous or categorical.

In developing a multivariable risk prediction model, the most influential covariates are identified from a preselected set of candidate predictors. A modeling technique is then specified and used to determine how much of the variation in the outcome can be explained by the predictors. Relative weights are assigned to each predictor as a measure of the magnitude and direction of their relationship with the outcome. These are then combined to produce an overall risk score.

Before model development begins, a number of issues must be addressed including:

- Selection of relevant predictors
- How to handle missing data
- How to model each predictor
- Strategy for inclusion of covariates
- How to evaluate the performance of the model

Several strategies exist for determining which covariates should be included a prediction model. Often a sequence of hypothesis tests is applied and predictors are

chosen on the basis of statistical significance. However, this approach confuses hypothesis testing with prediction modeling and can lead to the omission of useful predictors, thus selection based on effect estimates is preferable (Royston, Moons, Altman, & Vergouwe, 2009). Regression analyses may show statistically significant associations between risk factors and lung cancer but may fail to improve the predictive capacity of a model to any useful extent. The opposite is also true in that non-significant predictors may still improve prediction if included. Selecting predictors based on prior scientific knowledge is preferable to arbitrary inclusion of potentially useful factors.

Models are sometimes built up from a constant and a single covariate (forward selection) but backwards elimination (initially including all potential predictors and sequentially removing the least influential ones) is often preferred since suppressor effects, that is, effects which only occur when another variable is held constant, can otherwise be missed (Field, 2009). Backwards elimination usually performs better than forward selection when some covariates are highly correlated (collinearity) and examines a full model fit which is the only method that provides accurate standard errors and *p*-values (Harrell, 2001).

The most widely used modeling technique for dichotomous outcomes is perhaps logistic regression (Weisburd & Britt, 2014). Logistic regression fits the data according to an S-shaped *logistic* curve which follows a linear model in the middle of its distribution but flattens as it as it approaches 0 or 1 such that the outcome is constrained to values between 0 and 1. This ensures that predictions cannot be outside the range of actual values of what we are trying to predict (the probability of a binary outcome). In OLS regression, the prediction of Y is represented by the following equation: $Y = Y_0 +$

$\beta_1 X_1 \ldots \beta_n X_n$ which can yield values greater than 1 or less than 0. In logistic regression, rather than predicting Y itself, we predict the natural logarithm (ln) of the odds of the event occurring, also known as the logit of Y. Mathematically, this model is represented as:

$$ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_1 X_1 \ldots \beta_n X_n$$

Regression coefficients (B_i) on the right hand side of the equation still represent the change in the dependent variable for every one unit increase in the independent (X_i) and can be exponentiated to produce odds ratios (ORs), i.e. the relative odds of the outcome occurring given a particular exposure compared to being unexposed (Szumilas, 2010).

The estimates produced by logistic regression are based on maximum likelihood estimation (MLE) technique which attempt to maximize the probability that the regression estimates will follow a distribution similar to that of the observed data. This process attempts to improve on tentative mathematical solutions based on a likelihood function. This function measures the probability of the observed results given the current regression coefficients. Typically, logistic regression uses -2 times the natural logarithm of the likelihood function (-2LL). Successive iterations of models are compared until the change in likelihood function is negligible – the solutions have 'converged' (Weisburd & Britt, 2014).

2.4.1 Model Performance

Model performance refers to the quality of the predictions made from it. The overall predictive performance of a model can be assessed in a variety of ways, with the distance

between the predicted outcome and the actual outcome $\left(Y - \hat{Y}\right)$ playing a central role. These distances relate to the 'goodness-of-fit' of a model - better models have smaller distances. (Steyerberg et al., 2010).

The amount of total variation in the outcome variable explained by the model (R^2) is the most common overall fitness score for continuous outcomes (linear regression models) but for dichotomous outcomes, models are scored with the logarithm of predictors:

$$Y * \log(p) + (Y - 1) * (\log(1 - p)).$$

R^2 statistics cannot be calculated for models with binary outcomes as modeled using logistic regression models. Several pseudo-R^2 s have been devised as analogous alternatives but none is equivalent and they can vary widely for the same data. Instead, measures based on the difference in -2 log likelihood between models are used (Steyerberg et al., 2010). Overall performance measures comprise two main characteristics, discrimination and calibration, both of which are usually evaluated individually. The Brier score, a measure of average prediction error, is also used to assess overall model performance. It is calculated by squaring the absolute difference between the observed and predicted probabilities for each individual and then taking the average of this across the sample (Brier, 1950). A lower Brier score indicates lower average error and thus better prediction and a Brier score of 0.25 represents random classification.

2.4.1.1 Discrimination

Discrimination refers to the accuracy with which a predictive model classifies those who experience the outcome and those who do not. Put simply, it is the probability that a

single randomly selected participant will be correctly classified by the model as either experiencing the outcome (being diagnosed with lung cancer in this case) or not. Several measures are available to assess this for a binary outcome. The area under the receiver operating characteristic (ROC) curve (AUC) statistic is the most common indicator of discrimination for general linear models. The statistic describes how well a model can rank order predicted values in order to separate those who experienced the outcome from those who did not. Since rank order is used, the statistic is not a function of the actual predicted probabilities and as such is insensitive to any distances between them. The AUC represents the probability that a positive observation ($Y = 1$) will be ranked higher than a negative one ($Y = 0$) (Fawcett, 2006).

The AUC is represented graphically by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) (Steyerberg et al., 2010). The sensitivity and specificity are paired at all consecutive cut-offs points (the predicted probability for each observation). The (theoretical) maximum value of the AUC is 1.0, indicating perfect classification, whereas a value of 0.5 indicates the classifications are no better than random chance. An AUC of between 0.5 and 0.7 is considered by Hosmer and Lemeshow as poor discrimination, with an AUC of 0.7 to < 0.8 indicating acceptable discrimination, and an AUC of 0.8 to <0.9 representing excellent discrimination and anything higher to be outstanding discrimination (Hosmer et al., 2013).

2.4.1.2 Calibration

Calibration refers to the agreement between the probabilities of the outcome as estimated by the model versus the observed event probabilities (Steyerberg et al., 2010). Calibration

plots provide a graphical representation of this agreement, with predictions on the x-axis and observed outcomes on the y-axis. Perfect predictions lie on a 45-degree line (theoretically). For continuous outcomes, this basic scatter plot will suffice but for binary outcomes, the y-axis only contains the values 0 and 1, so smoothing techniques can be used to approximate the observed outcome probabilities, $p(Y = 1)$ in relation to the predicted probabilities using locally weighted linear regression algorithms (Steyerberg et al., 2010). The intercept (α) of a calibration plot indicates whether the predictions are systematically too high or too low (ideally, $\alpha = 0$) and the slope (β) should ideally be equal to 1, indicating perfect prediction. Numerical measures of calibration can include median and 90th percentile absolute error (differences between observed and predicted values) (Hyndman & Koehler, 2006) as well as a sub-component of the Brier score, the Spiegelhalter statistic.

The correspondence between the predictions of a logistic regression model and observations or its 'goodness-of-fit' is often assessed by the Hosmer-Lemeshow test (Hosmer & Lemeshow, 1980). A cross-tabulation of observed and expected values is created, usually per decile of predicted probabilities by binary outcome status. However, the Hosmer-Lemeshow statistic has been criticized for several reasons (Desai, Bruce, Desai, & Druss, 2001; Hosmer, Hosmer, Le Cessie, & Lemeshow, 1997; Kramer & Zimmerman, 2007; Steyerberg, 2009; Vittinghoff, Glidden, Shiboski, & McCulloch, 2011):

- The standard binning of probabilities by decile is arbitrary and using a different number of groups, even 9 or 11 can completely change the result.

- The statistic does not possess considerable power to detect lack of calibration and does not adequately penalize overfitting.
- Adding non-linear or interaction terms which ought to improve model fit may, paradoxically, result in lower p-values (a poorer result).
- With a large enough sample size, statistical significance (a failed test) can be achieved even when observed and predicted probabilities closely agree.

Overfitting is a situation in which a statistical model describes too much random error or noise rather than the underlying relationship and generally occurs when a model is overly complex, with too many predictors relative to the number of observations. In short, they fit the dataset used to estimate the parameters better than different data and generally have poor predictive performance in external or new or test data as a result. This can be problematic with large sample sizes where the statistical significance of relatively innocuous differences between observed and expected values can be misinterpreted as poor calibration when in reality this is expected to occur by chance. Overfitting can be corrected for using statistical resampling techniques such as the Quenouille-Tukey jackknife (Efron & Stein, 1981) and bootstrapping (Efron, 1979). Bootstrapping allows for repeated resampling of data to produce numerous alternate datasets of the same size as the original. A statistic can be re-calculated for each of these bootstrapped datasets and then averaged to produce bias-corrected estimates and confidence intervals. To help minimize the likelihood of overfitting, the number of outcome events per the number of predictors in the model should be at least 10 but preferably >50 (Steyerberg et al., 2010).

2.4.2 Notable Lung Cancer Prediction Models

Relatively few models have been developed to estimate lung cancer risk. One of the earliest models, the Harvard Cancer Risk Index (Colditz et al., 2000), was based on literature reviews and general consensus. Key risk prediction models based on more reliable data were subsequently developed: the Bach model, (Bach et al., 2003) the Spitz model, (Spitz et al., 2008) and the Liverpool Lung Project (LLP) model (Cassidy et al., 2008). Smoking duration and occupational asbestos exposure are factors common to all three models. However, they differ in terms of other risk factors such as lung-related comorbidities (emphysema vs pneumonia), history of prior malignancies, and definition of family cancer history. A summary table of the characteristics of these models is available for ease of comparison (Etzel & Bach, 2011).

The Bach model was developed using data from the β-Carotene and Retinol Efficacy Trial (CARET), a randomized controlled trial of β-Carotene and vitamin A supplements among 14,254 heavy smokers and 4,060 asbestos-exposed participants (Bach et al., 2003). Bach et al used Cox proportional hazards regression to develop 1-year probability models for lung cancer diagnosis. The models contain the same variables: number of cigarettes per day, number of years smoked, number of years quit (among former smokers), age, sex, and exposure to asbestos. These 1-year models were run recursively 10 times to predict 10-year absolute lung cancer risk. Bach et al. estimated 10-year absolute risk "because it is probably in excess of the time it takes for lung cancer to progress from an undetectable size to an untreatable stage; consequently, it is a useful perspective from which to counsel patients about screening" (Bach et al., 2003).

The Spitz model was based on 1,851 lung cancer cases matched with 2001 controls according to age, sex, race and smoking status (never, former, current). Logistic regression was used to create separate models for each smoking status category and estimate the relative risk of developing lung cancer. Relative risks from these models were combined with age- and sex-specific incidence rates to present 1-year absolute lung cancer risk (Spitz et al., 2007).

The LLP model was derived using data from 579 lung cancer cases and 1,157 age-, sex- matched controls taken from a case-control study that was part of the Liverpool Lung Project. The cases were also frequency matched on smoking status so the impact of smoking could not be accurately assessed. Logistic regression models were used to estimate relative lung cancer risk. Case-control data and regional age-standardized lung cancer incidence rates were combined to present 5-year absolute lung cancer risk (Cassidy et al., 2008). Prior to the PLCO Trial, lung cancer risk prediction models were limited in terms of the number of possible predictors, and demonstrated low predictive performance (Tammemagi et al., 2011). Data from the PLCO trial allowed improved models to be developed using a wider range of risk factors, a prospective study design, and evaluation of non-linear effects. Factors comprised age, socioeconomic status (using education as a proxy), BMI, family history of lung cancer, COPD, recent chest x-ray, smoking status (never, former, or current), pack-years smoked, and smoking duration were used to predict the probability of lung cancer diagnosis during the ongoing follow-up in both the general population (model 1) and a sub-cohort of ever-smokers (n=38,254) (model 2). Model 2 also incorporated smoking quit-time (the time in years since the person has permanently quit smoking). Both models demonstrated high discrimination.

The AUC for models 1 and 2 were 0.857 and 0.805 respectively and the values for the calibration slopes were 0.987 and 0.784, respectively (Tammemagi et al., 2011).

The 2011 lung cancer risk prediction model based on the PLCO trial data was later modified to facilitate compatibility with data from the NLST. The resultant logistic regression model predicting the probability of lung cancer diagnosis during the 6-year study period ($PLCO_{M2012}$) demonstrated an AUC of 0.803 in the development dataset (PLCO control-group smokers) and 0.797 in the validation dataset (PLCO intervention-group smokers). Mean absolute error and 90^{th} percentile absolute error were 0.009 and 0.042, respectively. (Tammemagi et al., 2013). Predictors included: age, race/ethnicity, education level (a surrogate for socioeconomic circumstance), BMI, COPD, personal history of cancer, family history of lung cancer, smoking status, smoking intensity, smoking duration, and smoking quit time. $PLCO_{M2012}$ was later modified to produce a model applicable to never-smokers as well as ever-smokers (PLCOall2014) (Tammemagi et al., 2014). Since the current analysis is based on data from the PLCO and NLST, it makes the most sense to use the use the models developed using PLCO data as a basis for potential improvement. These models have shown superior performance compared to previous efforts, they have been thoroughly described and validated, and are well accepted by the research community.

CHAPTER 3: METHODOLOGY

3.1 Introduction

The design of the study and its methodology is summarized in this chapter including acquisition of the data, specific measures investigated and the steps taken to organize and utilize the data as appropriate. Details of the specific analytic strategies employed to address the research questions proposed in the first chapter conclude this section.

3.2 Design and Context of the Original Studies

This research is derived from two multi-center randomized controlled trials conducted in the United States, details of which are provided in the following sections. These original studies collected a comprehensive array of patient characteristics and extensive data with particular respect to risk factors for lung cancer. Additional variables were synthesized from the available data as part of this project in order to better assess the effects of having a previous history of cancer.

3.2.1. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial

The PLCO randomized controlled trial investigated the effect of screening for lung cancer using chest radiography versus usual care with respect to lung cancer specific mortality (Oken et al., 2011). The main outcome measure was mortality from lung cancer, with lung cancer incidence, complications associated with diagnostic procedures and all-cause mortality as secondary endpoints. A detailed description of the design and operation of the PLCO can be found elsewhere (Prorok et al., 2000) but a brief summary is presented here.

3.2.1.1 Recruitment and Randomization (PLCO)

Enrollment began in 1993 at 10 screening centers in the United States and continued until 2001. A total of 154,942 participants were recruited by mass mailing, of which 154,901 were assigned to either the intervention or usual care group using block randomization, stratifying them by screening center, sex and age. Of those randomized, 77,445 received usual care and 77,456 in the intervention arm were screened with annual posterior-anterior view chest X-rays, 67,038 of whom received an initial screening at baseline followed by three additional annual screenings (except for never-smokers randomized after April 1995 who were not offered the third screen) (Oken et al., 2011). All participants in the PLCO trial signed informed consent documentation approved by both the National Cancer Institute and their local institutional review board (Oken et al., 2011).

3.2.1.2 Eligibility Criteria (PLCO)

Eligibility was restricted to patients not participating in other cancer screening or primary prevention trial, aged between 55 and 74 years. Among other criteria, those who had been previously diagnosed with prostate, lung, colorectal or ovarian cancer at any time, or receiving treatment for any cancer other than basal cell or squamous cell skin cancer were excluded (Oken et al., 2005). Some survivors of prostate, lung and colorectal cancer remained in the dataset however, so it was possible to investigate the effects of these cancer types to some extent in the current study.

3.2.1.3 Data Collection (PLCO)

Upon entry to the study, participants completed a self-administered baseline questionnaire comprising questions regarding their sociodemographic characteristics (age, race/ethnicity, sex, marital status, education), family history of lung cancer, personal medical history, cigarette smoking history and cancer screening history (Oken et al., 2011).

Cancers diagnosed at any site and all deaths were ascertained by a mailed annual update questionnaire. Repeat questionnaires were sent if initial ones were not returned, else the participant was contacted by telephone. Periodic checks against the National Death Index complimented the follow-up (Oken et al., 2011). All deaths potentially related to a PLCO cancer were assessed by reviewers blinded to the trial group.

Lung cancer-specific deaths were defined as those whose underlying cause was lung cancer itself or a result of treatment for it. Among those in the intervention group, cancers were defined as *screening-detected* if they were diagnosed within a 9 month period after a positive screen result (or the diagnostic evaluation linked to it). Non-screening detected cancers were classified as *interval* if they occurred up to one year after the most recent scheduled screening or *post-screening* if they occurred more than one year after the final screening (Oken et al., 2011).

3.2.2 The National Lung Screening Trial

The NLST is a multicenter randomized controlled trial (RCT) designed to compare two methods of screening for lung cancer: low-dose computed tomography (LDCT) and chest radiography. The effects of these techniques on lung cancer mortality rates among current and former smokers were assessed. The primary endpoint of the study was lung cancer

mortality, and secondary endpoints included lung cancer incidence, treatment or screening related morbidity, quality of life and cost-effectiveness (National Lung Screening Trial Research Team, 2011). The design of the NLST has been described previously (National Lung Screening Trial Research Team, 2011) but a summary is presented here. Initiated in 2002, the NLST is a study funded by the U.S. National Cancer Institute (NCI), Division of Cancer Prevention and is based on a smaller-scale pilot study known as the Lung Screening Study. The NLST was jointly conducted by Lung Screening Study (LSS) screening centres (mostly PLCO centres), and the American College of Radiology Imaging Network (ACRIN), funded by the NCI Division of Cancer Treatment and Diagnosis. The study protocols and data collection forms used by LSS and ACRIN were harmonized (National Lung Screening Trial Research Team, 2011).

3.2.2.1 Recruitment and Randomization (NLST)

Thirty-three medical centres throughout the United States took part in the NLST (Clinicaltrials.gov, NCT 00047385). Participants were recruited using a variety of strategies including direct mailings and use of mass media with support from both the NCI's Cancer Information Service and the American Cancer Society. The Census Department's Tobacco Use Supplement of the Continuing Population Survey for 2002-2004 was used to ascertain whether the study population was representative of those eligible to participate from the whole US population. Enrolment was completed between 2002 and 2004. A total of 53,452 participants were included. Targeted recruitment strategies were used to enrol minority populations (Aberle et al., 2010). Participants were randomly assigned to either the CT or chest radiograph arm, stratified by site, sex and 5-year age group using a block size of six or eight (Aberle et al., 2011). The study was

conducted after approval by an institutional review board at each institution. Each participant provided written informed consent (Aberle et al., 2010).

3.2.2.2. Eligibility Criteria (NLST)

The NLST required participants to meet the following basic criteria (among others described in further detail by Aberle et al. (2011)) in order to be eligible to take part in the trial:

- Aged between 55 – 74 years
- Have a minimum of 30 pack-years history of cigarette smoking
- Former smokers must have quit within 15 years of randomization
- No history of lung cancer
- No participation in other cancer screening trial or cancer prevention study
- No evidence of any cancer other than non-melanoma skin cancer or carcinoma in situ within the preceding 5 years
- Signed informed consent form

Those who still smoked cigarettes regularly were defined as current smokers. Some lung cancer survivors remained in the dataset and were included in the current analysis.

3.2.2.3 Data Collection (NLST)

Low-dose helical CT scans were conducted using NLST-approved equipment according to a protocol appropriate at the time of the trial. The obtained LDCT images were electronically transmitted to ACRIN- or LSS-maintained central repositories (National

Lung Screening Trial Research Team, 2011). These scans were interpreted by NLST-approved radiologists.

NLST-certified chest radiographic machines were used to collect X-ray images. NLST-ACRIN images were permanently archived at ACRIN headquarters and NLST-LSS images at the screening centres. Protocol requirements for chest radiograph interpretation were identical to those for used for interpreting LDCT images (National Lung Screening Trial Research Team, 2011).

Data on patient characteristics were collected using epidemiological questionnaires upon study entry (National Lung Screening Trial Research Team, 2011). Each person was offered three lung cancer screenings – one at baseline and two annual screenings thereafter. LDCT images in the NLST were assessed for the presence of lung nodules, masses or other abnormalities suspicious for lung cancer (positive suspicious results) as well as other findings of potential clinical relevance. Morphological features were recorded for all non-calcified nodules and masses with a minimum diameter of 4mm (National Lung Screening Trial Research Team, 2011).

The datasets from the PLCO and NLST trials were harmonized to the extent possible, that is, common information collected on the participants of both trials so that associations based on these factors could be analyzed using a far greater sample size. This was done prior to this study by Information Management Systems, Inc. (IMS) on behalf of the NCI.

3.3 Analytical Strategy

All statistical analyses for the current study were performed using the Stata software package (StataCorp. 2012. *Stata Statistical Software: Release 13.1*. College Station, TX: StataCorp LP.). Prior to analysis, some existing variables were modified and several new variables were created. The variables representing the age and BMI of the participant were centered to the mean (mean subtracted from the value), rounded to the nearest integer (62 years and 27 kg/m^2, respectively) in order to make the regression coefficients for those variables more straightforward to interpret. Data cleaning consisted of looking for outliers and handling them as appropriate. Missing values had already been defined as such in the dataset by IMS.

Descriptive statistics for the risk factors analyzed consisted of calculating the mean and standard deviation for continuous variables and frequency and percentage for categorical ones. These statistics were stratified by trial (PLCO vs. NLST) and outcome status (lung cancer diagnosed during follow-up or no lung cancer). To test for statistically significant differences across each pair of strata, two-sample independent t-tests or Mann-Whitney U were performed for continuous variables as appropriate, depending on whether normality could be assumed. Chi-squared tests were conducted for categorical variables or Fisher's exact tests in cases where any expected cell count was less than 5.

Logistic regression models were used to assess the association of each variable under investigation with subsequent primary lung cancer (SPLC) before and after adjustment. This approach was chosen because the outcome is dichotomous and logistic regression would facilitate direct comparisons with established models. Also, individual-level predicted values are straightforward to compute from such models. Follow-up was

truncated at six years to further ensure comparability with previous models as well as between the participants of the PLCO and NLST.

All univariate and multivariable analyses were conducted for PLCO and NLST participants separately followed by re-testing any notable associations with harmonized data, where appropriate, i.e., heterogeneity of effects did not exist. The associations of all study variables with lung cancer were first assessed in univariate analysis and then adjusted for the covariates in the $PLCO_{all2014}$ (Tammemagi et al., 2014)): age, race/ethnicity, education, BMI, COPD, family history of lung cancer, smoking status, smoking intensity, duration and quit time. Personal history of cancer was omitted since the variables being tested are variants of the same factor. Ordinal variables were assessed for evidence of a dose-response pattern where applicable. Ordinal variables were treated as continuous if it was sensible to do so, in order to minimize degrees of freedom in the model.

Multivariable fractional polynomials (MFPs) were used to determine whether any of the predictors from the $PLCO_{all2014}$ should be modeled using non-linear transformations and which transformations were most suitable. Using this approach, the best power transformation for a variable x (x^p) is found where the power p is chosen from the following set of candidates {-2, -1, -0.5, 0, 0.5, 1, 2, 3} (Royston & Altman, 1994). It has been acknowledged by others that low order polynomials (such as quadratic or cubic) can model only a few possible curve shapes and high order ones fit the data more closely but often do so poorly for extreme values of x. This restricted set of fractional polynomials have been shown to have extensive flexibility and since iterative algorithms for testing possible candidates are implanted in most statistical software

packages, this approach is simple to carry out. Transformations are straightforward to fit into standard regression models (Royston & Altman, 1994). Suggested transformations were used in place of the original linear terms when conducting adjusted analyses. The –mfp- command in Stata was used to conduct the MFP analysis. The number of degrees of freedom for the transformations was capped at 2 in order to exclude non-linear relationships that were implausible - this limits the complexity of the curve representing the relationship between the continuous predictor and the probability of the outcome (i.e. the number of times it can change direction). For example, if the probability of lung cancer increases sharply beyond a certain smoking intensity threshold, the gradient should be allowed to become steeper to reflect this but it does not make biological sense that the curve should be allowed to become inverted at even higher smoking intensities.

Since the influence of the covariates from $PLCO_{all2014}$ was established prior to the study, conclusions were only drawn from adjusted analyses. Effects observed in unadjusted analyses may be attributable to known factors hence associations seen after accounting for these influences are more important. Additionally, all associations were also adjusted for clustered sampling based on screening center, in order to control for both known and unknown differences between centers such as the proficiency of the medical staff, access to technology, sociodemographics of their patients or even localized environmental factors which could be carcinogenic, all of which could introduce bias into the analyses. This adjustment increases the robustness of the standard errors.

Personal history of cancer

1. Type of cancer

Are any specific types of previous cancers strong independent predictors of subsequent lung cancer?

The dataset contained variables that show whether or not the participant reported having previously been diagnosed with cancer (0=no, 1=yes) at the following anatomical sites:

Breast

Digestive: Colorectal, esophageal, pancreatic, stomach

Endocrine: Thyroid

Genitorurinary: Bladder (including transitional cell), kidney, prostate, testicular

Gynecologic: Cervical, uterine

Head and neck: Laryngeal, nasopharyngeal, oral

Hematologic: Hodgkin lymphoma, non-Hodgkin lymphoma

Lung

The variables for Hodgkin lymphoma, non-Hodgkin lymphoma, prostate, testicular and uterine were created manually based on data provided by the PLCO participants for up to four previous cancers. Nasal, oral, pharyngeal and transitional cell cancers were reported for NLST participants only, but the PLCO did contain data on previous cancers classified as 'nasopharyngeal' so nasal and pharyngeal cancers were combined for the NLST participants such that one variable would be applicable to those in either trial. Since transitional cell cancers are a subset of bladder cancers, these were combined into a single variable. The variable denoting previous COPD diagnosis was

modified to also include diagnosis of emphysema and chronic bronchitis since COPD encompasses these conditions.

Logistic regression models were used to quantify the association of each of these variables with lung cancer before and after adjustment and the sample was restricted to males or females only when studying the effects of sex-specific cancer types. Associations were evaluated separately among PLCO and NLST participants. When testing the associations for each previous cancer type, some models would not converge due to complete separation. In these instances, to circumvent the problem, a single participant who developed lung cancer during follow-up was treated as having survived the type of cancer being tested, even though they did not.

2. Number of previous cancers

Is the number of previous cancers with which a person has been diagnosed an important predictor of future lung cancer?

A score based on the sum of all dichotomous (0/1) variables representing each specific cancer type was generated for all participants and coded as (0, 1, 2+). Only 9 people reported more than two different previous cancer types, hence grouping them together to form the 2+ category. Further questions specific to PLCO participants gathered information regarding up to four previous cancers. This information was incorporated into the 'number of previous cancers' variable so as to capture as much data as possible. After conducting adjusted analyses for this variable, post-estimation Wald tests were performed to determine whether any difference between the 1 and 2+ categories was statistically significant for either the PLCO or NLST sub-groups.

3. Age at cancer diagnosis

Is being diagnosed with cancer before age 60 associated with future lung cancer risk more strongly than at 60 or older?

PLCO participants reported their age at diagnosis for up to four cancers. In both trials, age at diagnosis was also recorded for each of the specific previous cancer types listed in Q1.

Age at diagnosis was categorized in ordinal form as follows: no previous cancer, previous cancer(s) diagnosed at 60 years or over, and at least one cancer diagnosis before age 60, respectively). A post-estimation Wald test was performed to determine whether the any difference between the 60+ and <60 categories was statistically significant.

Do any of these associations lead to improved prediction models?

Variables for which positive effect sizes of at least 15% (i.e. ORs $\geq$ 1.15) were observed after adjustment were considered for inclusion in multivariate modeling. Only cancers that showed strong effects in the participants of both trials (where possible) were considered as it did not make sense to include associations that were inconsistent. Although ultimately down to a priori knowledge and judgment, the threshold was chosen because it is large enough to be clinically relevant and far enough away from the null (no effect) so as to be unlikely due to residual confounding, but sensitive enough so that important effects were unlikely to be ignored. A dichotomous summary variable was created to indicate a previous diagnosis of any of these previous cancer types with elevated risks for inclusion in later models. This was done in order to minimize the degrees of freedom in the model and maximize the predictive utility of the information.

Two further binary personal history of cancer variables were created. The first represented a history of smoking-related cancers since tobacco use likely increases a tendency towards multiple cancers at such sites: lung, nasopharyngeal, oral, laryngeal, pancreatic, esophageal, kidney, stomach, bladder, cervical or colorectal (American Cancer Society, 2014)) and the second represented a history of head and neck cancers as these are the most strongly associated with SPLC according to previous studies. All three of these dichotomous variables were considered as candidates to replace the original 'personal history of any cancer' variable in the $PLCO_{all2014}$ model. Additionally, interaction effects between the best performing personal history of cancer variable and age at previous cancer diagnosis (60+/<60 years) was evaluated. Other interaction effects based on a priori knowledge included sex-race/ethnicity, sex-smoking intensity and sex-smoking duration, but none of these interaction terms were found to be significant so were not carried forward to more complex models. Ideally, interactions with age of previous cancer diagnosis should have been tested for each specific previous cancer site but this would have made the analysis somewhat verbose and statistical power was minimal so priority was given to testing the interaction with only the original summary term.

Further models were then built by including all variables that improved discrimination alongside the $PLCO_{all2014}$ covariates, and then those variables which increased discrimination by negligible amounts were successively removed to produce the most parsimonious model.

Predicted probabilities were calculated using the formula:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots \beta_n X_n)}}$$

The model error was calculated as the absolute difference between observed probabilities and those predicted by the model. Calibration of the final model was assessed based on median and 90th percentile absolute error and the p-value of the Spiegelhalter's z-statistic which is provided by Stata when computing the Brier score; this statistic tests the extent to which an individual Brier component is extreme, with a larger p-value (> 0.05) indicating better prediction (Spiegelhalter, 1986).

Bootstrapping techniques were used to produce bias-corrected confidence intervals for both the AUC and Brier scores; in each case, 1000 replications were used, which is generally considered to be a reasonable number for such estimates (Duval, 1993); internal validity was gauged based on the width of these confidence intervals.

The amount of missing data was assessed by comparing the number of observations used in the estimation of the final model to the total number available for that analysis. If more than 10% of the participants were excluded, multiple imputation of missing data was considered in order to maximize the sample size used in estimation. Finally, the number of events in the sample will be divided by the number of predictors to ensure that the ratio between the two is sufficient to minimize overfit (> 50).

3.4 Modeling Assumptions

Unlike linear regression, binary logistic regression does not require a linear relationship between predictors and the outcome variable and the independent variables do not need to be normally distributed. Homoscedasticity of residuals is not required and variances do

not need to be equal across all values or levels of the predictors. Levels of measurement for the predictors are not limited to interval or ratio only. However, assumptions that do apply include the following (Menard, 2002; Pregibon, 1981; Tabachnick & Fidell, 2007):

- The model is fitted correctly, meaning that no important variables are omitted and no extraneous variables are included. The inclusion of all predictors established in previous models as well as using a backwards selection method for new covariates should ensure that this is the case. Misspecification of the logistic function can be tested using the -linktest- post-estimation command in Stata. The specification link test is based on the principle that if a model is properly specified, additional predictors should only achieve statistical significance by chance. Linktest rebuilds the model using the linear predicted value (_hat) and the squared form of this value (_hatsq) as predictors. If the model is properly specified, _hatsq should not be statistically significant. If it is so, this may suggest that relevant predictors are omitted or the logit of the outcome is not a linear combination of the predictors and thus logit function is not the correct approach to use. Suggested transformations from the MFP analysis will ensure that non-linear relationships are modeled appropriately.
- Graphical plots of the residuals from the model against the predicted values can be an effective way to check whether the level of measurement error for the predictors is problematic. Standardized Pearson residuals represent the relative deviation between observed and fitted values. The deviance residual measures discrepancy between the limits of the observed and fitted likelihood functions, analogous to the raw residual in OLS regression. Finally, the Pregibon leverage

statistic measures the influence of each observation which is useful when deciding whether to remove a problematic observation from the model. Each of these statistics will be plotted against the predicted values in order to assess this assumption.

- The observations are independent, that is, the data from one participant should not depend on that of another or be from any dependent sampling design.
- The predictors should be independent of one another, that is, no strong multicollinearity should be present. Centering variables can help reduce collinearity and problems with collinearity will be checked with variance inflation factor (VIF) statistics. Any strongly collinear variables will either be combined if possible or removed from the model as appropriate.
- The independent variables are not linear combinations of one another. If this is the case for a predictor, Stata will automatically drop the variable from the model

CHAPTER 4: RESULTS

Descriptive statistics for the study participants are presented separately for each screening trial in Tables 4.1 and 4.2. The former describes factors included in the $PLCO_{all2014}$ model, while the latter summarizes characteristics central to this thesis project. Results from the current analyses is described below. Results for PLCO and NLST participants are provided separately. Both unadjusted and adjusted results are presented in the summary tables, however only the latter will be discussed here given that this thesis is concerned with associations that persist after accounting for pre-established risk factors. The written summary is presented in the following manner:

- Comparisons of participant characteristics first between screening trials based on crude estimations using the data from Tables 4.1 and 4.2.
- Assessment of the impact of participant characteristics on cumulative lung cancer incidence.
- Key study findings (Tables 4.3 – 4.9) summarized separately by research question and discussed firstly among PLCO participants and then for those in the NLST as appropriate.
- Evaluation of whether assumptions applicable to the modelling procedures were met.

Table 4.1 - Sample characteristics (variables used in PLCOall2014) for PLCO (N=154,900) and NLST (N=53,452) participants

			Screening Trial		Developed Lung Cancer During Follow-up		
	Harmonized (N = 208,352)	PLCO (N = 154,900)	NLST (N = 53,452)		PLCO (n = 1,636)	NLST (n = 1,925)	
	N (%)* / Mean (SD)	**N (%)* / Mean (SD)**	**N (%)* / Mean (SD)**	**P-value†**	**N (%)§ / Mean (SD)**	**N (%)§ / Mean (SD)**	**P-value‡**
Age	62.3 (5.3)	62.6 (5.4)	61.4 (5.0)	<.001	64.7 (5.3)	63.7 (5.3)	<.001
Sex				<.001			
Male	108,214 (51.9%)	76,684 (49.5%)	31,530 (59.0%)		990 (1.3%)	1,147 (3.6%)	<.001
Female	100, 138 (48.1%)	78,216 (50.5%)	21,922 (41.0%)		646 (0.8%)	778 (3.6%)	<.001
Race/ethnicity				<.001			
White	180, 484 (86.6%)	132,582 (85.6%)	47,902 (89.6%)		1,377 (1.0%)	1,739 (3.6%)	<.001
Black	10,069 (4.8%)	7,708 (4.9%)	2,361 (4.4%)		128 (1.7%)	98 (4.2%)	<.001
Hispanic	3,480 (1.7%)	2,818 (1.8%)	662 (1.2%)		20 (0.7%)	11 (1.7%)	.034
Asian	6,671 (3.2%)	5,576 (3.6%)	1,095 (2.1%)		32 (0.6%)	32 (2.9%)	<.001
Native Americans	579 (0.3%)	389 (0.3%)	190 (0.4%)		5 (1.3%)	9 (4.7%)	.018
Pacific Islanders	1,028 (0.5%)	835 (0.5%)	193 (0.4%)		10 (1.2%)	5 (2.6%)	.176
Unknown	6,041 (2.9%)	4,992 (3.2%)	1,049 (2.0%)		64 (1.3%)	31 (3.0%)	<.001
Education level				<.001			
Less than high school	14,330 (7.1%)	11,081 (7.2%)	3,249 (6.1%)		229 (2.1%)	172 (5.3%)	<.001
High school graduate	47,106 (23.3%)	34,394 (22.2%)	12,712 (23.8%)		432 (1.3%)	545 (4.3%)	<.001
Post high school training	26,261 (13.0%)	18,827 (12.2%)	7,434 (13.9%)		197 (1.1%)	285 (3.8%)	<.001
Some college	44,970 (22.2%)	32,693 (21.1%)	12,277 (23.0%)		348 (1.1%)	427 (3.5%)	<.001
College graduate	34,290 (16.9%)	25,344 (16.4%)	8,946 (16.8%)		200 (0.8%)	247 (2.8%)	<.001
Postgraduate	34, 830 (17.2%)	27,230 (17.6%)	7,600 (14.2%)		160 (0.6%)	203 (2.7%)	<.001
Other	966 (0.5%)	0	966 (1.8%)		64 (1.3%)	31 (3.2%)	<.001
Family history of lung cancer	27,532 (13.2%)	15,911 (10.3%)	11,621 (21.7%)	<.001	284 (1.8%)	497 (4.3%)	<.001
No	169,726 (81.5%)	128,786 (83.1%)	40,940 (76.6%)		1,195 (0.9%)	1,386 (3.4%)	<.001
Body mass index, Kg/m^2	27.5 (5.0)	27.3 (4.9)	28.0 (5.1)	<.001	26.5 (4.4)	26.9 (4.7)	.002
Diagnosed with COPD‖	19,149 (9.2%)	9,823 (6.3%)	9,326 (17.5%)	<.001	297 (3.0%)	517 (5.5%)	<.001
No	183,020 (87.8%)	139,089 (89.8%)	43,931 (82.2%)		1,258 (0.9%)	1,399 (3.2%)	<.001
Cigarette smoking status				<.001			
Never	69,183 (33.2%)	69,183 (44.7%)	0		110 (0.2%)	0	
Former	92,383 (44.3%)	64,691 (41.8%)	27,692 (51.8%)		807 (1.3%)	766 (2.8%)	<.001
Current	41,817 (20.1%)	16,057 (10.4%)	25,760 (48.2%)		655 (4.1%)	1,159 (4.5%)	.041
Smoking intensity, cigarettes/day ¶	26.2 (13.6)	24.7 (14.6)	28.4 (11.5)	<.001	29.9 (15.3)	29.6 (11.7)	.447
Smoking duration, years ¶	32.6 (13.1)	27.7 (13.8)	39.8 (7.3)	<.001	40.0 (10.7)	44.2 (7.0)	<.001
Smoking quit time, years **	16.3 (12.0)	20.2 (12.0)	7.3 (4.8)	<.001	12.5 (10.6)	6.6 (4.8)	<.001

Abbreviations: BMI: Body mass index; CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; N: Number; NLST: National Lung Screening Trial; PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.

* Percentages for 'all' are out of the total N for that column. Percentages may not add up to 100% due to missing data which is not shown here.

† P-value for two sample t-test/Mann Whitney U test or Chi-sq./ Fisher's exact test (PLCO vs. NLST for the corresponding variable).

‡ P-value for two sample t-test/Mann Whitney U test or Chi-sq./ Fisher's exact test (the difference in cumulative incidence between screening trials for that row).

§ Percentages are row percentages, i.e. the percentages of participants for that row who developed lung cancer during follow-up.

‖ Diagnosis of COPD encompasses diagnoses of emphysema or chronic bronchitis; ¶ among ever-smokers; ** among former smokers.

Table 4.2 - Personal cancer history characteristics for study participants (N=208,352) by trial and lung cancer diagnosis

		Screening Trial			Developed Lung Cancer During Follow-up		
Personal cancer history variable	Combined (N = 208,352)	PLCO (N = 154,900)	NLST (N = 53,452)		PLCO (n = 3,592)	NLST (n = 2,058)	
	N (%)*	**N (%)***	**N (%)***	**P-value†**	**N (%)§**	**N (%)§**	**P-value‡**
Personal history of any cancer	9,205 (4.4%)	6,897 (4.5%)	2,308 (4.3%)	.007	112 (1.6%)	126 (5.5%)	<.001
No	194,193 (93.2%)	143,049 (92.4%)	51,144 (95.7%)		1,460 (1.0%)	1,799 (3.5%)	<.001
Number of previous cancers				.001			
1	8,699 (4.2%)	6,491 (4.2%)	2,208 (4.1%)		105 (1.6%)	120 (5.4%)	<.001
2+	506 (0.2%)	406 (0.3%)	100 (0.2%)		7 (1.7%)	6 (6.0%)	.027
Previous cancer age category				<.001			
60+	1,925 (0.9%)	1,606 (1.0%)	319 (0.6%)		31 (1.9%)	36 (11.3%)	<.001
< 60	7,088 (3.4%)	5,123 (3.3%)	1,965 (3.7%)		77 (1.5%)	86 (4.4%)	<.001
Breast‖	3,441 (3.4%)	2,702 (3.5%)	739 (3.4%)	.617	29 (1.1%)	39 (5.3%)	<.001
No					617 (0.8%)	735 (3.5%)	<.001
Digestive							
Colorectal	254 (0.1%)	24 (<0.1%)	230 (0.4%)	<.001	0	9 (3.9%)	1.00
No					1,636 (1.1%)	1,907 (3.6%)	<.001
Esophageal	48 (<0.1%)	28 (<0.1%)	20 (<0.1%)	.011	0	1 (5.0%)	.417
No					1,636 (1.1%)	1,915 (3.6%)	<.001
Pancreatic	25 (<0.1%)	18 (<0.1%)	7 (<0.1%)	.782	0	0	
No					1,636 (1.1%)	1,916 (3.6%)	<.001
Stomach	82 (<0.1%)	55 (<0.1%)	27 (0.1%)	.128	0	1 (3.7%)	.329
No					1,636 (1.1%)	1,915 (3.6%)	<.001
Endocrine							
Thyroid	425 (0.2%)	342 (0.2%)	83 (0.2%)	.004	1 (0.3%)	3 (3.6%)	.025
No					1,635 (1.1%)	1,913 (3.6%)	<.001
Genitourinary							
Bladder	590 (0.3%)	323 (0.2%)	267 (0.5%)	<.001	19 (7.1%)	6 (1.9%)	<.001
No					1,906 (3.6%)	1,630 (1.1%)	.001
Kidney	282 (0.1%)	214 (0.1%)	68 (0.1%)	.571	3 (1.4%)	4 (5.9%)	.060
No					1,633 (1.1%)	1,912 (3.6%)	<.001
Prostate¶	35 (<0.1%)	35 (<0.1%)	N/A		1 (2.9%)	N/A	
No					989 (1.3%)	N/A	
Testicular¶	78 (0.1%)	78 (0.1%)	N/A		0	N/A	
No					990 (1.3%)	N/A	
Gynecologic‖							
Cervical	1,365 (1.4%)	600 (0.8%)	765 (3.5%)	<.001	10 (1.7%)	30 (3.9%)	.014
No					636 (0.8%)	744 (3.5%)	<.001
Uterine	532 (0.7%)	532 (0.7%)	N/A		11 (2.1%)	N/A	
No					635 (0.8%)	N/A	
Head and neck							
Laryngeal	138 (0.1%)	92 (0.1%)	46 (0.1%)	.037	4 (4.4%)	2 (4.4%)	.654
No					1,632 (1.1%)	1,914 (3.6%)	<.001
Nasopharyngeal	41 (<0.1%)	12 (<0.1%)	29 (0.1%)	<.001	1 (8.3%)	2 (6.9%)	.872
No					1,635 (1.1%)	1,923 (3.6%)	<.001
Oral	103 (0.1%)	N/A	103 (0.2%)		N/A	13 (12.6%)	
No					N/A	1,903 (3.6%)	
Hematologic							
Hodgkin lymphoma	73 (<0.1%)	73 (0.1%)	N/A		1 (1.4%)	N/A	
No					1,635 (1.1%)	N/A	
Non-Hodgkin lymphoma	215 (0.1%)	215 (0.1%)	N/A		4 (1.9%)	N/A	
No					1,632 (1.1%)	N/A	
Lung	27 (<0.1%)	7 (<0.1%)	20 (<0.1%)	<.001	2 (28.6%)	10 (50.0%)	.326
No					1,634 (1.1%)	1,906 (3.6%)	<.001

Abbreviations: BMI: Body mass index; CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; N: Number; N/A Not present in dataset for participants of that screening trial; NLST: National Lung Screening Trial; PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.

* Percentages for 'all' are out of the total N for that column. Percentages may not add up to 100% due to missing data which are not shown here.

† P-value for two sample t-test/Mann Whitney U test or Chi-sq. / Fisher's exact test (PLCO vs. NLST for the corresponding variable).

‡ P-value for two sample t-test/Mann Whitney U test or Chi-sq. / Fisher's exact test (the difference in cumulative incidence between screening trials for that row).

§ Percentages are row percentages, i.e. the percentages of participants for that row who developed lung cancer during follow-up.

‖ ¶ Percentages presented for females (‖) or males (¶) only.

Overall characteristics of participants in PLCO and NLST pooled data

A personal history of any cancer was reported by 4.4% of the pooled samples, the overwhelming majority of whom (94.5%) had only one previous cancer. The most common previous cancer type reported among both trials combined was female breast cancer (n=3,441), representing 44.4% of all previous cancers for which the specific site was available (n=7,754). A total of 506 (0.2%) people had two or more previous cancers. Of those with a personal history of cancer and for whom the age at diagnosis was reported, 78.6% reported at least one cancer diagnosis before the age of 60. SPLC developed in 4.0% of the 9,205 cancer survivors in the study and 4.0% of ever-smoking cancer survivors developed SPLC compared to only 0.4% of never-smokers.

4.1 Participant characteristics

Comparing PLCO and NLST participants

The mean age in each trial was similar: 62.6 for the PLCO group and 61.4 for the NLST participants. A higher proportion of the NLST participants were male compared to the PLCO trial (59.0% vs. 49.5%). In both cohorts, the majority of the patients were white (85.6% and 89.6% for PLCO and NLST participants respectively) and the distribution of education levels was similar between both trials. Most (69.6% of PLCO and 68.1% of NLST participants) had some form of post high school education. A higher proportion of NLST participants had been diagnosed with COPD (17.5% vs. 6.3% among PLCO patients). Compared to the PLCO group, the NLST had a higher proportion of current smokers (48.2% vs. 10.4%). NLST ever-smokers, on average, smoked more cigarettes per day (28.4 vs. 24.7) for a greater number of years (39.8 vs. 27.7) but the mean quit

time among former smokers was higher in the PLCO (20.2 vs. 7.3 years). All contrasts for these characteristics were statistically significant ($p < .001$).

In terms of previous cancer, a similar proportion of participants in each trial had been previously diagnosed with at least one type of cancer (4.5% of PLCO participants and 4.3% of the NLST cohort) though the difference was statistically significant ($p = .007$). The most common types of previous cancer reported were breast cancer (3.5% and 3.4% in the females in the PLCO and NLST trials, respectively) and cervical cancer, which was statistically significantly higher ($p < .001$) in the NLST trial (3.5%) compared to the PLCO (0.8%). Previous bladder cancers were more than twice as frequent in the NLST (0.5% vs 0.2%; $p < .001$) compared to the PLCO and colorectal cancers were more frequent in the NLST than in the PLCO (0.4% vs. <0.1%; $p < .001$). The latter finding is expected since previous colorectal cancer was an exclusion criterion for the PLCO, but some people who reported having it were inadvertently included in the trial. In general, absolute differences in proportions of previous cancer types were generally very small. Lung cancer survivors were supposed to be excluded from both screening trials so the small numbers that were included (n=7 and n=20 for the PLCO and NLST, respectively) may not be representative of all participants who would otherwise have been eligible. The NLST contained a higher proportion of lung cancer survivors but in absolute terms the number was still small (n=20). A similar proportion of PLCO and NLST participants were previously diagnosed with cancer before the age of 60 (3.3% vs 3.7%, respectively). All contrasts for these characteristics were statistically significant ($p < .001$).

Impact of participant characteristics on cumulative incidence of lung cancer.

In the PLCO, the cumulative lung cancer incidence was higher among males than females, 1.3% vs. 0.8%, respectively, but no difference between men and women was seen for NLST participants (3.6% for both sexes). In the PLCO, lung cancer incidence was highest for Blacks (1.7%) whereas in the NLST, Native Americans showed the highest incidence (4.7%). The cumulative lung cancer incidence was statistically different between the PLCO and NLST for each race/ethnicity except Pacific Islanders ($p = .176$). In both trials, those who did not have at least a high school education had the highest incidence of lung cancer (2.1% and 5.3% in the PLCO and NLST respectively). Those with a family history of lung cancer in the PLCO had a 100% increased risk of lung cancer relative to those with no family history of the disease. The increase in lung cancer risk among NLST participants with a family history of lung cancer was 26%. Participants in the PLCO with COPD had a 233% increased risk of lung cancer compared to those without COPD. Among NLST participants, the increase was 72% between those with and without COPD. Current smokers in the PLCO had 20 times the risk of lung cancer compared to never-smokers (cumulative incidence of 9.0% vs. 0.4%).

Having a personal history of cancer increased the risk of lung cancer by 60% and 57% in PLCO and NLST participants, respectively. Being previously diagnosed with two or more previous cancers increased the risk of lung cancer, compared to being previously diagnosed with only one, by 6% for PLCO participants and 11% for those in the NLST. In both trials, lung cancer incidence was higher in those previously diagnosed with cancer at age 60 or above, compared to those diagnosed before age 60. Compared to those not previously diagnosed with the disease, PLCO lung cancer survivors had 26 times the risk of SPLC and lung cancer survivors in the NLST had 14 times the risk of SPLC. The

second and third largest effects were seen for survivors of nasopharyngeal and laryngeal cancers in the PLCO with relative risks for SPLC of 7.5 and 4.0, respectively, compared to having no history of that type of cancer.

4.2 Key study findings

Are any specific types of previous cancers strong independent predictors of subsequent lung cancer?

Previous cancers at several different tumour sites were associated with lung cancer diagnosis during follow-up. Results from several separate logistic regression models, one for each previous cancer type, are presented in Table 4.3.

Table 4.3 – Odds ratios for future lung cancer by previous cancer type (relative to no previous cancer of that type) for PLCO (N=154,900) and NLST (n=53,452) participants

	PLCO			NLST		
		Unadjusted	Adjusted *		Unadjusted	Adjusted *
Previous cancer type	**N**	**Odds Ratio (95% CI); P-value**		**N**	**Odds Ratio (95% CI); P-value**	
Breast†	2,702	4.03 (0.54 – 30.1); .175	1.22 (0.70 – 2.13); .480	739	1.54 (1.11 – 2.15); .010	1.25 (0.88 – 1.76); .213
Digestive						
Colorectal	24	3.92 (0.53 – 29.0); .181§	2.94 (0.23 – 37.0); .404§	230	1.09 (0.55 – .2.13); <.001	0.77 (0.27 – 2.15); .623
Esophageal	28	3.36 (0.46 – 24.7); .234§	1.82 (0.21 – 15.8); .588§	20	1.41 (0.19 – 10.5); .739	1.09 (0.17 – 7.10); .931
Pancreatic	18	-‖	-‖	7	-‖	-‖
Stomach	55	1.71 (0.24 – 12.4); .596§	1.06 (0.14 – 8.02)§	27	1.03 (0.14 – 7.59); .977	0.71 (0.10 – 5.19); .734
Endocrine						
Thyroid	342	0.28 (0.04 – 1.97); .200	0.46 (0.08 – 2.85); .406	83	1.00 (0.32 – 3.18); .995	1.11 (0.39 – 3.23); .842
Genitourinary						
Bladder	323	1.79 (0.80 – 4.01); .160	0.96 (0.57 – 1.61); .875	272	2.06 (1.29 – 3.29); .002	1.43 (0.86 – 2.40); .169
Kidney	214	1.35 (0.43 – 4.22); .607	0.93 (0.35 – 2.49); .884	68	1.67 (0.61 – 4.60); .318	1.22 (0.50 – 2.98); .663
Prostate‡	35	2.32 (0.32 – 16.96); .408	2.88 (0.21 – 39.1); .426§	N/A	N/A	N/A
Testicular‡§	78	0.98 (0.14 – 7.05); .984§	1.04 (0.16 – 6.72); .969§	N/A	N/A	N/A
Gynecologic†						
Cervical	600	2.08 (1.11 – 3.90); .023	1.26 (0.82 – 1.95);.291	765	1.11 (0.77 – 1.62); .567	0.96 (0.68 – 1.37); .840
Uterine	532	2.58 (1.41 – 4.71); .002	2.01 (1.09 – 3.71); .026	N/A	N/A	N/A
Head and neck						
Laryngeal	92	4.28 (1.57 – 11.67); .005	1.36 (0.58 – 3.18); .481	46	1.22 (0.29 – 5.02); .786	0.73 (0.20 – 2.68); .634
Nasopharyngeal	12	8.54 (1.10 – 66.22); .040	7.62 (0.35 – 166.20); .197	29	1.98 (0.47 – 8.35); .350	1.11 (0.89 – 2.39); .169
Oral	N/A	N/A	N/A	103	3.88 (2.17 – 6.96); <.001	3.02 (1.41 – 6.45); .004
Hematologic						
Hodgkin lymphoma	73	1.32 (0.18 – 9.53); .781	1.39 (0.17 – 11.64); .762	N/A	N/A	N/A
Non-Hodgkin lymphoma	215	1.78 (0.66 – 4.80); .253	1.05 (0.14 – 8.11) .962	N/A	N/A	N/A
Lung	7	37.63 (7.30 – 194.09); <.001	17.39 (3.14 – 96.13); <.001	20	26.89 (11.18 – 64.69); <.001	35.12 (10.8 – 114.6); <.001
Any	6,897	1.60 (1.32 – 1.95); <.001	1.39 (1.13 – 1.72); .002	2,308	1.58 (1.32 – 1.91); <.001	1.32 (1.08 – 1.61); .006

Abbreviations: BMI: Body mass index; CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; N: Number; N/A – Association could be not calculated as data for this cancer were not present; NLST: National Lung Screening Trial; PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.

* Adjusted for age, race/ethnicity, education level, BMI, COPD diagnosis, family history of lung cancer, smoking status, intensity and duration and quit time. Robust standard errors were used to account for correlations resulting from cluster sampling by study centres.

† ‡ Estimates calculated among males only (‡ prostate, testicular; n=108,214) or females only († breast, cervical, uterine, n= 100,138).

§ A single participant who developed lung cancer but did not report having this kind of previous cancer was treated as if they did have it to avoid complete separation of odds ratio estimation.

‖ Risk estimates could not be obtained for survivors of pancreatic cancer because of problems with collinearity.

Previous cancer types among PLCO participants

Breast cancer was the most common type of previous cancer reported among PLCO participants (n=2,702), followed by cervical cancer (n=600). Other frequent previous cancer types were uterine (n=532), thyroid (n=342) and bladder cancers (n=323). The largest clinically relevant effect was observed, after adjustment, for a history of previous nasopharyngeal (OR = 7.62; 95% CI: 0.35 – 166.20). Prostate cancer survivors had an almost three-fold increased risk of SPLC but the small sample size (n=35) meant the confidence interval for the estimate was imprecise (OR = 2.88; 95% CI: 0.21 – 39.1). Despite being an exclusion criterion for the PLCO trial, previous lung cancer was reported among some participants and was strongly associated with subsequent primary lung cancer (SPLC) (OR = 17.39; 95% CI: 3.14 – 96.13) though an estimate from such a small number of observations (n=7) should be interpreted with caution. In descending order of magnitude, other notable effects were seen for survivors of colorectal (OR = 2.94; 95% CI: 0.23 – 37.0), uterine (OR = 2.01; 95% CI: 1.09 – 3.71) and esophageal carcinomas (OR = 1.82; 95% CI: 0.21 – 15.8), Hodgkin lymphoma (OR = 1.39; 95% CI: 0.17 – 11.64) and laryngeal cancer (OR = 1.36; 95% CI: 0.58 – 3.18). Women with cervical (OR = 1.26; 95% CI: 0.82 – 1.95) or breast cancer (OR = 1.22; 95% CI: 0.70 – 2.13) had elevated odds of developing lung cancer, however neither association was statistically significant.

Previous cancer types among NLST participants

Cervical cancer was also the most common type of previous cancer reported among NLST participants (n=765), closely followed by breast cancer (n=739). Other notable

previous cancer types were bladder (n=272) and colorectal (n=230). The strong association between previous lung cancer and SPLC among PLCO participants was replicated in the NLST cohort (OR 35.12; 95% CI: 10.8 – 114.6). The second strongest association for NLST participants was observed for those who reported having previous oral cancer, for whom the odds of SPLC was three times as high relative to those who had never had oral cancer oral (OR = 3.02; 95% CI: 1.41 – 6.45). In descending order of magnitude, an increased SPLC risk was also seen following previous bladder cancer (OR 1.43; 95% CI: 0.86 – 2.40), breast cancer (OR = 1.25; 95% CI: 0.88 – 1.76), kidney (OR 1.22; 95% CI: 0.50 – 2.98) and esophageal cancer (OR = 1.09; 95% CI: 0.17 – 7.10). Information on previous oral cancer was only available for participants of the NLST so this association could not be tested among PLCO participants.

Is the number of previous cancers with which a person has been diagnosed an important predictor of future lung cancer?

The results for logistic regression models to assess the association between the number of previous cancers and subsequent lung cancer risk are shown in Table 4.4.

Table 4.4 - Odds ratios for SPLC by number of previous cancers (relative to none) for PLCO (N=154,900) and NLST (N=53,452) participants

	PLCO			NLST		
		Unadjusted	Adjusted*		Unadjusted	Adjusted*
Number of previous cancers	**N**	**Odds Ratio (95% CI); P-value**		**N**	**Odds Ratio (95% CI); P-value**	
1	6,491	1.60 (1.31 – 1.95); <.001	1.40 (1.14 – 1.71); .001	2,208	1.58 (1.30 – 1.91); <.001	1.32 (1.09 – 1.60); .005
2+	406	1.70 (0.80 – 3.60); .164	1.35 (0.57 – 3.22); .498	100	1.75 (0.77 – 4.00); .184	1.10 (0.42 – 2.88); .849

Abbreviations: BMI: Body mass index; CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; N: Number; NLST: National Lung Screening Trial; PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; SPLC: Subsequent primary lung cancer.
* Adjusted for age, race/ethnicity, education level, BMI, COPD diagnosis, family history of lung cancer, smoking status, intensity and duration and quit time. Robust standard errors were used to account for correlations resulting from cluster sampling by study centres.

The unadjusted effect of having a previous cancer on SPLC risk was higher, in both trials, for those with more than one previous cancer than with just one. However, the

effect estimates for the latter group were imprecise (wide confidence intervals) and the increase in effect size compared to having only had one previous cancer did not persist after adjustment. Among PLCO participants, in adjusted analysis relative to those with no previous cancer the odds ratio for SPLC for those diagnosed with two or more previous cancers was 1.35 (95% CI: 0.57 – 3.22) and for those diagnosed with only one previous cancer was 1.40 (95% CI: 1.14 – 1.71). A post-estimation test of equality between the two categories showed this difference to be statistically insignificant (χ^2 = 0.01, p = 0.934). The odds of developing lung cancer relative to those with no previous cancer was lower for NLST participants with two or more previous cancers (OR = 1.10; 95% CI: 0.42 – 2.88) than for those with only one (OR = 1.32; 95% CI: 1.09 – 1.60) but again the difference was nonsignificant (χ^2 = 0.16, p = 0.692). In summary, the effect of a personal history of cancer as a risk factor for future lung cancer was not more pronounced in those with more than one previous cancer compared with having only one.

Is being diagnosed with cancer before age 60 years associated with future lung cancer risk more strongly than being diagnosed at 60 years or older?

Risk of lung cancer in those with a personal cancer history compared to those without, stratified by age at diagnosis category, are presented in Table 4.5.

Table 4.5 - Odds ratios for lung cancer (relative to having never had cancer) by age of previous cancer diagnosis for PLCO (n=154,900) and NLST (n=53,452) participants

	PLCO			NLST		
		Unadjusted	Adjusted*		Unadjusted	Adjusted*
Age of previous cancer onset	**N**	**Odds Ratio (95% CI); P-value**		**N**	**Odds Ratio (95% CI); P-value**	
60+	1,606	1.91 (1.34 – 2.74); <.001	1.20 (1.00 – 1.46); .054	319	3.49 (2.46 – 4.95); <.001	1.80 (1.09 – 2.98); .021
<60	5,123	1.48 (1.18 – 1.87); .001	1.45 (1.07 – 1.96); .016	1,965	1.26 (1.00 – 1.57); .043	1.15 (0.97 – 1.37); .107

Abbreviations: BMI: Body mass index; CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; N: Number; NLST: National Lung Screening Trial. PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; SPLC: Subsequent primary lung cancer.
* Adjusted for age, race/ethnicity, education level, BMI, COPD diagnosis, family history of lung cancer, smoking status, intensity and duration, quit time. Robust standard errors were used to account for correlations resulting from cluster sampling by study centres.

For PLCO participants, the odds ratio for SPLC compared to those with no previous cancer was higher among those diagnosed before age 60, compared to those diagnosed at 60 or above (OR = 1.45; 95% CI: 1.07 – 1.96 vs. 1.20; 95% CI: 1.00 – 1.46) after adjustment for known risk factors, but the difference was not statistically significant (χ^2 = 1.18, p = 0.277). Among NLST participants, the OR for SPLC, relative to those with no previous cancer, was *lower* in those diagnosed before age 60 (1.15; 95% CI: 0.97 – 1.37) compared to those diagnosed at 60 or older (1.80; 95% CI: 1.09 – 2.98); the difference was not statistically significant (χ^2 = 3.43, p = 0.064). In summary, the association between a personal history of cancer and increased future lung cancer risk did not appear to be stronger in those diagnosed before age 60, compared to those diagnosed at 60 or above.

Do any of these associations lead to improved lung cancer risk prediction models?

149,247 participants from the PLCO trial and 53,202 from the NLST were eligible for inclusion into risk prediction modeling, when taking the personal history of cancer criterion into consideration. Multivariable fractional polynomial analysis of the PLCOall2014 model applied to all PLCO participants indicated that smoking intensity and the length of time since former smokers had quit smoking both had non-linear relationships with 6-year lung cancer risk (Figures 4.1 and 4.2). Suggested transformations for both variables were included in all further models as 'PLCOall2014mfp'. An interaction term between quit time and former smoker status

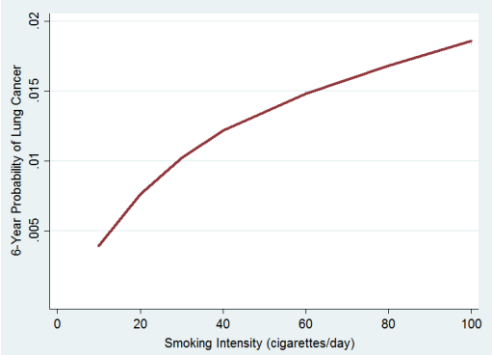

Figure 4.2 Non-linear relationship between smoking quit time and lung cancer risk among former smokers in the PLCO

Probabilities were based on an age of 62 years, white race/ethnicity, some college education, a body-mass index of 27 kg/m^2, no history of COPD/emphysema/chronic bronchitis, neither personal nor family history of lung cancer, former smoker status, 27 year smoking history and an average smoking intensity of 20 cigarettes/day.

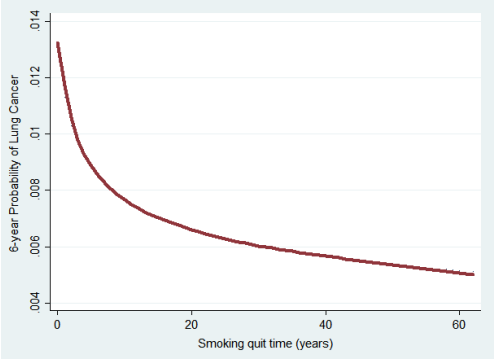

Figure 4.1 Non-linear relationship between smoking intensity and lung cancer risk among PLCO smokers

Probabilities were based on an age of 62 years, white race/ethnicity, some college education, a body-mass index of 27 kg/m^2, no history of COPD/emphysema/chronic bronchitis, neither personal nor family history of cancer, former smoker status, 27 year smoking history and 10 years quit time.

(0/1) was used to incorporate quit time into the model such that the contribution of quit time would only be taken into account for former smokers.

A test of the PLCOall2014 model applied to all PLCO participants was statistically significant, indicating that the set of predictors reliably discriminated between those who developed lung cancer and those who did not (-2LL = -6530,0294, χ^2 (16) = 2705.37, p <0.0001). The model demonstrated excellent discrimination among PLCO participants (AUC = 0.8552, bias-corrected 95% CI: 0.8450 – 0.8643). The three refined measures of a personal history of cancer described in Section 3.3 were substituted for the original variable and discrimination was re-tested separately for each candidate. A consistent sample size was maintained for each model by ascertaining which one included the fewest observations (N=139,053) and re-running all models using only these participants. The discrimination performance for these models is summarized in Table 4.6.

Table 4.6 Discrimination performance for logistic 6-year lung cancer models applied to PLCO participants (N=139,053 for all models).

Model +/- covariates	AUC (95% CI)*
PLCOall2014mfp†	**0.8552 (0.8450 – 0.8643)**
PLCOall2014mfp - personal history of cancer (0/1)	0.8545 (0.8443 – 0.8636)
+ personal history of strongly-associated cancers ‡ (0/1)	0.8548 (0.8447 – 0.8640)
+ personal history of head and neck cancer § (0/1)	0.8546 (0.8441 – 0.8637)
+ personal history of smoking-related cancer ‖ (0/1)	0.8546 (0.8444 – 0.8637)
PLCOall2014mfp†	
+ number of previous cancers (none, 1, 2+)	0.8552 (0.8450 – 0.8642)
+ age at previous cancer diagnosis (N/A, 60+,<60)	0.8551 (0.8448 – 0.8641)

Abbreviations: AUC: Area under the curve; CI: Confidence interval; PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.

* Bootstrap bias-corrected.

† Includes non-linear transformations for smoking intensity and smoking status-quit time interaction term.

‡ Breast, prostate, uterine, non-Hodgkin lymphoma or lung.

§ Nasopharyngeal, laryngeal or oral cancer.

‖ Lung, nasopharyngeal, oral, laryngeal, pancreatic, esophageal, kidney, stomach, bladder, cervical or colorectal.

After removing the variable representing a personal history of cancer at baseline, the AUC dropped to 0.8545 (95% CI: 0.8443 – 0.8636). Adding the variable comprising cancer types that showed a strong association with SPLC in adjusted analyses (breast, prostate, non-Hodgkin lymphoma and lung) only raised the AUC by .0003 (0.8548; 95% CI: 0.8447 – 0.8640). Even smaller increases in AUC were seen when substituting a personal history of head and neck cancers (0.8546; 95% CI: 0.8444 – 0.8637) or a history of smoking-related cancer into the model (0.8546; 95% CI: 0.8444 – 0.8637). None of the three refined binary measures of personal cancer history improved model discrimination beyond what was achieved using the original term.

The number of previous cancers (none, 1, 2+) reported did not improve discrimination (AUC = 0.8552; 95% CI: 0.8450 – 0.8642), nor did the age at previous cancer diagnosis category (AUC = 0.8551; 95% CI: 0.8448 – 0.8641). No evidence of interaction between a personal history of cancer and age of previous cancer onset (60+/<60 years) was observed when included alongside the PLCOall2014 covariates (p = 0.787). The PLCOall2014mfp model was re-evaluated among NLST participants as a

means of external validation. Performance statistics for each case are summarized in Table 4.7.

Table 4.7 - Performance statistics for the PLCOall2014mfp model by screening trial.

Statistic	PLCO (N=139,212)	NLST (N=51,033)
AUC (95% CI)*	0.855 (0.845 – 0.864)	0.710 (0.698 – 0.721)
Brier score (95% CI)*	0.0097 (0.0092 – 0.0102)	0.0338 (0.0324 – 0.0353)
Spiegelhalter's z-statistic (p-value)	0.2437 (0.4037)	0.1984 (0.4214)
Median absolute error	0.0026	0.0275
90th percentile absolute error	0.0289	0.0808

Abbreviations: AUC: Area under the curve. CI: Confidence interval; N: Number; NLST: National Lung Screening Trial; PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.
* Bias-corrected using bootstrapping of 1000 samples.

Discrimination was considerably lower when the model was applied to the NLST (AUC = 0.710). The performance as a whole, as measured by the Brier score, showed a similar pattern. The score for the NLST was higher (0.0338; 95% CI: 0.0324 – 0.0353), indicating worse performance compared to the PLCO (0.0097; 95% CI: 0.0092 – 0.0102). In terms of calibration, the PLCOall2014mfp model achieved a non-significant result for the Spigelhalter z-statistic in all cases, indicating a good model calibration in the PLCO and NLST. Median and 90th percentile absolute error between observed and predicted probabilities were also less favourable when assessed in NLST participants-only, though the values were still acceptably low (0.0275 and 0.0808, respectively) and were no cause for concern. Bootstrap bias-corrected confidence intervals for both the AUC and Brier score from the final model were relatively narrow, suggesting that the model was internally valid. In terms of external validity though, the excellent performance of the model was not matched when applied to the NLST. The reason for this is that the NLST consists only of high-risk ever-smokers, increasing the homogeneity in risk across all individuals compared to a general population-based sample containing both ever- and

never-smokers. Increased homogeneity makes it more difficult to discriminate between those who will get lung cancer (higher risk) and those who will not (lower risk).

The odds ratios and coefficients for the variables included in the final logistic regression model for 6-year lung cancer risk among PLCO participants is summarized in Table 4.8. 139,212 observations were used in the final model of a possible 154,900 PLCO participants, thus 10.1% of the sample did not have sufficient data for this set of predictors. While Bennet maintains that estimates are likely to be biased when more than 10% of data are missing (Bennett, 2001), 10.1% is on the edge of that threshold and a simulation study investigating the impact of missing data on secondary analyses of large surveys found that results with 10% missing were not biased to any major degree. Only at values of 20% or higher did missing data impact their results (Langkamp, Lehman, & Lemeshow, 2010). Therefore, imputation of missing data was not undertaken. Finally, there were 1,572 occurrences of lung cancer in the PLCO during six years of follow-up. This equates to 92 per each of the 17 predictors present in the model, which is higher than the recommended minimum of 50 per predictor to minimize overfitting.

Table 4.8 - Logistic regression model for 6-year lung cancer risk among PLCO cancer screening trial participants (N=136,212)

Variable	Odds Ratio (95% CI)	P-value	Beta Coefficient
Age, per 1-yr increase†	1.071 (1.057 – 1.085)	<0.001	0.068348
Race or ethnicity‡			
White	1.000		Reference group
Black	1.587 (1.294 – 1.946)	0.001	0.461738
Hispanic	0.679 (0.405 – 1.138)	0.142	-0.38783
Asian	0.651 (0.435 – 0.973)	0.036	-0.42949
Native American	1.242 (0.504 – 3.065)	0.638	0.216873
Pacific Islanders	1.117 (0.548 – 2.28)	0.761	0.110893
Education, per 1 level increase†§	0.929 (0.897 – 0.962)	<0.001	-0.07321
Body-mass index, per 1 kg/m^2 increase†	0.968 (0.956 – 0.980)	<0.001	-0.0328
Chronic obstructive pulmonary disease (yes vs. no)‖	1.406 (1.220 – 1.619)	<0.001	0.340466
Personal history of cancer (yes vs. no)	1.393 (1.128 – 1.719)	0.002	0.331297
Family history of lung cancer (yes vs. no)	1.716 (1.498 – 1.965)	<0.001	0.53973
Former smoker (yes vs. no)	17.154 (8.687 – 33.871}	<0.001	2.842204
Current smoker (yes vs. no)	16.666 (9.176 – 30.272}	<0.001	2.813396
Smoking intensity (cigarettes/day) ¶		<0.001	-0.7168
Smoking duration, per 1-yr increase†	1.044 (1.034 – 1.054)	<0.001	0.043233
Smoking quit time, per 1-yr increase†¶		<.0001	-0.23175
Model constant			-7.51463

Abbreviations: CI: Confidence interval; COPD: Chronic obstructive pulmonary disease; PLCO: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial.

* To calculate the 6-year probability of lung cancer for an individual, do the following: for categorical variables, multiply the beta coefficient for the variable by 1 if the person falls into that category and 0 if not. For variables other than smoking intensity and quit time, subtract the centering value from the individual's value for that factor and multiply the result by the beta coefficient of the variable. For smoking intensity, divide the person's value by 100, exponentiate it to the power −0.5, and subtract 2.010432048, then multiplying this number by the beta coefficient of the variable. For quit time duration, add 1 to the person's value and divide the total by 10, then take the natural log of this number and add .0407789134 to center the value. Multiplying the result by the smoking quit time beta coefficient. Sum together each of these components for each predictor as well as the model constant to obtain the logit. Finally, convert the logit to a probability using the formula: $e^{logit}/(1+ e^{logit})$.

† Age was centered on 62 years, education was centered on level 4, body-mass index was centered on 27 kg/m^2, smoking duration was centered on 27 years.

‡ As self-reported.

§ Education level consisted of ordinal groups: less than high-school graduate (1), high-school graduate (2), some training after high school (3), some college (4), college graduate (5), and postgraduate or professional degree (6).

‖ COPD variable also encompasses diagnoses of emphysema and chronic bronchitis.

¶ Smoking intensity (average number of cigarettes smoked per day) and smoking quit time (duration in years) had nonlinear associations with lung cancer, and these variables were transformed. For this reason, meaningful interpretations of their odds ratios directly are not possible.

Overall, considerable increases in SPLC risk were observed for those with previous lung cancer and female breast cancer survivors in both trials. Marked elevations in risk were observed for prostate and uterine cancer as well as non-Hodgkin lymphoma survivors but these effects were only estimable for PLCO. Similarly, the increased SPLC observed in NLST participants with previous oral cancer could not be re-evaluated in

PLCO participants. Only the associations with previous lung cancer (both trials), oral cancer (NLST only) and uterine cancer (PLCO) were statistically significant.

The number of previous cancers was not an important factor in future lung cancer risk, nor was being diagnosed with any cancer before age 60 compared with at 60 or above. More refined variables representing personal cancer history did not improve lung cancer risk prediction by a considerable amount when compared to the simple binary 'personal history of cancer (yes/no)' predictor present in the most current lung cancer risk prediction model. A detailed discussion of these findings in context will be presented in the following chapter.

4.3 Assumptions Testing

In order to test for misspecification error for the final model, a specification link test was performed following model estimation. The _hatsq term was statistically significant ($p = 0.001$) which suggests that important predictors or interaction effects might be missing from the model or the logit function is incorrectly specified. However, the significant result might be due to the large sample size used and excellent discrimination and calibration performance suggest that assumption of correct model fit holds for the model. Non-linear relationships were accounted for with transformations and all a priori interactions tested were found to be non-significant.

Pearson, deviance and Pregibon leverage residuals were plotted for all participants in the sample after fitting the final model (Figures. 4.3 – 4.5). Only a relatively small number of observations showed any deviation from the rest of the sample and the leverage statistics showed that these observations contributed negligible amounts

to the overall model performance individually, so removal of these participants was deemed unnecessary.

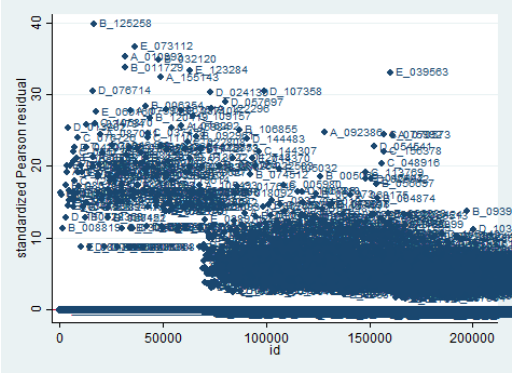

Figure 4.3. Pearson residuals for PLCOall2014mfp model applied to PLCO participants (n=139,212).

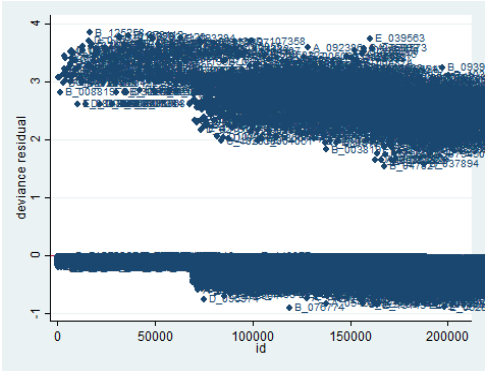

Figure 4-4. Deviance residuals for the PLCOall2014mfp model applied to PLCO participants (n=139, 212).

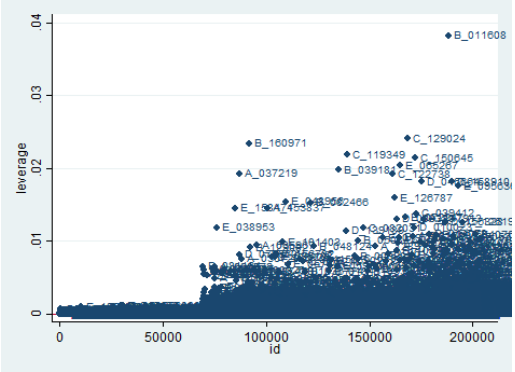

Figure 4.5. Leverage of each observation in the sample after applying PLCOall2014mfp to PLCO participants (n=139, 212).

Collinearity diagnostics were run for all variables included in the final model. No problematic degree of multicollinearity was observed since no variance inflation factor (VIF) exceeded 10, the accepted threshold for high multicollinearity (Kutner, Nachtsheim, & Neter, 2004). The highest VIF observed was for smoking duration (4.89), followed by smoking status (4.86) and the mean VIF was 1.64.

CHAPTER 5 - DISCUSSION AND CONCLUSIONS

5.1 Discussion

5.1.1 Sample characteristics

The approximate 60% increase in SPLC risk in cancer survivors relative to those with no personal history of cancer, was consistent between the two trials. Directly comparable estimates elsewhere are not readily available but this increase is much larger than the 6% higher SPLC risk reported for Japanese cancer survivors (SIR: 1.06; 95% CI: 1.01 – 1.11) (Utada et al., 2014). The difference could be attributable to several factors including differences in participants' characteristics and follow-up periods. Although no follow-up statistics were presented for the development of SPLC specifically, the median and mean follow-up durations for all cancers combined were both lower in the study by Utada and colleagues (1.8 years and 4.3 years, respectively) (Utada et al., 2014) than the 6-year follow-up duration for the current analysis.

5.1.2 Previous cancer sites associated with subsequent primary lung cancer risk

Those who reported having a previous history of any cancer have a 39% increased odds of SPLC after adjustment for known risk factors. The closest comparable estimate from population cancer registries was published by Utada and colleagues (2014); the researchers reported a 6% increased odds of lung cancer following previous cancer at any site (SIR = 1.06; 95% CI: 1.01 – 1.11) among Japanese cancer survivors (Utada et al., 2014) but the median follow-up of 8 months was substantially shorter than the current study.

The effects (or lack thereof) observed in the current study are discussed separately below. The majority of the previous cancer types investigated failed to show positive, significant associations with lung cancer which may indicate that many of the associations reported in previous studies are subject to confounding but could also be due to lack of statistical power compared to other studies that used larger SEER and population-wide cancer registry datasets.

Lung

The site of previous cancer most strongly associated with SPLC was the lung with odds ratios of 17.4 (95% CI: 3.14 – 96.13) and 35.1 (95% CI: 10.8 – 114.6) in the PLCO and NLST, respectively, but both estimates should be interpreted with caution since they are derived from very small numbers of lung cancer survivors (n=7 and n=20, respectively). The risk of developing SPLC following resection of non-small cell lung cancer (NSCLC) has been estimated at approximately 1-2% per person per year (Johnson, 1998). Lou and colleagues reviewed the outcomes of 1,294 patients with early-stage NSCLC who underwent surgical resection and found that 7% developed SPLC; SPLC risk varied from 3-6% per person-year (Lou et al., 2013). The effects of chest irradiation and continued cigarette smoking after an initial lung cancer have both been shown to contribute to SPLC risk (Heyne, Lippman, Lee, Lee, & Hong, 1992; Johnson, 1998; Richardson et al., 1993; Tucker et al., 1997).

Breast

Although previous studies have investigated the risk of subsequent primary cancer (SPC) among breast cancer survivors (Andersson et al., 2008; Mellemkjær et al., 2006), few

have investigated SPLC specifically and many of those compared the effects of treatments (Andersson et al., 2008; Catsburg, Kirsh, Soskolne, Kreiger, & Rohan, 2014; Grantzau, Mellemkjær, & Overgaard, 2013; Lorigan, Califano, Faivre-Finn, Howell, & Thatcher, 2010), smoking (Catsburg et al., 2014; Ford et al., 2003; Kaufman et al., 2008; Neugut et al., 1994) and estrogen receptor (ER) status among survivors rather than comparing those with previous breast cancer to those without (Schonfeld, Curtis, Anderson, & de Gonzalez, 2012). This makes comparisons between the current study and previous research less than straightforward. Studies that reported increased risks for SPLC, typically only did so among sub-groups, particularly those whose breast cancer was diagnosed before menopause (Evans et al., 2001; Harvey & Brinton, 1985). Evans and colleagues reported an increased SPLC risk among women diagnosed before age 50 but found a protective effect for women diagnosed at age 50 or older (Evans et al., 2001). Raymond and Hogue (2006) reported increased SPLC risk among breast cancer survivors but standardized incidence ratios (SIRs) were presented separately by age bracket so an overall estimate is not available for comparison, despite statistically significant positive associations which increased with each age bracket (Raymond & Hogue, 2006).

In the current analysis, the adjusted associations were not statistically significant in either PLCO (OR = 1.22; 95% CI: 0.70 – 2.13) or NLST participants (OR = 1.25; 95% CI: 0.88 – 1.76). This result is similar to the findings of recent studies which reported increased SPLC risk among breast cancer survivors when age at breast cancer diagnosis was not considered (Andersson et al., 2008; Hemminki et al., 2005; Utada et al., 2014). Explanations for the elevation in SPLC among breast cancer survivors include the effects of radiation exposure during cancer treatment and continued cigarette smoking, as well as

ER status as those with ER^- tumours are at higher risk of SPLC compared to those with ER^+ breast cancers (SIR = 1.22; 95% CI: 1.10 – 1.37) (Kaufman et al., 2008; Neugut et al., 1994; Schonfeld et al., 2012).

Digestive

In the NLST, a protective effect on SPLC risk was observed among colorectal cancer survivors but the effect was statistically nonsignificant with a wide confidence interval, thus not allowing for any meaningful conclusions to be drawn (OR = 0.77; 95% CI: 0.27 – 2.15). Should a true association exist between colorectal cancer and SPLC, it may be attributable to several possible mechanisms including misclassification of metastases as new primary cancers as well as shared genetic risk factors (Emi et al., 1992; Korošec, Glavač, Rott, & Ravnik-Glavač, 2006). Colorectal and lung cancers both share smoking as a risk factor (Botteri et al., 2008; Liang, Chen, & Giovannucci, 2009) and the epithelia of both tissues may respond similarly to environmental carcinogens and epigenetic changes since they both of endodermal origin.

In PLCO participants, the increase in SPLC risk seen among esophageal cancer survivors (OR = 1.82; 95% CI: 0.21 – 15.8) fell within the range of statistically significant hazardous associations reported in other studies (Chuang, Hashibe, et al., 2008; Jégu et al., 2014; Zhu et al., 2012) but the effect was statistically nonsignificant and markedly lower than the three- and four-fold increases in risk reported by Zhu et al. (2012) (OR = 3.19; 95% CI: 2.12 – 4.61) and Jégu et al. (2014) (OR = 4.25, CI not presented), respectively. Large increases in SPLC risk have been reported in several previous studies among esophageal cancer survivors in the general population (Chuang,

Hashibe, et al., 2008; Jégu et al., 2014; Utada et al., 2014; Zhu et al., 2012) but these studies were not well adjusted for lung cancer risk factors, especially smoking-related ones. The unadjusted effect in the PLCO was similar (OR = 3.36; 95% CI: 0.46 – 24.7) but the effect was non-significant and was not replicated in the NLST. After adjustment, only a small increase in lung cancer risk was observed among esophageal cancer survivors and the effect was not statistically significant (OR = 1.09; 95% CI: 0.17 – 7.10) however, the confidence interval is wide and includes effects found to be significant in previous studies. Associations between esophageal cancer and lung cancer may be attributable to a "field cancerization" effect whereby the carcinogenic effects of alcohol and tobacco may act on different parts of the aerodigestive tract simultaneously causing multiple independent cancers to occur at various sites (Slaughter, Southwick, & Smejkal, 1953).

The adjusted effects observed for stomach cancer survivors in both the PLCO (OR = 1.06; 95% CI: 0.14 – 8.02) the NLST (OR = 0.71; 95% CI: 0.10 – 5.19) were both statistically nonsignificant with wide confidence intervals, so no accurate sense of the true effects could be drawn from these estimates.

Endocrine

Estimates for the effect of previous thyroid cancer on SPLC risk vary in other studies from statistically nonsignificant null effects (Canchola et al., 2006; Rubino et al., 2003) to statistically significant hazard (C.-H. Lu et al., 2013; Utada et al., 2014) or protective effects (A. P. Brown et al., 2008; Verkooijen et al., 2006). Thyroid cancer survivors of the PLCO were not at increased risk for SPLC after adjustment (OR = 0.46; 95% CI: 0.08

– 2.85). A small increase in risk was seen among NLST thyroid cancer survivors but this was not statistically significant (OR = 1.11; 95% CI: 0.39 – 3.23). The effects in both the PLCO and NLST had wide confidence intervals, thus the estimates are imprecise, but the upper limit tells us that the association between thyroid cancer and SPLC is unlikely to be of a very large magnitude.

Genitourinary

An increased lung cancer risk was observed for kidney cancer survivors in previous studies by Czene & Hemminki (2002) (SIR = 1.36) and Utada and colleagues (2014) (SIR = 1.29); both types of cancer share smoking as a risk factor. An increase in lung cancer risk of similar magnitude to those reported previously was seen among kidney cancer survivors in the NLST after adjustment but was not statistically significant (OR = 1.22; 95% CI: 0.50 – 2.98) and no significant elevation in SPLC risk was seen in PLCO participants previously diagnosed with kidney cancer (OR = 0.93; 95% CI: 0.35 – 2.49).

A clinically relevant increase in SPLC risk was seen following prior bladder cancer in NLST participants (OR = 1.43; 95% CI:: 0.86 – 2.40) but the effects was not statistically significant and not replicated in PLCO participants (OR = 0.96; 95% CI:: 0.57 – 1.61). The almost three-fold increased risk of SPLC (OR = 2.88; 95% CI: 0.21 – 39.1) seen in prostate cancer survivors is unusually high but the confidence interval for the estimate is so wide that the estimate may not be reflective of any true effect.

Gynecologic

Tobacco smoking could partly explain increases in SPLC among cervical cancer survivors since it is an established cofactor in HPV-mediated cervical carcinogenesis

(Fonseca-Moutinho, 2011). The increases in SPLC risk among survivors of cervical cancer reported in previous studies (Hemminki et al., 2000; Jégu et al., 2014) were replicated before controlling for confounding in both the PLCO and NLST cohorts. A marked increase in risk persisted for cervical cancer survivors in the PLCO after adjustment (OR = 1.26; 95% CI: 0.82 – 1.95) but the effect was nonsignificant and not reproduced among NLST participants (OR = 0.96; 95% CI: 0.68 – 1.37).

Among PLCO participants, those with prior uterine cancer (n=532) were twice as likely to develop subsequent lung cancer compared to those without (OR = 2.01; 95% CI: 1.09 – 3.71). The magnitude of this effect was roughly 40-70% higher than has been reported in previous studies among the general population relative to the null of 1.00 (Jégu et al., 2014; Koivisto-Korander et al., 2012). As suggested by Koivisto-Korander et al., miscoding of recurrent uterine cancers may explain part of the association since the lungs are one of the most common sites of uterine cancer metastases occurring in around 7-8% of patients (AlHilli & Mariani, 2013) but this is unlikely to account for such a large effect. Smoking has a protective effect on uterine cancer through antiestrogenic effects such as weight loss and earlier onset of menopause (Michnovicz, Hershcopf, Naganuma, Bradlow, & Fishman, 1986) but this appears to be exclusive to postmenopausal women (Zhou et al., 2008). A large European cohort showed a twofold increased risk of endometrial cancer in premenopausal current smokers compared to never-smokers (Al-Zoughool et al., 2007). It has been suggested that polycyclic aromatic hydrocarbons in tobacco smoke may cause chronic anovulation, an established risk factor for endometrial cancer (Haidopoulos et al., 2010; Kaaks, Lukanova, & Kurzer, 2002). Information on

previous uterine cancer was not present in the NLST dataset so the association could not be re-evaluated.

Head and Neck

Previous head and neck cancers have been associated with some of the highest increases in SPLC risk based on estimates from previous studies (Chen et al., 2011; Chuang, Scelo, et al., 2008; Jégu et al., 2014; Morris, Sikora, Patel, Hayes, & Ganly, 2010; Utada et al., 2014) and lung cancer is the most common subsequent primary malignancy following squamous cell cancer of the head and neck (45.8%) (Griffioen et al., 2015). Although the etiology of SPLC following cancer of the head and neck remains unclear, most cancers of the head and neck are strongly associated with a history of cigarette smoking (Licciardello et al., 1989). It has been suggested that metastatic SCCs of the head and neck are particularly prone to being misclassified as SPLC (Geurts et al., 2005; Kuriakose et al., 2002; Ritoe et al., 2002).

When combined into a single group, survivors of previous head and neck cancers in the PLCO had a 75% increased odds of SPLC after adjustment but the effect was nonsignificant (OR = 1.75; 95% CI: 0.62 – 4.92). This is markedly lower than the effects reported by Chuang et al. (SIR = 3.30) and Jégu et al. (SIR = 8.71 for males and 18.81 for females) but neither of these studies controlled for smoking-related exposures which are potential confounders of an association with SPLC (Chuang, Scelo, et al., 2008; Jégu et al., 2014). Previous nasopharyngeal and (OR = 7.62; 95% CI: 0.35 – 166.2) laryngeal cancers (OR = 1.36; 95% CI: 0.58 – 3.18) were associated with considerable increases in SPLC risk among PLCO participants, but the effect was of much smaller magnitude in

NLST cohort (OR = 1.11; 95% CI: 0.89 – 2.39). A greatly elevated risk of SPLC was observed for those with previous oral cancer (OR = 3.02; 95% CI: 1.41 – 6.45) and the effect was statistically significant but this was only measured in the NLST.

Hematologic

Several studies, including a meta-analysis of 21 investigations of survivors of Hodgkin lymphoma have reported a 4-fold increased risk of SPLC compared to the general population (Ibrahim et al., 2013; Jégu et al., 2014). The effects of smoking and radiotherapy are both established risk factors for SPLC in Hodgkin lymphoma survivors (Ibrahim et al., 2013). The increase in SPLC risk seen among Hodgkin lymphoma survivors in the PLCO was markedly lower than previous studies but still of clinically important magnitude (OR = 1.39; 95% CI: 0.17 – 11.64).

In addition to smoking and radiation exposure, a distinct dose-response relationship between the use of alkylating chemotherapy agents and SPLC risk has been demonstrated among those treated for non-Hodgkin lymphoma (Travis et al., 2002). Survivors of non-Hodgkin lymphoma in the PLCO were not at increased risk for SPLC after adjustment (OR = 1.05; 95% CI: 0.14 – 8.11). This result contrasts with several other studies that reported SPLC as the most common type of SPC among non-Hodgkin lymphoma survivors (Travis et al., 2002).

5.1.3 Number of previous cancers

Being previously diagnosed with more than one cancer had a greater impact on SPLC risk for NLST participants whose cumulative incidence of SPLC increased by 11% (relative to those with only one previous cancer) compared to PLCO participants for

whom the increase was only 6%. However, these effects did not persist after adjustment for other major risk factors and the differences between the groups were not statistically significant. Thus, there was no evidence in the current study that those with more than one previous cancer are at a higher risk of lung cancer than those with only one. To the best of our knowledge, this is the first study to specifically investigate the influence of the frequency of previous cancers of any type with respect to future lung cancer risk. The proportion of participants with more than one previous cancer was smaller than in other studies so it could be that there were too few instances present to detect any considerable elevation in risk.

5.1.4 Age of previous cancer onset

Being diagnosed before age 60 does not appear to affect the association between previous cancer and future lung cancer. This conclusion is supported by the lack of any effect modification between personal cancer history and age of previous cancer onset when testing a priori interactions. The lack of any increased lung cancer risk seen in the younger age at diagnosis group could in part be due to the fact that the vast majority (78.6%) of cancer survivors in the study were diagnosed before age 60. There is a much greater duration over which to develop lung cancer up to age 60 than between 60 and 74 years, the upper limit for enrolment in both trials, so it could be that those diagnosed after age 60 were too few in number for any difference between the age groups to manifest itself.

5.1.5 Predictive utility of these factors

None of the factors investigated here improved the discrimination of the existing PLCOall2014 model to any considerable degree. None of the refined variables representing a personal history of specific cancer types demonstrated as much predictive utility as the original term included in the current model. When applying the final model (PLCOall2014mfp) to all PLCO participants, the point estimate for the AUC (0.855) is slightly higher than that reported by Tammemagi (2014) (AUC = 0.848); the latter was derived from participants in the PLCO intervention arm only (n = 77,445) so this is likely as a result of the increased sample size used to test the model and the point estimate lies within the 95% CI for the AUC reported by Tammemagi (2014) (0.833 – 0.861) suggesting that the increase is unlikely to be a significant one. The difference could also be in part due to slightly different non-linear transformations used as well as the larger sample and updated dataset. The final model did not perform as well when applied to NLST participants (AUC = 0.710). The narrowness of the bootstrap bias-corrected 95% CIs suggest that the performance statistics for the final model are robust, i.e. they are not sensitive to small changes in the data and therefore the model should perform similarly in populations with characteristics that are comparable to those of the development sample. Low median and 90th percentile absolute errors indicate good calibration when applied to the PLCO, NLST and both samples combined.

5.1.4 Limitations

Perhaps one of the most important limitations of any investigation into the effects of previous cancers on SPLC risk is that multiple primary malignancies in the same individual is uncommon. Population-wide cancer registries afford researchers the largest

numbers of occurrences but are limited in their ability to adjust for influential cofactors. Even though the PLCO and NLST are large trials, low numbers of specific previous cancer types limit the number of outcome events, leading to wide confidence intervals for the effect estimates – often too large to allow meaningful conclusions to be drawn. Further limitations are discussed below.

5.1.4.1 Self-reported occurrences of previous cancer

A possible source of measurement bias in this study is the accuracy of self-reported information on personal cancer history. A few studies investigated this (Bergmann, Calle, Mervis, Miracle-McMahill, & Thun, 1998; Desai et al., 2001; Manjer, Merlo, & Berglund, 2004; Parikh-Patel, Allen, & Wright, 2003; Schrijvers, Stronks, van de Mheen, Coebergh, & Mackenbach, 1994). Each of these studies used mailed questionnaires or in-person interviews and compared participants' responses to cancer registry data. Breast cancer is the most accurately reported previous cancer type in terms of true positive responses (Bergmann et al., 1998; Berthier et al., 1997; Desai et al., 2001). In their study of 133,479 current and former Californian teachers, Parikh-Patel and colleagues found that the accuracy of self-reported cancer data varied by site. Sensitivity (proportion of true positives classified as positive) varied greatly by cancer site, from 94.6% and 92.9% for breast and thyroid cancers, to between 80.0 and 85.9% for lung, colorectal and ovarian cancers and a markedly poorer 44.3% for cervical cancer (Parikh-Patel et al., 2003). Specificity (proportion of true negatives classed as negative) ranged from 90.2-99.9%, but was above 98% for all but one cancer type ('other skin').

Non-white race has been associated with increased false-positive responses (OR = 1.34), as has older age (among those 45+ compared to <45) after adjusting for race and socioeconomic status (Parikh-Patel et al., 2003). In contrast, other researchers found non-white race and older age to be predictors of increased false-negative reporting of cancer history but the latter finding was only true in those older than the maximum age of the participants in the current analysis (74) (Desai et al., 2001). Dominguez and colleagues assessed the accuracy of self-reported data in an outpatient breast cancer center. Among 3,614 women who had a single cancer according to the registry and reported none or one cancer on their questionnaire, the overall sensitivity was 85.7%. This varied by cancer type from 92.1% for breast cancer to 42.9% for leukemia and white women reported breast cancer more accurately (in terms of sensitivity) than did Asian/Pacific Islanders (p = 0.008), but no statistically significant differences were found in comparison to black, Caribbean or West Indian, or Native American. Older age at the time of the study did not have any effect on sensitivity for any group ≥ 45 compared with those aged 44 or younger (Dominguez et al., 2007).

Although the general accuracy of self-reported cancer history data is suspect to much variation, this study was not susceptible to the problems associated with reports from those aged 75+ as they were not included and since 86.6% of the participants were white the influence of inaccuracies observed in non-whites should be minimal (though the findings may consequently be less generalizable to them). Although cervical cancers were frequently underreported in the study of Californian teachers, the prevalence in the current analysis was relatively high in NLST participants, possibly due to the proportion of heavy smokers at high risk for the disease. Thus, problems associated with

underrepresentation should be of little concern. Reporting the wrong type of previous cancer seems more likely than misreporting its occurrence in general, so problems with misclassification ought to have a great impact on associations with specific cancer types than the crude overall measure. Errors in self-reporting of personal cancer history should lead to decreased performance in terms of discrimination and calibration since they are dependent upon changes in both sensitivity and specificity. As sensitivity decreases, false-negatives increase and as specificity decreases, false-positives increase. Either of these scenarios (together or separately) will reduce predictive performance. Although no unexpectedly large effects that may have been attributable to differential misclassification were seen, it remains possible that some associations apparent in large registry studies that have more accurate cancer history classification, may have been underestimated in the current study.

5.1.4.2 Creation of new variables

For the majority of the cancer types included in the current analyses, dichotomous variables representing previous diagnosis had already been created by Information Management Systems (IMS). Most of these were measured in both trials such that harmonized variables could be created, with the exception of oral cancer which was only available for the NLST. Variables for previous prostate, uterine and both non-Hodgkin and Hodgkin lymphoma had to be created manually using data from questions only asked in the PLCO so associations for these cancers could not be tested among NLST participants.

The score representing the number of previous cancers was created using as much data as possible but differences in the information present for PLCO and NLST participants could have led to problems with construct validity (how well the variable measures what it intends to measure) to some degree. PLCO participants reported the type and age at diagnosis for up to four previous cancers, irrespective of cancer type (i.e. subsequent cancers were considered separate even in they were of the same type as the first). A score of 0, 1 or 2+ was assigned to PLCO participants according to the number of cancers for which the participant had reported the anatomical site (regardless of what the sites were) from the variables 'ph_type1'-'ph_type4'. However, the information for NLST participants was not analogous. Instead, several variables simply indicated the presence or absence of specific cancer types. Therefore, a participant with two previous cancers of the same type, would be assigned a 1 despite having had two separate previous cancers. This means that the number of NLST participants with two or more previous cancers may be underestimated since subsequent primary cancers of the same cancer type are not considered separate. This underestimation could decrease the statistical power to detect differences in risk between those with one previous cancer and those with two or more. This could be important since even if 1% of the 2,208 NLST participants denoted as having one cancer were misclassified, those 22 people would represent an increase of 22% to the size of the 2+ group (n=100).

5.1.4.3 Age at diagnosis cut-off

Much has been written about the hazards associated with dichotomizing continuous data (Baneshi & Talei, 2011; Cumberland et al., 2014; Royston, Altman, & Sauerbrei, 2006). Dichotomization of a continuous variable results in a loss of information and in turn,

decreases statistical power. It may also lead to an inflation of type I error (Austin & Brunner, 2004). Finally there is also the problem of how to choose the most appropriate cut-point.

Age at diagnosis was dichotomized for several reasons, including ease of interpretation. If the variable was left in its original continuous form and included in prediction models, requiring the participant to know their exact age at diagnosis of their initial cancer may result in higher information bias compared to recalling whether their initial cancer was diagnosed before a given age. Categorizing the age at diagnosis also enables the results to be more easily compared to previous studies. Age 60 was chosen based on the sharp increase in 10-year lung cancer risk seen after this age and to maximize the number of observations in the < 60 category.

5.1.4.4 Modeling limitations

Logistic regression allowed the results of the prediction modeling portion of this study to be directly comparable with previous work, however, this modeling technique does not capture time-to-event data unlike Cox proportional hazards models (Cox, 1972). Harrell's c statistic (Harrell, Califf, Pryor, Lee, & Rosati, 1982; Harrell, Lee, & Mark, 1996) is the equivalent to the AUC. The c statistic for the comparable Cox model for 6-year lung cancer risk was only slightly higher, at 0.856 compared to an AUC of 0.855 for the logistic.

The c statistic for a Cox model for 10-year lung cancer risk was computed to ascertain whether loss of information by truncating follow-up to 6 years was detrimental to predictive performance, but it was found to be lower than the 6-year model (c = 0.850).

It is unclear whether the performance deterioration over the longer follow-up period is due to a weakening of the association between all of the predictors and the outcome or only for some subset of predictors in particular.

Several of the associations of interest in this study were initially unable to be estimated due to problems with complete separation but the solution of intentionally misclassifying one participant as a workaround is not without its disadvantages, as have been previously alluded to. The likelihood of complete separation diminishes with increasing sample size but since these events are rare, minimizing the problem would require even greater sample sizes than were available in the PLCO and NLST. Many of the associations that were not estimable among PLCO participants were able to be calculated among NLST participants.

5.1.5 Strengths

There is a widely-held assumption that a prior cancer diagnosis could affect the conduct or outcomes of cancer-related studies, despite a lack of conclusive evidence to support the claim (Gerber, Laccetti, Xuan, Halm, & Pruitt, 2014). Gerber and colleagues reviewed SEER data and found that 80% of 51 lung cancer clinical trials excluded patients with a prior cancer diagnosis and 95% among trials with survival as their primary endpoint. The researchers estimate that the proportion of participants excluded due to this criterion ranged from 0-18% which is larger than the total ineligibility proportion in other cancer trials (Gerber et al., 2014). Further research needs to be conducted to better understand the implications of this practice, but since neither the PLCO nor NLST employed this policy, they provided an excellent opportunity to study the effects of

personal cancer history in detail. High quality data collection protocols ensured that data was accurate and the amount of missing data was minimized. The number of important confounding factors that were able to be adequately controlled for when investigating these associations is also unprecedented. Many of the more considerable effects seen were for smoking-related cancers. This underlines cigarette smoking as a modifiable risk factor for lung cancer.

This study is of clinical importance in that it suggests that when trying to identify those at highest risk for lung cancer, simply establishing whether or not they are a cancer survivor is sufficient compared to more sophisticated information, an accurate report of which may be difficult to obtain. With respect to research surrounding the risk of second primary lung cancer, our study underlines the importance of adjusting for key cofactors, particularly smoking-related information which is commonly omitted.

5.1.7 Future directions

Future studies should investigate the impact of previous cancer types that were considered exclusion criteria for the two screening trials analyzed here (prostate, lung, colorectal and ovarian). It may also be prudent to consider relaxing this criterion for future lung cancer screening trials. The effects of previous prostate and testicular cancers on SPLC risk after adjustment for known lung cancer risk factors require clarification as they were inestimable in the current analysis. Clarification is also required as to whether the marked increase in risk of SPLC observed among lung cancer survivors persists after controlling for other known lung cancer risk factors. Patterns of inaccurate reporting also

need to be better understood as well as the reliability of self-reported cancer data as a whole (Harlow & Linet, 1989).

The impact of treatments for previous cancer should be studied in more detail as data to assess the factors were not available here. Specifically, to assess whether the use of chemotherapeutic agents and/or radiation dose modify the association between a personal history of cancer and future lung cancer risk. Little is known about the extent to which smoking interacts with ionizing radiation in terms of lung cancer carcinogenesis so this also requires further study. Since it has been estimated that around 30% of cancers worldwide are attributable to tobacco use, smoking status is considered by some researchers as 'another vital sign' that ought to be measured at several stages of all future oncology trials: at registration, diagnosis, throughout treatment and during follow-up if possible (Gritz, Dresler, & Sarna, 2005).

Future efforts to improve lung cancer risk prediction should pursue alternative strategies such as refining risk factors other than personal cancer history or investigating novel predictors as candidates for inclusion in future models.

5.2 Conclusions

Despite having far lower absolute numbers of cancer survivors compared to studies that have used all cases found in cancer registries, this analysis is novel in that no other study has evaluated the effects of previous cancers on lung cancer risk after adjustment for many important covariates. Contrary to many of the hazardous effects reported for specific cancer types in previous studies, relatively few cancer types, save for oral and

uterine, were associated with a statistically significant increase in SPLC risk after controlling for known lung cancer risk factors.

No evidence of a trend between the number of previous cancers reported and lung cancer risk was seen for the participants of either screening trial. The risk of SPLC relative to those with no cancer was not statistically different between those diagnosed with an initial cancer before age 60 compared with those whose first cancer was diagnosed after 60.

None of the more sophisticated measures of personal cancer history improved lung cancer risk prediction. The original binary measure representing a personal history of any cancer demonstrated comparable predictive value to the more refined summary variables based on specific cancer types, head and neck cancers only or smoking-related cancers. Neither the number of previous cancers nor the age at first cancer diagnosis improved prediction when included alongside the established predictors. This more sophisticated information regarding personal history does not appear to improve lung cancer risk prediction compared to simply capturing its overall presence.

REFERENCES

Abdel-Hamid, M., Chen, J.-J., Constantine, N., Massoud, M., & Raab-traub, N. (1992). EBV strain variation: geographical distribution and relation to disease state. *Virology*, *190*(1), 168–175.

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

Aberle, D. R., Adams, A., Berg, C. D., Clapp, J. D., Clingan, K. L., Gareen, I. F., … Pinsky, P. F. (2010). Baseline Characteristics of Participants in the Randomized National Lung Screening Trial, (April 2004), 1771–1779. doi:10.1093/jnci/djq434

AlHilli, M. M., & Mariani, A. (2013). The role of para-aortic lymphadenectomy in endometrial cancer. *International Journal of Clinical Oncology*, *18*(2), 193–199.

Al-Zoughool, M., Dossus, L., Kaaks, R., Clavel-Chapelon, F., Tjønneland, A., Olsen, A., … Linseisen, J. (2007). Risk of endometrial cancer in relationship to cigarette smoking: results from the EPIC study. *International Journal of Cancer*, *121*(12), 2741–2747.

American Cancer Society. (2012). *Second Cancers Caused by Cancer Treatment*.

American Cancer Society. (2013a). *Cancer Facts and Figures 2013*. Atlanta.

American Cancer Society. (2013b). *Lifetime Risk of Developing or Dying From Cancer*.

American Cancer Society. (2013c). *Lung Cancer (Non-Small Cell): Detailed Guide*.

American Cancer Society. (2014). Cancer Facts & Figures 2014.

American Cancer Society. (2015). Non-small cell lung cancer survival rates by stage.

Andersson, M., Jensen, M.-B., Engholm, G., & Henrik Storm, H. (2008). Risk of second primary cancer among patients with early operable breast cancer registered or randomised in Danish Breast Cancer cooperative Group (DBCG) protocols of the 77, 82 and 89 programmes during 1977-2001. *Acta Oncologica*, *47*(4), 755–764.

Arnold, M., Liu, L., Kenter, G. G., Creutzberg, C. L., Coebergh, J.-W. W., & Soerjomataram, I. (2014). Second primary cancers in survivors of cervical cancer in the Netherlands: Implications for prevention and surveillance. *Radiotherapy and Oncology*.

Aschengrau, A., & Seage, G. (2008). *Essentials of epidemiology in public health*. Jones & Bartlett Learning.

Austin, P. C., & Brunner, L. J. (2004). Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine*, *23*(7), 1159–1178.

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

Bach, P. B., Mirkin, J. N., Oliver, T. K., Azzoli, C. G., Berry, D. A., Brawley, O. W., … Jett, J. R. (2012). Benefits and Harms of CT Screening for Lung CancerA Systematic ReviewBenefits and Harms of CT Screening for Lung Cancer. *Jama*, *307*(22), 2418–2429.

Bae, M. K., Byun, C. S., Lee, C. Y., Lee, J. G., Park, I. K., Kim, D. J., & Chung, K. Y. (2011). The role of surgical treatment in second primary lung cancer. *The Annals of Thoracic Surgery*, *92*(1), 256–262.

Baneshi, M. R., & Talei, A. R. (2011). Dichotomisation of continuous data: review of methods, advantages, and disadvantages. *Iranian Journal of Cancer Prevention*, *4*(1), 26–32.

Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*(5), 464–469.

Bergmann, M. M., Calle, E. E., Mervis, C. A., Miracle-McMahill, H. L., & Thun, M. J. (1998). Validity of self-reported Cancers in a Propsective Cohort Study in Comparison with Data from State Cancer Registries. *American Journal of Epidemiology*, *147*(6), 556–562.

Berthe, E., Henry-Amar, M., Michels, J.-J., Rame, J.-P., Berthet, P., Babin, E., … Mahoudeau, J. (2004). Risk of second primary cancer following differentiated thyroid cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, *31*(5), 685–691.

Berthier, F., Grosclaude, P., Bocquet, H., Faliu, B., Cayla, F., & Machelard-Roumagnac, M. (1997). Prevalence of cancer in the elderly: discrepancies between self-reported and registry data. *British Journal of Cancer*, *75*(3), 445.

Birgisson, H., Påhlman, L., Gunnarsson, U., & Glimelius, B. (2005). Occurrence of second cancers in patients treated with radiotherapy for rectal cancer. *Journal of Clinical Oncology*, *23*(25), 6126–6131.

Boice Jr, J. D., Land, C. E., & Preston, D. L. (1996). Ionizing radiation. Cancer Epidemiology and Prevention. New York: Oxford University Press.

Botteri, E., Iodice, S., Bagnardi, V., Raimondi, S., Lowenfels, A. B., & Maisonneuve, P. (2008). Smoking and colorectal cancer: a meta-analysis. *JaMa*, *300*(23), 2765–2778.

Brawley, O. W. (2012). Prostate cancer epidemiology in the United States. *World Journal of Urology*, *30*(2), 195–200.

Brenner, D. J., Curtis, R. E., Hall, E. J., & Ron, E. (2000). Second malignancies in prostate carcinoma patients after radiotherapy compared with surgery. *Cancer*, *88*(2), 398–406.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Brown, A. P., Chen, J., Hitchcock, Y. J., Szabo, A., Shrieve, D. C., & Tward, J. D. (2008). The risk of second primary malignancies up to three decades after the treatment of differentiated thyroid cancer. *The Journal of Clinical Endocrinology & Metabolism*, *93*(2), 504–515.

Brown, L. M., Pottern, L. M., & Hoover, R. N. (1987). Testicular cancer in young men: the search for causes of the epidemic increase in the United States. *Journal of Epidemiology and Community Health*, *41*(4), 349–354.

Caini, S., Boniol, M., Botteri, E., Tosti, G., Bazolli, B., Russell-Edu, W., … Gandini, S. (2014). The risk of developing a second primary cancer in melanoma patients: A comprehensive review of the literature and meta-analysis. *Journal of Dermatological Science*, *75*(1), 3–9.

Canadian Cancer Society's Advisory Committee on Cancer Statistics. (2014). *Canadian Cancer Statistics 2015*. Toronto, ON.

Canchola, A. J., Horn-Ross, P. L., & Purdie, D. M. (2006). Risk of second primary malignancies in women with papillary thyroid cancer. *American Journal of Epidemiology*, *163*(6), 521–527.

Cantwell, M. M., Murray, L. J., Catney, D., Donnelly, D., Autier, P., Boniol, M., … Gavin, A. T. (2009). Second primary cancers in patients with skin cancer: a population-based study in Northern Ireland. *British Journal of Cancer*, *100*(1), 174–177.

Cassidy, A., Myles, J. P., van Tongeren, M., Page, R. D., Liloglou, T., Duffy, S. W., & Field, J. K. (2008). The LLP risk model: an individual risk prediction model for lung cancer. *British Journal of Cancer*, *98*(2), 270–276.

Catsburg, C., Kirsh, V. A., Soskolne, C. L., Kreiger, N., & Rohan, T. E. (2014). Active cigarette smoking and the risk of breast cancer: a cohort study. *Cancer Epidemiology*, *38*(4), 376–381.

Centers for Medicare & Medicaid Services. (2015). Final National Coverage Determination on Screening for Lung Cancer with Low Dose Computed Tomography (LDCT) (CAG-00439N). Retrieved June 10, 2015, from http://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=274

Chaturvedi, A. K., Engels, E. A., Gilbert, E., Chen, B. E., Storm, H., Lynch, C. F., … Kaijser, M. (2007). Second cancers among 104760 survivors of cervical cancer: evaluation of long-term risk. *Journal of the National Cancer Institute*, *99*(21), 1634–1643.

Chen, M.-C., Chen, P.-T., Chan, C. H., Yang, C.-T., Chen, C.-C., Huang, C.-E., … Lee, K.-D. (2011). Second primary esophageal or lung cancer in patients with head and neck carcinoma in Taiwan: incidence and risk in relation to primary index tumor site. *Journal of Cancer Research and Clinical Oncology*, *137*(1), 115–123.

Chuang, S.-C., Hashibe, M., Scelo, G., Brewster, D. H., Pukkala, E., Friis, S., … Tamaro, S. (2008). Risk of second primary cancer among esophageal cancer patients: a pooled analysis of 13 cancer registries. *Cancer Epidemiology Biomarkers & Prevention*, *17*(6), 1543–1549.

Chuang, S.-C., Scelo, G., Tonita, J. M., Tamaro, S., Jonasson, J. G., Kliewer, E. V, … Tracey, E. (2008). Risk of second primary cancer among patients with head and neck cancers: a pooled analysis of 13 cancer registries. *International Journal of Cancer*, *123*(10), 2390–2396.

Chung, H. T. (2008). Prostate Cancer. In J. Lu & L. Brady (Eds.), *Radiation Oncology SE - 20* (pp. 299–319). Springer Berlin Heidelberg. doi:10.1007/978-3-540-77385-6_20

Colditz, G. A., Atwood, K. A., Emmons, K., Monson, R. R., Willett, W. C., Trichopoulos, D., & Hunter, D. J. (2000). Harvard report on cancer prevention volume 4: Harvard Cancer Risk Index. *Cancer Causes & Control*, *11*(6), 477–488.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187–220.

Cumberland, P. M., Czanner, G., Bunce, C., Doré, C. J., Freemantle, N., & García-Fiñana, M. (2014). Ophthalmic statistics note: the perils of dichotomising continuous variables. *British Journal of Ophthalmology*, *98*(6), 841–843.

Czene, K., & Hemminki, K. (2002). Kidney cancer in the Swedish Family Cancer Database: familial risks and second primary malignancies. *Kidney International*, *61*(5), 1806–1813.

Desai, M. M., Bruce, M. L., Desai, R. A., & Druss, B. G. (2001). Validity of self-reported cancer history: a comparison of health interview data and cancer registry records. *American Journal of Epidemiology*, *153*(3), 299–306.

Dominguez, F. J., Lawrence, C., Halpern, E. F., Drohan, B., Grinstein, G., Black, D. M., … Kopans, D. B. (2007). Accuracy of self-reported personal history of cancer in an outpatient breast center. *Journal of Genetic Counseling*, *16*(3), 341–345.

Dong, C., Lönnstedt, I., & Hemminki, K. (2001). Familial testicular cancer and second primary cancers in testicular cancer patients by histological type. *European Journal of Cancer*, *37*(15), 1878–1885.

Dossus, L., Boutron-Ruault, M., Kaaks, R., Gram, I. T., Vilier, A., Fervers, B., … Overvad, K. (2014). Active and passive cigarette smoking and breast cancer risk: results from the EPIC cohort. *International Journal of Cancer*, *134*(8), 1871–1888.

Duval, R. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Sage.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 1–26.

Efron, B., & Stein, C. (1981). The jackknife estimate of variance. *The Annals of Statistics*, 586–596.

Emi, M., Fujiwara, Y., Nakajima, T., Tsuchiya, E., Tsuda, H., Hirohashi, S., … Nakamura, Y. (1992). Frequent loss of heterozygosity for loci on chromosome 8p in hepatocellular carcinoma, colorectal cancer, and lung cancer. *Cancer Research*, *52*(19), 5368–5372.

Etzel, C. J., & Bach, P. B. (2011). Estimating individual risk for lung cancer. In *Seminars in respiratory and critical care medicine* (Vol. 32, pp. 3–9). © Thieme Medical Publishers.

Evans, H. S., Lewis, C. M., Robinson, D., Bell, C. M. J., Møller, H., & Hodgson, S. V. (2001). Incidence of multiple primary cancers in a cohort of women diagnosed with breast cancer in southeast England. *British Journal of Cancer*, *84*(3), 435.

Ewertz, M., & Mouridsen, H. T. (1985). Second cancer following cancer of the female breast in Denmark, 1943-80. *National Cancer Institute Monograph*, *68*, 325–329.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Fonseca-Moutinho, J. A. (2011). Smoking and Cervical Cancer. *ISRN Obstetrics and Gynecology*, *2011*, 847684. doi:10.5402/2011/847684

Ford, M. B., Sigurdson, A. J., Petrulis, E. S., Ng, C. S., Kemp, B., Cooksley, C., … Bondy, M. L. (2003). Effects of smoking and radiotherapy on lung carcinoma in breast carcinoma survivors. *Cancer*, *98*(7), 1457–1464.

Gaudet, M. M., Gapstur, S. M., Sun, J., Diver, W. R., Hannan, L. M., & Thun, M. J. (2013). Active smoking and breast cancer risk: original cohort data and meta-analysis. *Journal of the National Cancer Institute*, djt023.

Gerber, D. E., Laccetti, A. L., Xuan, L., Halm, E. A., & Pruitt, S. L. (2014). Impact of Prior Cancer on Eligibility for Lung Cancer Clinical Trials. *Journal of the National Cancer Institute*, *106*(11), dju302.

Geurts, T. W., Nederlof, P. M., van den Brekel, M. W. M., van't Veer, L. J., de Jong, D., Hart, A. A. M., … van Velthuysen, M.-L. F. (2005). Pulmonary squamous cell carcinoma following head and neck squamous cell carcinoma: metastasis or second primary? *Clinical Cancer Research*, *11*(18), 6608–6614.

Gilbert, E., Stovall, M., Gospodarowicz, M. K., Van Leeuwen, F. E., Andersson, M., Glimelius, B., … Holowaty, E. (2009). Lung cancer after treatment for Hodgkin's disease: focus on radiation effects.

Grantzau, T., Mellemkjær, L., & Overgaard, J. (2013). Second primary cancers after adjuvant radiotherapy in early breast cancer patients: a national population based study under the Danish Breast Cancer Cooperative Group (DBCG). *Radiotherapy and Oncology*, *106*(1), 42–49.

Grantzau, T., & Overgaard, J. (2014). Risk of second non-breast cancer after radiotherapy for breast cancer: A systematic review and meta-analysis of 762,468 patients. *Radiotherapy and Oncology*.

Grantzau, T., Thomsen, M. S., Væth, M., & Overgaard, J. (2014). Risk of second primary lung cancer in women after radiotherapy for breast cancer. *Radiotherapy and Oncology*.

Griffioen, G. H. M. J., Louie, A. V, de Bree, R., Smit, E. F., Paul, M. A., Slotman, B. J., … Senan, S. (2015). Second primary lung cancers following a diagnosis of primary head and neck cancer. *Lung Cancer*.

Gritz, E. R., Dresler, C., & Sarna, L. (2005). Smoking, the missing drug interaction in clinical trials: ignoring the obvious. *Cancer Epidemiology Biomarkers & Prevention*, *14*(10), 2287–2293.

Haidopoulos, D., Simou, M., Akrivos, N., Rodolakis, A., Vlachos, G., Fotiou, S., … Protopappas, A. (2010). Risk factors in women 40 years of age and younger with endometrial carcinoma. *Acta Obstetricia et Gynecologica Scandinavica*, *89*(10), 1326–1330.

Harlow, S. D., & Linet, M. S. (1989). Agreement between questionnaire data and medical records: the evidence for accuracy of recall. *American Journal of Epidemiology*, *129*(2), 233–248.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*, 361–387.

Harvey, E. B., & Brinton, L. A. (1985). Second cancer following cancer of the breast in Connecticut, 1935-82. *National Cancer Institute Monograph*, *68*, 99–112.

Hemminki, K., Dong, C., & Vaittinen, P. (2000). Second primary cancer after in situ and invasive cervical cancer. *Epidemiology*, *11*(4), 457–461.

Hemminki, K., Li, X., & Dong, C. (2001). Second primary cancers after sporadic and familial colorectal cancer. *Cancer Epidemiology Biomarkers & Prevention*, *10*(7), 793–798.

Hemminki, K., Scelo, G., Boffetta, P., Mellemkjaer, L., Tracey, E., Andersen, A., … Kliewer, E. V. (2005). Second primary malignancies in patients with male breast cancer. *British Journal of Cancer*, *92*(7), 1288–1292.

Henderson, B. E., Benton, B., Jing, J., Yu, M. C., & Pike, M. C. (1979). Risk factors for cancer of the testis in young men. *International Journal of Cancer*, *23*(5), 598–602.

Heyne, K. H., Lippman, S. M., Lee, J. J., Lee, J. S., & Hong, W. K. (1992). The incidence of second primary tumors in long-term survivors of small-cell lung cancer. *Journal of Clinical Oncology*, *10*(10), 1519–1524.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, *327*(7414), 557.

Hinnen, K. A., Schaapveld, M., van Vulpen, M., Battermann, J. J., van der Poel, H., van Oort, I. M., … Monninkhof, E. M. (2011). Prostate brachytherapy and second primary cancer risk: a competitive risk analysis. *Journal of Clinical Oncology*, *29*(34), 4510–4515.

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, *16*(9), 965–980.

Hosmer, D. W., & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics-Theory and Methods*, *9*(10), 1043–1069.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Howlander, N., Noone, A. M., Krapcho, M., Garshell, J., Miller, D., & Altekruse, S. F. (n.d.). SEER cancer statistics review, 1975–2011. Bethesda, MD: National Cancer Institute; 2014.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Ibrahim, E. M., Kazkaz, G. A., Abouelkhair, K. M., Al-Mansour, M. M., Al-Fayea, T. M., Al-Foheidi, M., … Elmasri, O. A. (2013). Increased risk of second lung cancer in hodgkin's lymphoma survivors: a meta-analysis. *Lung*, *191*(1), 117–134.

Ikeda, Y., Saku, M., Kawanaka, H., Nonaka, M., & Yoshida, K. (2002). Features of second primary cancer in patients with gastric cancer. *Oncology*, *65*(2), 113–117.

Inskip, P. D., Stovall, M., & Flannery, J. T. (1994). Lung cancer risk and radiation dose among women treated for breast cancer. *Journal of the National Cancer Institute*, *86*(13), 983–988.

Jégu, J., Colonna, M., Daubisse-Marliac, L., Trétarre, B., Ganry, O., Guizard, A.-V., … Woronoff, A.-S. (2014). The effect of patient characteristics on second primary cancer risk in France. *BMC Cancer*, *14*(1), 94.

Johnson, B. E. (1998). Second lung cancers in patients after treatment for an initial lung cancer. *Journal of the National Cancer Institute*, *90*(18), 1335–1345.

Kaaks, R., Lukanova, A., & Kurzer, M. S. (2002). Obesity, Endogenous Hormones, and Endometrial Cancer Risk A Synthetic Review. *Cancer Epidemiology Biomarkers & Prevention*, *11*(12), 1531–1543.

Kaufman, E. L., Jacobson, J. S., Hershman, D. L., Desai, M., & Neugut, A. I. (2008). Effect of breast cancer radiotherapy and cigarette smoking on risk of second primary lung cancer. *Journal of Clinical Oncology*, *26*(3), 392–398.

Kleinerman, R. A., Liebermann, J. V, & Li, F. P. (1985). Second cancer following cancer of the male genital system in Connecticut, 1935-82. *National Cancer Institute Monograph*, *68*, 139–147.

Koivisto-Korander, R., Scélo, G., Ferro, G., Mellemkjær, L., Hemminki, K., Weiderpass, E., … Brewster, D. H. (2012). Second primary malignancies among women with uterine sarcoma. *Gynecologic Oncology*, *126*(1), 30–35.

Koppe, M. J., Zoetmulder, F. A. N., van Zandwijk, N., Hart, A. A. M., Baas, P., & Rutgers, E. J. T. (2001). The prognostic significance of a previous malignancy in operable non-small cell lung cancer. *Lung Cancer*, *32*(1), 47–53.

Korošec, B., Glavač, D., Rott, T., & Ravnik-Glavač, M. (2006). Alterations in the ATP2A2 gene in correlation with colon and lung cancer. *Cancer Genetics and Cytogenetics*, *171*(2), 105–111.

Kovalchik, S. A., Tammemagi, M. C., Berg, C. D., Caporaso, N., Riley, T. L., Korch, M., … Katki, H. A. (2013). Targeting of low-dose CT screening according to the risk of lung-cancer death. *New England Journal of Medicine*, *369*(3), 245–254.

Kramer, A. A., & Zimmerman, J. E. (2007). Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited*. *Critical Care Medicine*, *35*(9), 2052–2056.

Kuriakose, M. A., Loree, T. R., Rubenfeld, A., Anderson, T. M., Datta, R. V, Hill, H., … Hicks, W. L. (2002). Simultaneously presenting head and neck and lung cancer: a diagnostic and treatment dilemma. *The Laryngoscope*, *112*(1), 120–123.

Kutner, M. H., Nachtsheim, C., & Neter, J. (2004). *Applied linear regression models*. McGraw-Hill/Irwin.

Langkamp, D. L., Lehman, A., & Lemeshow, S. (2010). Techniques for handling missing data in secondary analyses of large surveys. *Academic Pediatrics*, *10*(3), 205–210.

León, X., Quer, M., Diez, S., Orús, C., López-Pousa, A., & Burgués, J. (1999). Second neoplasm in patients with head and neck cancer. *Head & Neck*, *21*(3), 204–210.

Liang, P. S., Chen, T., & Giovannucci, E. (2009). Cigarette smoking and colorectal cancer incidence and mortality: Systematic review and meta-analysis. *International Journal of Cancer*, *124*(10), 2406–2415.

Licciardello, J. T. W., Spitz, M. R., & Hong, W. K. (1989). Multiple primary cancer in patients with cancer of the head and neck: second cancer of the head and neck, esophagus, and lung. *International Journal of Radiation Oncology* Biology* Physics*, *17*(3), 467–476.

Liskow, A. S., Neugut, A. I., Benson, M., Olsson, C. A., Birkhoff, J., & Chang, C. H. (1987). Multiple primary neoplasms in association with prostate cancer in black and white patients. *Cancer*, *59*(2), 380–384.

Lorigan, P., Califano, R., Faivre-Finn, C., Howell, A., & Thatcher, N. (2010). Lung cancer after treatment for breast cancer. *The Lancet Oncology*, *11*(12), 1184–1192.

Lou, F., Huang, J., Sima, C. S., Dycoco, J., Rusch, V., & Bach, P. B. (2013). Patterns of recurrence and second primary lung cancer in early-stage lung cancer survivors followed with routine computed tomography surveillance. *The Journal of Thoracic and Cardiovascular Surgery*, *145*(1), 75–82.

Lu, C.-H., Lee, K.-D., Chen, P.-T., Chen, C.-C., Kuan, F.-C., Huang, C.-E., … Chen, M.-C. (2013). Second primary malignancies following thyroid cancer: a population-based study in Taiwan. *European Journal of Endocrinology*, *169*(5), 577–585.

Manjer, J., Merlo, J., & Berglund, G. (2004). Validity of self-reported information on cancer: determinants of under-and over-reporting. *European Journal of Epidemiology*, *19*(3), 239–247.

Mariotto, A. B., Rowland, J. H., Ries, L. A. G., Scoppa, S., & Feuer, E. J. (2007). Multiple cancer prevalence: a growing challenge in long-term survivorship. *Cancer Epidemiology Biomarkers & Prevention*, *16*(3), 566–571.

Martini, N., Bains, M. S., Burt, M. E., Zakowski, M. F., McCormack, P., Rusch, V. W., & Ginsberg, R. J. (1995). Incidence of local recurrence and second primary tumors in resected stage I lung cancer. *The Journal of Thoracic and Cardiovascular Surgery*, *109*(1), 120–129.

Mellemkjær, L., Friis, S., Olsen, J. H., Scélo, G., Hemminki, K., Tracey, E., … McBride, M. L. (2006). Risk of second cancer among women with breast cancer. *International Journal of Cancer*, *118*(9), 2285–2292.

Menard, S. (2002). *Applied logistic regression analysis* (Vol. 106). Sage.

Michnovicz, J. J., Hershcopf, R. J., Naganuma, H., Bradlow, H. L., & Fishman, J. (1986). Increased 2-hydroxylation of estradiol as a possible mechanism for the anti-estrogenic effect of cigarette smoking. *New England Journal of Medicine*, *315*(21), 1305–1309.

Moon, K., Stukenborg, G. J., Keim, J., & Theodorescu, D. (2006). Cancer incidence after localized therapy for prostate cancer. *Cancer*, *107*(5), 991–998.

Moons, K. G. M., Kengne, A. P., Woodward, M., Royston, P., Vergouwe, Y., Altman, D. G., & Grobbee, D. E. (2012). Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart*, *98*(9), 683–690.

Morris, L. G. T., Sikora, A. G., Hayes, R. B., Patel, S. G., & Ganly, I. (2011). Anatomic sites at elevated risk of second primary cancer after an index head and neck cancer. *Cancer Causes & Control*, *22*(5), 671–679.

Morris, L. G. T., Sikora, A. G., Patel, S. G., Hayes, R. B., & Ganly, I. (2010). Second primary cancers after an index head and neck cancer: subsite-specific trends in the era of human papillomavirus–associated oropharyngeal cancer. *Journal of Clinical Oncology*, JCO–2010.

Morton, L. M., Curtis, R. E., Linet, M. S., Bluhm, E. C., Tucker, M. A., Caporaso, N., … Fraumeni, J. F. (2010). Second malignancy risks after non-Hodgkin's lymphoma and chronic lymphocytic leukemia: differences by lymphoma subtype. *Journal of Clinical Oncology*, JCO–2010.

Moyer, V. A. (2014). Screening for lung cancer: US Preventive Services Task Force recommendation statement. *Annals of Internal Medicine*.

Mudie, N. Y., Swerdlow, A. J., Higgins, C. D., Smith, P., Qiao, Z., Hancock, B. W., … Linch, D. C. (2006). Risk of second malignancy after non-Hodgkin's lymphoma: a British Cohort Study. *Journal of Clinical Oncology*, *24*(10), 1568–1574.

National Lung Screening Trial Research Team. (2011). The National Lung Screening Trial: overview and study design. *Radiology*, *258*(1), 243–253. doi:10.1148/radiol.10091808/-/DC1

Neugut, A. I., Lee, W. C., Murray, T., Robinson, E., Karwoski, K., & Kutcher, G. J. (1993). Lung cancer after radiation therapy for breast cancer. *Cancer*, *71*(10), 3054–3057.

Neugut, A. I., Murray, T., Santos, J., Amols, H., Hayes, M. K., Flannery, J. T., & Robinson, E. (1994). Increased risk of lung cancer after breast cancer radiation therapy in cigarette smokers. *Cancer*, *73*(6), 1615–1620.

Nieder, C., & Baumann, M. (2011). Fractionation Concepts. In C. Nieder & J. A. Langendijk (Eds.), *Re-irradiation: New Frontiers SE - 77* (pp. 13–26). Springer Berlin Heidelberg. doi:10.1007/174_2010_77

Nugent, Z., Demers, A. A., Wiseman, M. C., Mihalcioiu, C., & Kliewer, E. V. (2005). Risk of second primary cancer and death following a diagnosis of nonmelanoma skin cancer. *Cancer Epidemiology Biomarkers & Prevention*, *14*(11), 2584–2590.

Oken, M. M., Hocking, W. G., Kvale, P. A., Andriole, G. L., Buys, S. S., Church, T. R., … Reding, D. J. (2011). Screening by chest radiograph and lung cancer mortality. *JAMA: The Journal of the American Medical Association*, *306*(17), 1865–1873.

Oken, M. M., Marcus, P. M., Hu, P., Beck, T. M., Kvale, P. A., Cordes, J., … Gohagan, J. K. (2005). Baseline Chest Radiograph for Lung Cancer Detection in the Randomized Prostate , Lung , Colorectal and Ovarian Cancer Screening Trial, *97*(24). doi:10.1093/jnci/dji430

Oliver, M. C. (1994). Social, behaviouraland medical factors in the aetiology of testicular cancer: results from the UK study. *Br. J. Cancer*, *70*, 513–520.

Parikh-Patel, A., Allen, M., & Wright, W. E. (2003). Validation of self-reported cancers in the California Teachers Study. *American Journal of Epidemiology*, *157*(6), 539–545.

Pawlish, K. S., Schottenfeld, D., Severson, R., & Montie, J. E. (1997). Risk of multiple primary cancers in prostate cancer patients in the Detroit metropolitan area: a retrospective cohort study. *The Prostate*, *33*(2), 75–86.

Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E., & Doll, R. (2000). Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *Bmj*, *321*(7257), 323–329.

Phipps, A. I., Chan, A. T., & Ogino, S. (2013). Anatomic subsite of primary colorectal cancer and subsequent risk and distribution of second cancers. *Cancer*, *119*(17), 3140–3147.

Pickles, T., & Phillips, N. (2002). The risk of second malignancy in men with prostate cancer treated with or without radiation in British Columbia, 1984–2000. *Radiotherapy and Oncology*, *65*(3), 145–151.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 705–724.

Preston, D. L., Mattsson, A., Holmberg, E., Shore, R., Hildreth, N. G., & Boice Jr, J. D. (2009). Radiation effects on breast cancer risk: a pooled analysis of eight cohorts.

Prorok, P. C., Andriole, G. L., Bresalier, R. S., Buys, S. S., Chia, D., David Crawford, E., … Hasson, M. A. (2000). Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials*, *21*(6), 273S–309S.

Raymond, J. S., & Hogue, C. J. R. (2006). Multiple primary tumours in women following breast cancer, 1973–2000. *British Journal of Cancer*, *94*(11), 1745–1750.

Reynolds, S. (2012). Crunching Numbers: What Cancer Screening Statistics Really Tell Us. *NCI Cancer Bulletin*, *9*(23).

Richardson, G. E., Tucker, M. A., Venzon, D. J., Linnoila, R. I., Phelps, R., Phares, J. C., … Johnson, B. E. (1993). Smoking cessation after successful treatment of small-cell lung cancer is associated with fewer smoking-related second primary cancers. *Annals of Internal Medicine*, *119*(5), 383–390.

Ritoe, S. C., Krabbe, P. F. M., Jansen, M. M. G., Festen, J., Joosten, F., Kaanders, J., … Marres, H. A. M. (2002). Screening for second primary lung cancer after treatment of laryngeal cancer. *The Laryngoscope*, *112*(11).

Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, 429–467.

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, *25*(1), 127–141.

Royston, P., Moons, K. G., Altman, D. G., & Vergouwe, Y. (2009). Prognosis and prognostic research: Developing a prognostic model. *Bmj*, *338*(mar31_1), b604–b604.

Rubino, C., De Vathaire, F., Dottorini, M. E., Hall, P., Schvartz, C., Couette, J. E., … Schlumberger, M. (2003). Second primary malignancies in thyroid cancer patients. *British Journal of Cancer*, *89*(9), 1638–1644.

Sawka, A. M., Thabane, L., Parlea, L., Ibrahim-zada, I., Tsang, R. W., Brierley, J. D., … Goldstein, D. P. (2009). Second Primary Malignancy Risk After Radioactive Iodine Treatment for Thyroid Cancer : A Systematic Review and Meta-analysis, *19*(5).

Scélo, G., Boffetta, P., Corbex, M., Chia, K.-S., Hemminki, K., Friis, S., … Tracey, E. (2007). Second primary cancers in patients with nasopharyngeal carcinoma: a pooled analysis of 13 cancer registries. *Cancer Causes & Control*, *18*(3), 269–278.

Schonfeld, S. J., Curtis, R. E., Anderson, W. F., & de Gonzalez, A. B. (2012). The risk of a second primary lung cancer after a first invasive breast cancer according to estrogen receptor status. *Cancer Causes & Control*, *23*(10), 1721–1728.

Schrijvers, C. T. M., Stronks, K., van de Mheen, D. H., Coebergh, J.-W. W., & Mackenbach, J. P. (1994). Validation of cancer prevalence data from a postal survey by comparison with cancer registry records. *American Journal of Epidemiology*, *139*(4), 408–414.

Secretan, B., Straif, K., Baan, R., Grosse, Y., El Ghissassi, F., Bouvard, V., … Galichet, L. (2009). A review of human carcinogens—Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. *The Lancet Oncology*, *10*(11), 1033–1034.

Shiels, M. S., Gibson, T., Sampson, J., Albanes, D., Andreotti, G., Freeman, L. B., … Elena, J. (2014). Cigarette Smoking Prior to First Cancer and Risk of Second Smoking-Associated Cancers Among Survivors of Bladder, Kidney, Head and Neck, and Stage I Lung Cancers. *Journal of Clinical Oncology*, *32*(35), 3989–3995.

Slaughter, D. P., Southwick, H. W., & Smejkal, W. (1953). "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. *Cancer*, *6*(5), 963–968.

Sonett, J. R., Suntharalingam, M., Edelman, M. J., Patel, A. B., Gamliel, Z., Doyle, A., … Krasna, M. (2004). Pulmonary resection after curative intent radiotherapy (> 59 Gy) and concurrent chemotherapy in non–small-cell lung cancer. *The Annals of Thoracic Surgery*, *78*(4), 1200–1205.

Song, F., Qureshi, A. A., Giovannucci, E. L., Fuchs, C. S., Chen, W. Y., Stampfer, M. J., & Han, J. (2013). Risk of a second primary cancer after non-melanoma skin cancer in white men and women: a prospective cohort study. *PLoS Medicine*, *10*(4), e1001433.

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, *5*(5), 421–433.

Spitz, M. R., Etzel, C. J., Dong, Q., Amos, C. I., Wei, Q., Wu, X., & Hong, W. K. (2008). An expanded risk prediction model for lung cancer. *Cancer Prevention Research*, *1*(4), 250–254.

Spitz, M. R., Hong, W. K., Amos, C. I., Wu, X., Schabath, M. B., Dong, Q., … Etzel, C. J. (2007). A Risk Model for Prediction of Lung Cancer, 715–726. doi:10.1093/jnci/djk153

Steyerberg, E. W. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation and Updating*. Rotterdam: Springer.

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, *21*(1), 128.

Sturgis, E. M., & Miller, R. H. (1995). Second primary malignancies in the head and neck cancer patient. *The Annals of Otology, Rhinology, and Laryngology*, *104*(12), 946–954.

Szumilas, M. (2010). Explaining odds ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *19*(3), 227.

Tabachnick, B. G., & Fidell, L. S. (2007). Multivariate analysis of variance and covariance. *Using Multivariate Statistics*, *3*, 402–407.

Tabuchi, T., Ito, Y., Ioka, A., Miyashiro, I., & Tsukuma, H. (2012). Incidence of metachronous second primary cancers in Osaka, Japan: Update of analyses using population-based cancer registry data. *Cancer Science*, *103*(6), 1111–1120.

Tabuchi, T., Ito, Y., Ioka, A., Nakayama, T., Miyashiro, I., & Tsukuma, H. (2013). Tobacco smoking and the risk of subsequent primary cancer among cancer survivors: a retrospective cohort study. *Annals of Oncology*, *24*(10), 2699–2704.

Tammemagi, M. C., Church, T. R., Hocking, W. G., Silvestri, G. A., Kvale, P. A., Riley, T. L., … Berg, C. D. (2014). 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*, *11*(12), e1001764.

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

Tammemagi, M. C., & Lam, S. (2013). Screening for lung cancer using low dose computed tomography. *BMJ (Clinical Research Ed.)*, *348*, g2253–g2253.

Tammemagi, M. C., Pinsky, P. F., Caporaso, N., Kvale, P. A., Hocking, W. G., Church, T. R., … Berg, C. D. (2011). Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. *Journal of the National Cancer Institute*, *103*(13), 1058–1068.

Teppo, L., Pukkala, E., & Saxén, E. (1985). Multiple cancer—an epidemiologic exercise in Finland. *Journal of the National Cancer Institute*, *75*(2), 207–217.

Thellenberg, C., Malmer, B., Tavelin, B., & Grönberg, H. (2003). Second primary cancers in men with prostate cancer: an increased risk of male breast cancer. *The Journal of Urology*, *169*(4), 1345–1348.

Thompson, D. E., Mabuchi, K., Ron, E., Soda, M., Tokunaga, M., Ochikubo, S., … Izumi, S. (1994). Cancer incidence in atomic bomb survivors. Part II: Solid tumors, 1958-1987. *Radiation Research*, *137*(2s), S17–S67.

Travis, L. B. (2006). The epidemiology of second primary cancers. *Cancer Epidemiology Biomarkers & Prevention*, *15*(11), 2020–2026.

Travis, L. B., Fosså, S. D., Schonfeld, S. J., McMaster, M. L., Lynch, C. F., Storm, H., … Pukkala, E. (2005). Second cancers among 40 576 testicular cancer patients: focus on long-term survivors. *Journal of the National Cancer Institute*, *97*(18), 1354–1365.

Travis, L. B., Gospodarowicz, M. K., Curtis, R. E., Clarke, E. A., Andersson, M., Glimelius, B., … Holowaty, E. (2002). Lung cancer following chemotherapy and radiotherapy for Hodgkin's disease. *Journal of the National Cancer Institute*, *94*(3), 182–192.

Tucker, M. A., Murray, N., Shaw, E. G., Ettinger, D. S., Mabry, M., Huber, M. H., … Grant, S. C. (1997). Second primary cancers related to smoking and treatment of small-cell lung cancer. *Journal of the National Cancer Institute*, *89*(23), 1782–1788.

U.S. Cancer Statistics Working Group. (2015). *United States Cancer Statistics: 1999-2010 Incidence and Mortality Web-based Report*. Atlanta. Retrieved from www.cdc.gov/uscs

UK National Collaborating Centre for Cancer. (2009). Early and Locally Advanced Breast Cancer.

Utada, M., Ohno, Y., Hori, M., & Soda, M. (2014). Incidence of multiple primary cancers and interval between first and second primary cancers. *Cancer Science*, *105*(7), 890–896.

Van Leeuwen, F. E., Klokman, W. J., Stovall, M., Hagenbeek, A., van den Belt-Dusebout, A. W., Noyon, R., … Somers, R. (1995). Roles of radiotherapy and smoking in lung cancer following Hodgkin's disease. *Journal of the National Cancer Institute*, *87*(20), 1530–1537.

Verkooijen, R. B. T., Smit, J. W. A., Romijn, J. A., & Stokkel, M. P. M. (2006). The incidence of second primary tumors in thyroid cancer patients is increased, but not related to treatment of thyroid cancer. *European Journal of Endocrinology*, *155*(6), 801–806.

Vittinghoff, E., Glidden, D. V, Shiboski, S. C., & McCulloch, C. E. (2011). *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. Springer Science & Business Media.

Volk, N., & Pompe-Kirn, V. (1997). Second primary cancers in breast cancer patients in Slovenia. *Cancer Causes & Control*, *8*(5), 764–770.

Weisburd, D., & Britt, C. (2014). *Statistics in criminal justice*. Springer Science & Business Media.

Welch, H. G., Birkmeyer, J. D., Schwartz, L. M., Black, W. C., & Woloshin, S. (1999). Finding and redefining disease. *Effective Clinical Practice*, *2*(2), 96–99.

Wheless, L., Black, J., & Alberg, A. J. (2010). Nonmelanoma skin cancer and the risk of second primary cancers: a systematic review. *Cancer Epidemiology Biomarkers & Prevention*, *19*(7), 1686–1695.

Wood, D. E., Eapen, G. A., Ettinger, D. S., Hou, L., Jackman, D., Kazerooni, E., … Leung, A. N. C. (2012). Lung cancer screening. *Journal of the National Comprehensive Cancer Network*, *10*(2), 240–265.

Yamamoto, E., Shibuya, H., Yoshimura, R., & Miura, M. (2002). Site specific dependency of second primary cancer in early stage head and neck squamous cell carcinoma. *Cancer*, *94*(7), 2007–2014.

Youlden, D., & Baade, P. (2011). The relative risk of second primary cancers in Queensland, Australia: a retrospective cohort study. *BMC Cancer*, *11*(1), 83.

Zablotska, L. B., & Neugut, A. I. (2003). Lung carcinoma after radiation therapy in women treated with lumpectomy or mastectomy for primary breast carcinoma. *Cancer*, *97*(6), 1404–1411.

Zhou, B., Yang, L., Sun, Q., Cong, R., Gu, H., Tang, N., … Wang, B. (2008). Cigarette smoking and the risk of endometrial cancer: a meta-analysis. *The American Journal of Medicine*, *121*(6), 501–508.

Zhu, G., Chen, Y., Zhu, Z., Lu, L., Bi, X., Deng, Q., … Guo, H. (2012). Risk of second primary cancer after treatment for esophageal cancer: a pooled analysis of nine cancer registries. *Diseases of the Esophagus*, *25*(6), 505–511.