MOLECULAR EPIDEMIOLOGY OF LUNG CANCER IN THE LIVERPOOL LUNG PROJECT (LLP) COHORT

Thesis submitted in accordance with the requirements of the

University of Liverpool for the degree of Doctor in Philosophy

by

Nicosha Belinda De Souza

July 2014

Molecular epidemiology of lung cancer in the Liverpool lung project (LLP) cohort

Summary: The primary aim of the project was to evaluate the epidemiological and genetic susceptibility factors associated with lung cancer, in the Liverpool Lung Project (LLP) population. The associated datasets available for research with the LLP dataset (questionnaire) were: Office of National Statistics (ONS), Health Episode Statistics (HES) data with comorbidity data, single nucleotide polymorphism (SNP) data of 570 cases from Liverpool, 3000 controls from the 1958 Birth Cohort.

The epidemiological (HES) data was used to study the effect of Charlson (CCI) and Elixhauser comorbidity index (ECI) on the incidence of lung cancer using the Cox proportional hazard regression and use the same HES data to design a 5-year sex specific incidence model for lung cancer with crucial covariates. The ECI and CCI were significant in both univariate and multivariate analyses adjusted for age at the start of the study, sex and smoking pack years. The developed models had a good discriminatory power (AUC_{male} = 0.73; AUC_{female} = 0.77) when internally validated using a 10-fold cross validation.

The genetic data for the LLP lung cancer cases was used in several contexts: i) to identify SNPS associated with lung cancer under a range of allelic models (additive, dominant, recessive and genotypic), using the Wellcome trust 1958 Birth Cohort as a control dataset; ii) to identify SNPs associated with cause specific and overall survival in lung cancer patients, utilising the Cox proportional hazard model with adjustment for various covariates; and iii) to identify gene pathways that are associated with lung cancer survival using the random forest survival method.

SNPs within the genes *PRDM11, ZNF382 and HMGA2* were identified in the genome wide case-control study when using the additive, dominant or genotypic models, whereas the recessive model identified the gene *ITIH2*.

Significant SNPs ($p \leq 10^{-6}$) associated with cause-specific survival in early stage cases were rs10230420 (*WIPF3*), rs3746619 and rs3827103 (both in *MC3R*). In advanced stage cases, significant SNPs were rs1868110 (*NEK10*) and rs2206779 (*AF357533*). For the overall survival analysis, significant SNPs were rs10230420 (*WIPF3*), rs2056533 (*ZBTB20*) and rs6708630 (*CYS1*) in early stage cases, whereas rs1868110 (*NEK10*) and rs2206779 (*AF357533*) were significantly associated with overall survival in advanced stage NSCLC cases.

The pathway analysis using the random survival forest method was undertaken on 18 pathways for both cause-specific and overall survival of lung cancer cases. The results were consistent with apoptosis, base excision repair and mismatch repair being pathways influencing survival.

ACKNOWLEDGEMENT

I would like to thank my supervisors, Prof. John Field and Dr. Russell Hyde for their guidance during my research. I would also like to thank Dr. Fiona McRonald and Dr. Michael Marcus for their insight and suggestions during the writing process of the thesis. I would also like to sincerely thank the Liverpool primary Care Trust for funding this studentship. And finally, I am grateful to my family for their constant support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

A

A	Adenine
ADC	Adenocarcinoma
AFB	Autofluorescence Bronchoscopy
AIDS	Acquired immunodeficiency syndrome
Arg	Arginine
ARTP	Adaptive Rank Truncated Product
ASR	Age standardised rate
AUC	Area under the receiver operating curve

B

BER	Base excision repair
BMI	Body mass index
BOD	Burden of Disease
bRSF-LR	Bivariate random survival forest with log-rank split

C

C	Cytosine
CAD	Coronary artery disease
CARET	The Beta-Carotene and Retinol Efficacy Trial
CCI	Charlson comorbidity index
CCIS	Charlson comorbidity index score
cDNA	Complementary deoxyribonucleic acid
CETO	Central European and Toronto
CEU	Utah residents with Northern and Western European ancestry from CEPH collection
CHF	Congestive heart failure
CHF	Cumulative hazard function
CIRS	Cumulative Illness Rating Scale
CNV	Copy number variation
COPD	Chronic obstructive pulmonary disease
CPD	Chronic pulmonary disease
CTD	Connective tissue disease
cRNA	Complementary ribonucleic acid
CT	Computerised tomography
CVD	Cerebrovascular disease

D

DM	Diabetes mellitus
DNA	Deoxyribonucleic acid

E

EBC	Exhaled breath condensate
EBUS	Endobronchial ultrasound
ECI	Elixhauser comorbidity index
ECIS	Elixhauser comorbidity index score
ECM	Extracellular matrix
EMT	Epithelial–Mesenchymal Transition
ERT	Estrogen replacement therapy
ETS	Environmental tobacco smoke

F

FDR	False discovery rate
FISH	Fluorescence in situ hybridization
FPRD	False positive report probability

G

G	Guanine
GDP	Guanosine diphosphate
GELCC	Genetic epidemiology of lung cancer consortium
Gln	Glutamine
GO	Gene ontology
GPRD	General practice research database
GRMD	Germany and Texas
GSEA	Gene set enrichment analysis
GST	Glutathione S-transferase
GSTM1	Glutathione s-transferase mu 1
GSTP1	Glutathione s-transferase pi 1
GSTT1	Glutathione S-transferase theta 1
GTP	Guanosine-5'-triphosphate
GWAS	Genome wide association studies

H

HES	Health episode statistics
HIV	Human immunodeficiency virus
HMM	Hidden Markov Model
HR	Hazard ratio
HWE	Hardy Weinberg equilibrium

I

IARC	International Agency for Research on Cancer
IBD	Identity by descent
IBS	Identity by state
ICD	International Classification of Diseases

Ile	Isoleucine

K

KEGG	Kyoto Encyclopaedia of Genes and Genomes

L

LCC	Large cell carcinoma
LD	Linkage disequilibrium
LCL	Lymphoblastic cell lines
LIFE	Light induced fluorescence endoscopy
LLP	Liverpool lung project
LLPC	Liverpool lung project cohort
LOH	Loss of heterozygosity
LR	Log-rank
LRS	Log-rank score
Lys	Lysine

M

MAF	Minor allele frequency
MCMC	Monte Carlo Markov Chain
MD	Monroe Dunaway
MDS	Multidimensional scaling
MI	Myocardial infarction
miRNA	Micro ribonucleic acid
MMR	Mismatch repair
mRNA	Messenger ribonucleic acid

N

NBI	Narrow band imaging
NCBI	National Centre for Biotechnology Information
NER	Nucleotide repair
NF-κβ	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGS	Next generation sequencing
NHS	National health service
NLST	National lung screening trial
NNK	4-(methylnitro-samino)-1-(3-pyridyl)-1-butanone
NSCLC	Non-small cell lung cancer

O

OCT	Optical coherence tomography
ONS	Office of national statistics
OOB	Out of bag
OR	Odds ratio

P

PAH	Polycyclic aromatic hydrocarbons
PCA	Principal component analysis
PCPT	Prostate cancer prevention trial
Pro	Proline
PSA	Prostate specific antigen
PVD	Peripheral vascular disease

R

RF	Random forest
ROCA	Risk of ovarian cancer algorithm
RSF	Random survival forest

S

SCLC	Small cell lung cancer
SNP	Single nucleotide polymorphism
SOLiD	Supported Oligonucleotide Ligation and Detection
SQC	Squamous cell carcinoma

T

T	Thymine
TB	Tuberculosis
TGF-α	Transforming growth factor alpha
TGFβ	Transforming growth factor beta
TKI	Tyrosine kinase inhibitor
TNF	Tumour necrotic factor
TRF	Two staged random forest
TSG	Tumour suppressor gene
TSNA	Tobacco-specific nitrosamines

U

UK	United Kingdom
US	United States
UTR	Untranslated region

V

Val	Valine

CHAPTER 1

INTRODUCTION

1.1 Lung Cancer Incidence and Mortality

Lung cancer was the most common cancer and the leading cause of cancer deaths in the world, in 2008[1]. In the same year, about 1.61 million new lung cancer cases were reported worldwide, representing 12.7% of all newly diagnosed cancer cases[1]. The incidence in 2008 was higher among males than females, accounting for 16.5% of all cancers in males and 8.5% of all cancers in females (worldwide figures)[1]. Within England, there are regional differences in lung cancer incidence, with rates of 59.7 per 100,000 for the whole of England, but 80.7 per 100,000 for Merseyside and Cheshire (2003-2005 figures) (UK Lung Cancer Coalition Commissioning Communications Toolkit - Merseyside and Cheshire Cancer Network)[2].

Lung cancer caused 1.38 million deaths worldwide, accounting for 18.2% of total cancer deaths (2008 figures); with higher mortality in males[1] and in 2010 the number of deaths increased to 1.5 million for trachea, bronchus and lung cancer[3]. Worldwide figures for 2008 indicate that 22.5% of male cancer deaths and 12.8% of female cancer deaths were due to lung cancer[1]. The standardised mortality ratio (SMR) for lung cancer from 2004-2006 for all ages was 186 in Liverpool, compared to 124 for the North West of England (Compendium of Clinical & Health Indicators – January 2008 release)[4].

1.2 Trends in Histopathological Types of Lung Cancer

The WHO histological typing of lung and pleural tumours (1999) recognises four major histological types of lung cancer, divided into two categories: small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC)[5-7]. The latter is further divided into lung

adenocarcinoma; squamous cell carcinoma and large cell carcinoma[5-7]. The classification of lung cancer has been updated in 2013[8].

Squamous cell carcinoma, small cell carcinoma and adenocarcinoma are common histological types associated with smoking, with adenocarcinoma showing a strong link to smoking[5, 6]. Nonetheless, adenocarcinoma is still the most common histological subtype in non-smokers[5]. Recent data have shown an increase in the incidence of adenocarcinoma and a corresponding decrease in the incidence of squamous cell carcinoma, a trend postulated to be due to the changes in cigarette composition and design (decreased nicotine/tar content and incorporation of cigarette filters)[5].

1.3 Causative Mechanisms Underlying Lung Cancer

Understanding biological processes leading to lung cancer is crucial in early detection and prevention research[9]. Aside from tobacco smoking, inflammation arising from various factors is an important mechanism and therefore understanding its role with relation to various carcinogens will shed light on this field[10]. Other hypothesized mechanisms are also noted.

Inflammation: The role of inflammation in lung cancer was reported in 1863, by Rudolf Virchow, who noted increased cell proliferation in inflamed tissue[11]. It is also seen in response to various injuries, infections and chemical or particle exposure[11]. A minor exposure such as an agent triggers an acute inflammatory response, resulting in the elimination of the agent and restoration of the affected site to its initial condition while a prolonged exposure would give rise to chronic inflammation and tissue damage[11].

Inflammation results in leukocyte production as a response to damaged tissue, leading to a complex process involving macrophages, cytokines such as IL1β and TNF, prostaglandins, etc. The inflammatory process results in the production of reactive oxygen or nitrogen that binds to DNA, resulting in its alteration. This process can either initiate or promote carcinogenesis[11]. Tobacco smoke and respiratory infections might be two of the carcinogenic agents in inflammations[11]. Furthermore, subsequent tissue repair processes linked to cellular proliferation involving DNA replication may lead to further DNA variations[11].

Two studies identified the association of chronic inflammation with lung cancer: one, an observational study carried out on 7,081 patients followed up for 10 years, and the other on 6,273 patients, for which inflammation was measured by C reactive protein quantity[12, 13].

Field Carcinogenesis theory: Tobacco smoke is the major environmental factor in lung cancer causation[14]. The carcinogens in tobacco smoke may cause multifocal premalignant lesions in the respiratory epithelium of the bronchial tree[15]. This is referred to as the field cancerisation effect, and refers to the ability of tobacco smoke to cause mutagenesis within respiratory epithelial cells[10].

Genetic changes have been observed at multiple focal points in the respiratory epithelium of former and current smokers as well as lung cancer patients[10]. These variations may last for many years after smoking cessation, and may be a cause of lung cancer in former smokers[10]. Studies have indicated that smoking causes field abnormalities in histologically normal lung cells, and gene expression studies of these normal cells could serve as a crucial biomarker in lung cancer studies[10].

Epithelial-mesenchyme transition: Developmental alterations leading to a change from an epithelial tissue-phenotype to a mesenchymal phenotype is referred to as epithelial mesenchymal transition (EMT) and is seen in embryonic development, chronic inflammation and fibrosis[12]. This process is crucial in carcinogenesis and is characterised by altered morphology, adhesion and migration capacity, anti-apoptosis, and increased expression of N-cadherin and vimentin; these can serve as important biomarkers in cancer studies[12].

EMT in tumourigenesis is an unregulated process, distinct from the normal transition[10]. Pathways influencing this process (including TGF-beta, PI3k/AKT, Ras signalling, Wnt) play a vital role in many malignancies including lung cancer[10].

EMT is promoted through the induction of Zinc finger transcriptional repressors of E-cadherin such as Zeb1, Snail and Slug, that not only promote invasion and metastasis, but also aid in the elimination of pre-cancerous cells to distant locations even before the actual cancer has progressed, suggesting its role in early lung cancer development and metastasis[10]. Tobacco exposure and chronic inflammation are processes that drive EMT, leading to malignancies[16]. For instance, the induction of EMT related genes by benzo-[a]-pyrene and promotion of EMT by NNK [4-(n-methyl-n-nitroamino)-1-(3-pyridyl)-1-butanone], in lung cancer cells[10].

Proliferative growth is enabled through a mesenchymal to epithelial transition[17]. EMT is associated with invasion and metastasis, early changes in lung cancer, acquiring stem cell like properties, cell death, senescence and conventional chemotherapy resistance[17].

1.4 Risk Factors for Lung Cancer

Risk factors for lung cancer range from exposure to carcinogens to previous respiratory disease[9, 18]. Other occupational and domestic exposures (eg, chemical compounds and solvents, paints, thinners and welding equipment) have been reported to increase lung cancer risk[19, 20]. Detailed occupational, dietary and hormonal factors implicated in lung cancer are tabulated below (Table 1). Following are some factors that have a major role in contributing to lung cancer risk.

Table 1.1: Important literature on lung cancer occupational, dietary and hormonal risk factors.

AUTHOR	AIM	POPULATION ETHNICITY; METHOD	RESULTS AND SUMMARY	CONCLUSION
Brenner *et al*., 2010[20]	To evaluate the risk factors associated with lung cancer in non-smokers.	**Caucasian**; Study conducted on 445 cases which were frequency matched on sex and ethnicity to 425 population controls and 523 hospital controls. Unconditional logistic regression was used to calculate the OR and the 95% CI to establish the association between risk factors and lung cancer.	There was a significant association between exposure to carcinogenic agents and lung cancer risk (OR = 1.6, 95% CI: 1.4-2.1) in the total population and in non-smokers (OR = 2.1, 95% CI: 1.3- 3.3). In never smokers, exposure to **solvents, paints or thinners** (OR = 2.8, 95% CI: 1.6-5.0); **welding equipment** (OR = 3.4, % CI: 1.1-10.4); **smoke, soot or exhaust** (OR =2.8, 95% CI: 1.4-5.3) were significant but not in the total population. **Emphysema** (OR = 4.8, 95 % CI: 2.0-11.1) was significantly associated with lung cancer risk in the total population. Asthma, chronic bronchitis, pneumonia, TB did not appear significant. Having a relative with lung cancer, >50 years of age (OR = 1.8, 95% CI: 1.0-3.2) was significantly associated with increased risk of lung cancer in non-smokers.	Occupational exposure depicted an essential impact on the risk of lung cancer. Family history also plays an important role in defining lung cancer risk.
Hosseini *et al*., 2009[19]	To evaluate the environmental risk factors associated with the risk of lung cancer in Iran.	**Asian**; Study conducted on 242 histologically confirmed cases and frequency matched for age, sex and place of residence to 242 hospital controls and	In the bivariate analyses, exposure to **heavy metals** (OR = 2.9, 95% CI: 1.4–5.9) , **inorganic dust** (OR = 4.2, 95% CI: 2.9–6.1), **chemical weapon exposure** (OR = 30.9, 95% CI: 1.8–542), **chemical**	Occupational exposure to various chemical agents and cigarette

		242 visiting healthy controls. Association was evaluated using unconditional logistic regression.	**compositions** (OR = 4.3, 95% CI: 2.8–6.6), **smoking** (OR = 4.7, 95% CI: 3.0–7.2) and **opium** (OR = 2.2 , 95% CI: 1.4–3.6) were significantly associated with lung cancer risk. Exposure to **inorganic dust** (OR = 4.9, 95% CI: 1.4–17.4) and **chemical compositions** (OR =5.1, 95% CI: 1.4–18.5) were significant in **non-smokers**. In the multivariate analyses, **cigarette smoking** (OR= 5.4, 95% CI: 3.2–8.9), exposure to **inorganic dust** (OR= 4.2, 95% CI: 2.8–6.7), **chemical compounds** (OR= 3.4, 95% CI: 2.1–5.6), **heavy metals** (OR =3.0, 95% CI: 1.3–7.0) were independent risk factors.	smoking are important risk factors in lung cancer.
Gao *et al*., 2009[21]	To investigate the role of family history and non-malignant lung diseases in the overall lung cancer risk.	**Caucasian** (Italian); Family history of smoking and histology on 1946 cases and 2116 controls were available. OR and 95% CI were calculated using the logistic regression adjusting for age, gender, residence, education and cigarette smoking.	**History of lung cancer** in father (OR =1.37; 95% CI: 1.01–1.87), mother (OR = 2.21; 95% CI: 1.11–4.41) sibling (OR = 1.53; 95% CI: 1.10–2.12) and overall (OR = 1.57; 95% CI: 1.25–1.98) was associated with increased risk. The association was stronger in younger members (OR = 3.26; 95% CI: 1.55–6.85), never smokers, adenocarcinoma (OR = 1.68; 95% CI: 1.28–2.20) and squamous cell carcinoma (OR =1.79; 95% CI: 1.25–2.55) subtypes. **History of bronchitis** in any family member and lung cancer risk was stronger in subjects <55 years (OR=1.76; 95% CI: 5	Family history of lung cancer and non-malignant lung disease affect the risk of lung cancer independently.

			1.003–3.08) than ≥55 years (OR = 1.48; 95% CI: 1.21–1.81). Similar results were observed for **emphysema** for subjects <55(OR = 1.34, 95% CI: 0.61–2.96) years and ≥55 years (OR = 1.18, 95% CI: 0.94–1.51).Protective effect was seen for the association with family history of **pneumonia**, stronger for ≥55 years (OR=0.78; 95% CI: 0.45–1.34) than <55 years (OR =0.71; 95% CI: 0.59–0.87).	
Mahabir *et al*., 2008[22]	To assess dietary magnesium and DNA repair capacity (DRC) as risk factors for lung cancer.	**Caucasian; Hispanic and African American**; 1139 cases and 1210 matched controls were used for this study. Multiple logistic regression analysis was used to calculate ORs and 95% CIs for associations between dietary Mg and lung cancer, adjusting for age, gender, race, smoking status, pack-years smoked family history of cancer, BMI, education, income, total calories and DRC.	The interaction between **Mg intake** and **DRC** was significant (p<0.0001). Joint analysis was carried out that compared high dietary intake of Mg and proficient DNA repair capacity with low dietary intake of Mg and suboptimal DRC produced an OR = 2.36 and 95% CI: 1.83-3.04. Within the low Mg intake and suboptimal DRC group, the risk was more pronounced in older subjects (OR= 3.0; 95% CI: 2.13–4.23), lower BMI(>25) (OR =2.77; 95% CI: 1.82–4.23), current smokers (OR =3.88; 95% CI: 2.46– 6.14), those with longer duration of smoking (OR = 2.90; 95% CI: 2.00–4.20) and heavy smokers (OR = 2.73; 95% CI: 1.79–4.15), small cell lung cancer (OR = 3.30; 95% CI: 1.69–6.46).	Increased dietary intake of Mg was associated with decrease in risk of lung cancer ranging from 17 to 53%.
Mahabir *et al*., 2008[23]	To evaluate the role of dietary boron intake	**Caucasian; Hispanic and African American;** 763 women	The OR for lung cancer increased with decreasing quartile for **boron** intake	Increased dietary intake of

	and hormonal replace therapy (HRT) in lung cancer risk.	with lung cancer and 838 healthy controls matched for diet and HRT were recruited. Logistic regression was used to calculate the OR and 95% CI adjusting for age, ethnicity, education, BMI, alcohol, years of smoking, number of cigarettes, use of vitamin/mineral supplements and family history in first degree relatives.	with a significant trend ($p<0.0001$). Joint analyses, the low dietary boron intake and no HRT showed an increased risk (OR = 2.07; 95% CI: 1.53-2.81) when compared with women with high boron intake and used HRT. For the low Boron intake and no HRT the risk of lung cancer adjusted for age >60 (OR =2.32; 95% CI: 1.50- 3.60), ≤60 (OR =1.84 ; 95% CI: 1.15-2.94); BMI >25 (OR = 1.99; 95% CI: 1.33- 2.98); BMI ≤25 (OR = 2.00; 95% CI: 1.24- 3.23) and smoking years >31 (OR = 2.26 ; 95% CI: 1.39 - 3.66) and smoking years ≤31(OR = 1.91; 95% CI: 1.27- 2.88).	boron reduced the risk of lung cancer. Furthermore, women who intake boron and used HRT are at a lower risk than those who are on low boron intake and no HRT.
Neuberger *et al*., 2006[24]	To evaluate the risk factors associated with lung cancer in Iowa women with respect to smoking habits.	**Caucasian**; 413 female lung cancer cases and 614 controls resident for at least 20 years were included in the study. Logistic regression was conducted to derive the OR and 95% CI after adjusting for age, education and cumulative radon exposure.	For unadjusted logistic analyses, association was seen for ever smoked (OR= 13.20; 95% CI: 9.50–18.33), current smokers compared to never smokers (OR = 25.98; 95% CI: 17.72–38.09), **family history of kidney** (OR = 2.95; 95% CI: 1.41–6.20), **family history of bladder** (OR =2.01; 95% CI: 1.04–3.91) and **family history of lung cancer** (OR =1.71; 95% CI: 1.25–2.35), **pre-existing bronchitis and emphysema**(OR =3.53; 95% CI: 2.45–5.08). After adjusting for radon exposure, age and education, current smokers (OR = 13.92, 95% CI: 7.40–26.18), ex-smokers	Active cigarette smoking is an important risk factor for lung cancer. Family history of smoking related cancers is an important factor in lung cancer.

			(OR = 13.47; 95% CI: 5.17–35.12), **asbestos exposure** (OR = 3.39; 95% CI: 1.18–9.75), **family history of bladder cancer** (OR = 3.08; 95% CI: 1.26–7.57), **family history of kidney cancer** (OR = 3.04; 95% CI: 1.13–8.18) were significant. Among current smokers, **family history of lung cancer** (OR = 2.43; 95% CI: 1.12–5.28) was significant. For ex-smokers, years since quitting (OR = 0.93; 95% CI: 0.88–0.98) was significantly protective and among never smokers, family history of kidney cancer (OR = 7.34; 95% CI: 1.91–28.18), **family history of bladder cancer** (OR = 5.02; 95% CI: 1.64–15.39) and **history of lung disease** (OR = 2.28; 95% CI: 1.24– 4.18) was significant.	
Kreuzer *et al*., 2003[25]	To evaluate the role of endocrine factors in the risk of lung cancer accounting for smoking status and histology.	**Caucasian** (Germany); Study conducted on histologically confirmed 811 lung cancer cases and 912 controls. Logistic regression was used to compute the OR and 95% CI adjusting for age, region, smoking and education.	The use of **oral contraceptives** (OR = 0.69; 95% CI: 0.51–0.92) and use of **hormonal replacement therapy** (OR = 0.83; 95% CI: 0.64–1.09), especially after long term use (≥ 7 years) (OR = 0.59; 95% CI: 0.37–0.93) depicted a decrease in lung cancer risk.	In women who smoke, the use of exogenous hormones depicted a reduced risk of lung cancer indicating the role of hormonal factors in the aetiology of lung

				cancer.
Takezaki *et al*., 2001[26]	To evaluate the influence of diet on lung cancer risk.	**Asian** (Japan); 367 male and 240 female cases with adenocarcinomas, and 381 males and 57 females with squamous cell and small cell carcinomas were recruited. Controls were cancer free individuals, matched for sex and age and comprised of 2964 male and 1189 female. Unconditional logistic regression was used to compute the OR and 95% CI.	**Raw/cooked fish** consumption was significantly associated with a decreased risk of lung cancer. The association between adenocarcinomas and raw/cooked fish consumption in males (OR = 0.51; 95% CI: 0.31– 0.84) showed a decrease in risk with respect to the highest quartile consumption with a statistically significant trend (p = 0.039) while squamous cell and small cell carcinoma were also associated with lower with a non-significant trend. A decreased OR was observed in females (OR = 0.48; 95% CI: 0.24 –0.94) with the highest quartile consumption of raw/cooked fish and adenocarcinomas.	The consumption of raw or cooked fish reduces the risk of adenocarcinomas in the Japanese.
Martin *et al*., 2000[27]	To study the occupational risk factors associated with French electricity and gas industry.	**Caucasian** (France); 310 male lung cancer cases were identified from the company's register and for each case four matched controls were randomly selected. Associations were assessed using conditional logistic regression from which the OR and 95% CI were obtained.	Exposure to 21 chemical compounds were assessed out of which **cutting fluids** (OR = 1.86; 95% CI: 1.14-3.06), **creosotes** (OR = 1.56; 95% CI: 1.08 - 2.27), and **chlorinated solvents** (OR = 1.37; 95% CI: 1.02 - 1.85) were significant after adjusting for socioeconomic status and asbestos. With respect to the level of exposure, after adjusting for socioeconomic status and asbestos exposure, the highest level of exposure of **coal gasification**	Emphasises the carcinogenic property of crystalline silica and the potential role of other agents like creosotes and chlorinated solvents.

			(OR = 3.87; 95% CI: 1.15 - 12.9),**cadmium** (OR = 1.69; 95% CI: 1.00-2.88) and **crystalline silica** (OR = 2.37; 95% CI: 1.25 - 4.49) was associated with lung cancer risk.	
Straif *et al*., 1999[28]	To study the occupational risk factors in **rubber workers** for mortality due to stomach and lung cancer.	**Caucasian** (Germany); to study the recent working conditions in the rubber the study cohort was restricted to recruitment after 1 January 1950 resulting in 8933 participants (1521 deaths).Standardised mortality ratios(SMR) and Cox proportional hazard model were calculated for each work area stratified by age of hire, year of employment in the specific area.	Compare to the reference population, mortality from lung cancer was increased (Observed 154; SMR =123; 95% CI: 104-144). Significant association of lung cancer risk was seen **preparation of material** (RR =2.3; 95% CI: 1.2-4.2), production of **technical rubber goods** (RR = 1.5; 95% CI: 1.1-2.1) and **production of tyres** (RR= 1.3; 95% CI: 0.9 -1.8)	Results depict an association between employment during the initial stages of rubber manufacturing process and excess mortality due to lung cancer.
Droste *et al*., 1999[29]	To investigate the occupational risk factors associated with lung cancer.	**Caucasian** (Belgium); 478 (male) histologically confirmed lung cancer cases and 536(male) controls were recruited from 10 hospitals. Logistic regression was used to calculate the OR and 95% CI adjusting for age, smoking history, marital and socio economic status.	Significant association of lung cancer risk was seen for industrial categories including **transport equipment other than automobiles** (OR= 2.3 95% CI: 1.3 - 4.0), **transport support services** (OR = 1.6; 95% CI: 1.1 to 2.4), and **manufacturing of metal goods** (OR =1.6; 95% CI: 1.0 - 2.5). Occupational exposure to potential carcinogens such as **Molybdenum** (OR = 2.1; 95% CI: 1.2 - 3.7), **mineral oils** (OR = 1.7; 95% CI: 1.1 to 2.7) and	The associations reported were independent of smoking, tobacco consumption and socioeconomic factors. This is the first study reporting the association

			Chromium (OR = 1.4; 95% CI: 1.0 - 1.9) were also significantly associated with lung cancer risk.	between Molybdenum and lung cancer risk.
Jockel *et al*., 1998[30]	To study the carcinogens and occupations related to lung cancer causation.	**Caucasian** (West Germany); 1004 incident lung cancer cases and 1004 controls matched for region, gender and age were recruited. Conditional logistic regression was used to compute the OR and the 95% CI adjusting for smoking and occupational asbestos exposure.	Industries with significant increase in OR include manufacturers of **grain products** (OR = 5.84; 95% CI: 1.06-32.15), **building installation** (OR = 1.60; 95% CI: 1.0-2.56), **seaport** (OR = 1.63; 95% CI: 1.04-2.56) and **life insurance** (OR = 5.31; 95% CI: 1.10-25.71). Occupations that depicted a significant increase include **plastic processing worker** (OR = 3.49; 95% CI: 1.07-11.37), **welder** (OR =1.93 ; 95% CI: 1.03-3.61), **sheet metal worker** (OR = 2.01; 95% CI: 1.14-3.55), **pipe fitter** (OR =2.76; 95% CI: 1.18-6.42), **structural metal worker** (OR = 2.37;95% CI: 1.13-4.96), **grain miller and related worker** (OR =9.61; 95% CI: 1.08-85.69), **docker and freight worker** (OR = 1.95; 95% CI: 1.11-3.42).	Looking into various occupational risks, importantly, after controlling for asbestos occupational exposure indicates the possibility of prevention and future research.
Ko *et al*., 1997[31]	To evaluate the risk factors for lung cancer in non-smoking women.	**Asian** (Taiwan); 117 (106 non-smokers) cases of female non-smokers and 117 matched hospital controls were used. Unconditional logistic regression was used to compute the OR and 95% CI, to evaluate the association	For non-smoking women, cooking in a kitchen without a fume extractor at the age between 20-40 was associated lung cancer (OR = 8.3, 95% CI: 3.1-22.7).**Cooking practices**, **history of pulmonary tuberculosis** and **low dietary intake of fresh vegetables** explained 78% of the risk in non-	Cooking oil fumes exposure in a room without an exhaust is an important risk factor for lung cancer.

		between the various factors and the lung cancer risk.	smokers.	
Benhamou *et al*., 1988[32]	To study the occupational risk factors of lung cancer in the given population.	**Caucasian** (French); 1625 histologically confirmed lung cancer cases and 3091 controls matched for age, gender, interviewer and hospital of admission. Logistic regression adjusting for cigarette smoking was used to evaluate the relationship between various occupational categories.	The risk was lower for professional, technical and related workers (RR = 0.59, p<0.0005) and administrative and managerial workers (RR = 0.68, p<0.02). The risk was higher for **production and related workers**, **transport equipment operators and labourers** (RR = 1.24, p<0.008), **agricultural, animal husbandry and forestry workers, fishermen, and hunters** (RR = 1.22, p < 0.07), **farmers** (RR = 1.24, p < 0.06), **miners and quarrymen** (RR = 2.14, p < 0.02), **plumbers and pipe fitters** (RR = 1.80, p < 0.04), **sheet metal workers** (RR = 1.51, p < 0.08), and **motor vehicle drivers** (RR = 1.42, p < 0.01).	Sample size is an issue and therefore any definite conclusion cannot be made. Certain cohort studies are needed to be carried out to evaluate the risk posed by respiratory carcinogens.
Buiatti *et al*., 1985[33]	To investigate the risk of occupational factors on lung cancer	**Caucasian** (Italian); 376 histologically confirmed cases of lung cancer and 892 controls recruited from the same hospital matched for age, gender, date of admission, smoking status and with diagnosis other than lung cancer and suicide attempt. Logistic regression adjusted for age, smoking and place of birth was used to calculate the OR	For men, four different occupational classes produced an odds ratio greater than 1, including **transportation**, **agriculture, construction and metal work**. However, the category of **stone, clay and glass** produced a significant result (OR = 1.8; 95% CI: 1.1-2.9). Further elucidating the above category, **bricklayers using firebrick and other refractory material** produced a significant result (OR = 6.5, 95% CI: 2.1 - 20.9). For women, **garment workers**	The number of cases for certain occupational categories was small hence further investigation is required.

		and 95% CI.	was the only category that appeared to be significant (OR= 3.5, 95% CI: 1.2-10.5) and within this category the **hat makers** had a significant risk of lung cancer (p=0.01).	

1.4.1 Age and Gender

Lung cancer mostly occurs in individuals above 65 years, with 70 years being the mean age at diagnosis[9]. Though the smoking prevalence is low among elderly patients, their high rate of cancer suggests a heavy smoking history[9]. The five year survival rate is inversely related to age in both genders[9]. The incidence and mortality for lung cancer in the US have decreased in the young (≤50) and increased in the old (≥70), in the past decade[9, 34].

Tobacco smoking is responsible for 80% of lung cancer cases in women[35]. Cigarette smoking increased during World War II, in men born in the 1920s and in women born a decade later[9]. The peak smoking period, for women in the US, was between 1930s-1960s followed by an increase in lung cancer around 1960[9, 36, 37].

The incidence of lung cancer is higher in men than women[38]. This gender specific difference that exists in lung cancer susceptibility may be due to the differences in metabolism and detoxification of carcinogens[9]. Furthermore, DNA adduct level differences have been noted, with higher levels in women than men[39]. Other factors that may cause gender discrepancies include estrogen replacement therapy (ERT) associated with an increased risk (Odds ratio =1.7) and early menopause associated with a decreased risk (OR =0.3) for adenocarcinomas[40]. Another reason for gender discrepancies may be due to greater susceptibility to tobacco-related non-malignant diseases in women than men[9].

1.4.2 Family History

Familial clustering of lung cancer can be explained by both shared environment and genetic factors[41]. The latter can be investigated using segregation and genome wide association analysis[41]. Familial aggregation of lung cancer, first demonstrated by Tokuhata and Lilienfeld[42], is associated with increased risk in both smokers and non-smokers[41]. Genetic contributions to lung cancer susceptibility include the capacity to metabolise and eradicate carcinogens[9]. Lung cancer has also been associated with rare Mendelian cancer syndromes such as Bloom's[43] and Werners' [44].

Many studies have demonstrated an increased risk of lung cancer in relatives[41, 45]. A meta-analysis conducted on 41 studies showed that having a family history of lung cancer increased the lung cancer risk [OR= 1.63; (95% CI: 1.31-2.01)].The risk was further increased if there are two or more affected relatives [OR= 3.60; (95% CI: 1.56-8.31)] [41]. The risk was also affected by the number of first degree family members affected and age of onset[41]. Furthermore, a meta-analysis conducted on 32 studies identified a two fold increase in lung cancer associated with familial aggregation[45].

The first familial linkage study in lung cancer was carried out by the Genetic Epidemiology of Lung Cancer Consortium (GELCC); this implicated a chromosomal region on 6q23-25[46]. Fine genotyping of this region found associations with three SNPs within the *RGS17* gene[47]. The study was conducted using the discovery dataset of 24 cases and 72 controls and validated in 2 independent datasets[47]. The validation dataset from the GELCC, containing 154 cases and 325 controls, produced an OR of 1.76 (95% CI: 1.17-2.68), 1.62 (95% CI: 1.07-2.41) and 1.53 (95% CI: 1.06-2.26) for SNPs rs6901126, rs4083914 and rs9479510, respectively, while the other

validation dataset from the Mayo clinic produced an OR of 1.28 (95% CI: 0.81-2.05), 1.62 (95% CI: 1.03-2.58) and 1.60 (95% CI: 1.02-2.55) for rs6901126, rs4083914 and rs9479510, respectively[47].

Including smoking history in the inheritance analysis produced a three-fold increased risk for lung cancer[46]. Risk models developed by Spitz[48, 49] and Cassidy[50], also identified the importance of family history in determining the risk of developing lung cancer.

1.4.3 Carcinogens

1.4.3.1 Cigarette smoke

A reduction in smoking in the population would lead to decreased lung cancer incidence[9]. Furthermore, by using smoking history to identify high risk individuals, and target screening to this group, early detection will be possible[9]. Lung cancer is indirectly linked to nicotine dependency, as the latter affects smoking behaviour[14]. To exert their carcinogenic effect, many tobacco components need to be activated and subsequently their effect nullified by detoxifying pathways[14]. The interindividual differences that balance the metabolism and detoxifying processes, affect lung cancer risk[14]. Activated carcinogens leads to DNA adduct formation, a covalently bonded product of DNA and carcinogenic metabolites[9]. A permanent mutation would result if the DNA adduct evades cellular repair, resulting in miscoding[9]. DNA damaged cells are eradicated by apoptosis, or programmed cell death, however, if such an irreversible

mutation occurs in an oncogene, leading to its activation and tumour suppressor gene, leading to its inactivation, this would contribute to tumourigenesis[9, 14].

The absolute risk of lung cancer in smokers is affected by duration of smoking and number of cigarettes smoked per day[9]. Other factors that relate lung cancer risk to cigarette smoking include age of smoking onset, degree of inhalation, tar and nicotine content as well as the competence of the filter[9]. Twenty percent of cancers worldwide could be prevented if tobacco smoking were to be eliminated[9]. Though 80% of lung cancers occur in tobacco exposed individuals, only around 20% of smokers develop lung cancer[9].

Cigarette smoke is an aerosol containing gaseous and particulate compounds[9]. Smoke is classified as mainstream smoke, produced by the smoker, by respiring air through the cigarette, and the sidestream smoke, the main source of environmental tobacco smoke, produced by cigarette smokes between puffs[9]. Mainstream tobacco smoke contains carcinogens such as polycyclic aromatic hydrocarbons (PAHs), aromatic and N-nitrosamines and other organic and inorganic compounds, such as benzene. It also contains vinyl chloride, arsenic, chromium and radioactive material like radon and its decayed products[9]. There are at least 50 carcinogens in tobacco smoke[51, 52] but the most important ones implicated in lung carcinoma are the tobacco specific N-nitrosamines (TSNAs)[9].

NNK-induced DNA mutations are associated with *KRAS* oncogene activation[53, 54], which has been detected in 24% of human lung adenocarcinomas[55]. Its detection in the lung adenocarcinomas of former smokers indicates its non-reversion after smoking cessation[56]. Also, benzo[a]pyrene, a constituent of tobacco smoke, can cause many mutations in the *TP53* tumour suppressor gene; such mutations are observed in 60% of primary lung cancer cases[57].

Chemical compounds such as formaldehyde, acetaldehyde, nitric oxide and free radicals in tobacco smoke cause structural alterations including inflammation, permeability, disruption and fibrosis of the respiratory system organs, thereby altering the immune response[58]. The immune response to smoking is still not well understood; however there is an observed decrease in immunoglobulin count and a reduction in antibody response and phagocyte activity[58]. Furthermore, the compounds in cigarette smoke form antigen antibody complexes, bringing about immunological changes, for instance, nicotine present in cigarette smoke is shown to be an immunosuppressant, thus predisposing an individual to various other respiratory, bacterial and viral infections[58].

1.4.1.2 Radiation

The link between lung carcinoma and radiation was first established in 1879 by Harting and Hesse, who noted the increased percentage of deaths from neoplasms among miners in the Schneeberg area of Europe[59]. Later in 1926, Rostoski ascertained the bronchial origin of these tumours and Evans noted that the average time for tumourigenesis was 17 years based on the gamma radiation given off from the inhalation of radon and radon-contaminated dust[59].

Radium 226 decays into radon (radon 222), which is a decay product of Uranium 238[9]. Uranium is naturally present at low levels in outdoor air, and accumulates in homes through fissures in floors, walls and foundations[9]. Uranium mines contain the highest level of radon[6]. Radon decays into polonium 218 and polonium 214, which emits alpha rays[9]. Radon is a ubiquitous,

well established carcinogen, present in soil and rock, causing occupational hazards as well as hazards from exposure to the general population[9].

Radon decays into active components that attach themselves to airborne particles which, when inhaled, adhere to the cells in the respiratory epithelium[6]. Further decay of this radon particle-cell combination may result in DNA damage[6]. Studies suggest there is a linear relationship between radon exposure and lung cancer risk in underground miners, mining being the oldest occupation linked to lung cancer[9]. German uranium miners, exposed for 15-24 years, younger than 55 years in age, have an increased risk of lung cancer[60]. Wagoner reported an increased number of pulmonary carcinomas in underground uranium miners of the Colorado plateau, and similar neoplastic lesions in Japanese factory workers exposed to mustard gas[61].

Lubin and Borce conducted a meta-analysis on 8 studies comprising of 4263 lung cancer cases and 6612 controls producing a relative risk of 1.14 for lung cancer[62]. Radon accounts for 2900 lung cancer deaths each year in never smokers (from residential exposure)[6]. Smoking has a synergistic effect, with radon further increasing the risk of lung cancer in smokers[9].

The lung cancer risk posed by domestic radon exposure is of increasing concern[9]. The intensity of radon gas depends on the concentration of the source of radium and the ventilation in the vicinity of the source[9]. Thus, environmental and indoor radon is a potential contributing agent to lung cancer risk to the public[9]. The concentration of indoor radon, i.e. in homes, concentration depends primarily on the concentration of radium in soils and rocks beneath, while building materials, well water and natural gas contribute a small proportion[9]. Hei and colleagues conducted a study to evaluate the exposure of radon in the general population and

concluded that environmental radon levels could cause mutations in small numbers of bronchial epithelia[63].

1.4.3.3 Asbestos

Asbestos exposure is the most common occupational cause of lung cancer with two naturally occurring types; serpentine (chrysotile) and amphibole (amosite, crocidolite and tremolite)[9]. It was commercially popular since the late 1800s as an insulating and construction material due to its strength and its non-susceptibility to fire[9].

The link between asbestos exposure and lung cancer had initially been suspected in 1934, and was first published in 1955 by Dr. Richard Doll [64]. Using autopsy data, Doll conducted a retrospective study on workers exposed to asbestos, and reported a 10 fold increase in death due to lung cancer compared to the overall population[64, 65]. The relative risk for lung cancer in asbestos-exposed individuals was 3.5 after controlling for age, smoking and vitamin intake[66]. Identical exposures of different asbestos fibres produced different risk estimates with amphibole exposed individuals having higher risks than chrysolite exposed individuals[66]. Also, amphibole fibres are more carcinogenic then chrysolite[67].

Asbestos-related disease can be manifested in pleura (as effusion, pleurisy or both), and pulmonary sites[9]. Pleural plaques, asbestosis and asbestomas are indicators of lung cancer risk[9]. Asbestos-related lung cancer is seen in asbestos textile workers, miners, millers, and asbestos insulation and shipyard workers; and has a dose-dependent relationship[68].

Parenchymal lung diseases resulting from inhalation of asbestos fibres are known as asbestosis[9]. This mainly occurs in workers exposed to an asbestos fibre dose of above 25 -105 fibre/mL/year, for example asbestos insulators, miners, millers and textile workers[68]. Interstitial fibrosis develops when an individual is exposed over periods of months to years[9]. The more intense the exposure, the shorter the time period for the presentation of the disease[9]. Lung fibrosis, including idiopathic pulmonary fibrosis and connective tissue associated interstitial disease, is linked to an increased risk of lung cancer[9, 69].

Interstitial fibrosis such as asbestosis is associated with an increased risk of lung cancer compared to those that are exposed but have no associated fibrosis: asbestosis is therefore a better predictor of excess lung cancer risk than asbestos exposure[66, 70]. Jones and colleagues noted that the risk of lung cancer for non-occupational exposure of asbestos is extremely low[69]. Furthermore, tobacco exposure enhances the carcinogenic effect of asbestos and increases the risk of lung cancer in asbestos workers[9]. The risk of lung cancer from exposure to asbestos alone, cigarette smoking alone and cigarette smoke and asbestos exposure combined is 6 fold, 11 fold and 59 fold, respectively[9].

The biological mechanisms causing lung cancer following asbestos exposure are yet to be determined, however, inhalation of asbestos fibres activates macrophage and airway epithelium causing inflammation and cell proliferation[65]. Evidence suggests that asbestos exposure induces *KRAS* mutations, with increased mutation risk after higher exposure[65].

1.4.3 Respiratory Conditions

Non-malignant diseases, such as pneumonia, tuberculosis and chronic obstructive pulmonary disease (COPD) have been extensively studied as risk factors for lung cancer[18]. In particular, COPD, characterised by chronic inflammation, which itself is being evaluated in lung cancer research, is considered to be associated with lung cancer[9]. These infections and lung diseases are described below.

Tuberculosis: The relationship between tuberculosis (TB) and lung cancer was first published in 1810 and has subsequently been elaborated upon[12]. Two recent studies carried out in China and Taiwan demonstrated that the risk of lung cancer increases due to tuberculosis and that tuberculosis causes chronic inflammation[12, 71, 72].

The lengthy period between symptom onset and diagnosis of TB and prolonged treatment of 6-9 months, results in substantial pulmonary inflammatory damage with the production of TNF, TGF-β, IL4 and IL13[11].

A retrospective study carried out on 42,422 farmers in rural Xuanwei County in China showed an association between tuberculosis and lung cancer mortality: this was more pronounced in the first 5 years after TB diagnosis but remained strong in subsequent years[72]. This study was performed using proportional hazard regression modelling and the association did not change after adjusting for demographic characteristics, lung disease and tobacco consumption[72].

Another population-based cohort study carried out in Taiwan on 5657 TB patients and 23,984 controls showed that the occurrence of lung cancer was significantly higher in TB patients than

in controls[71]. This analysis was carried out by calculating incidence rates and hazard ratios of lung cancer[71]. The studies suggest that the persistent risk of lung cancer, years after diagnosis of TB could likely be due to chronic pulmonary inflammation and scarring[12].

A 16 study analysis, conducted to study the effect of previous TB, produced a relative risk of 1.48 (95% CI: 1.17-1.87)[18]. When the analysis was conducted for ever smokers, the relative risk was 1.36 (95% CI: 1.05-1.75) and never smokers was 1.50 (95% CI: 1.03-2.19)[18].

Pneumonia*: Chlamydophila pneumoniaea,* transmitted through respiratory secretions causes many acute and chronic respiratory conditions, including pneumonia, and potentially, lung cancer[73]. Other microorganisms that cause pneumonia include *Streptococcus pneumoniaea*, *Haemophilus influenzae* and *Legionella pneumophila*[11]. This illness is time limited and most patients recover quickly; the pulmonary inflammation is of short duration, causing less severe scarring[11]. The mechanism by which this infection may elevate lung cancer risk is not established, although smoking may enhance the chance of contracting pneumonia[73].

The infection process involves a complex interplay between various inflammatory components such as superoxide oxygen radicals, TNF –α, IL1β and IL8[73]. These result in tissue and DNA alterations, and may lead to carcinogenesis[73].

A contradictory result is presented by Koshiol *et al*. (2010), who showed that multiple bouts of previously-reported pneumonia (≥3) appear to decrease the risk of lung cancer in an analysis conducted on 1846 cases and 2054 controls, producing an OR of 0.35 (95% CI: 0.16-0.75)[74]. Information about previous pneumonia for this study was collected through an interview[74]. Nonetheless, a meta-analysis conducted on 12 studies supports the suggestion that *C.*

pneumoniaea increases lung cancer risk[73]. This study was carried out using 12 studies comprising of 2595 cases and 2585 controls producing an OR of 1.48 (95% CI: of 1.32-1.67)[73].

Another 12 study analysis produced a relative risk of 1.57 (95% CI: 1.22-2.01) associated with lung cancer[18]. When stratified analyses were conducted, smokers produced a relative risk of 1.55 (95% CI: 1.16-2.06) compared to non-smokers who produced a relative risk of 1.35 (95% CI: 1.12-1.63)[18].

COPD: Chronic obstructive pulmonary disease (COPD) is an inflammatory condition of the lower airway characterised by emphysema and constriction of bronchi, provoked by inhalation of noxious particles or gases[12, 13, 75]. It is diagnosed in 50-70% of lung cancer patients and it has been shown to increase lung cancer risk by 4.5 fold in long term smokers[75].

The link between COPD and lung cancer was identified as early as 1939[12]. Studies carried out by Kishi *et al*. (2002) and Wilson *et al*. (2008), demonstrate that lung cancer risk increases with increasing airflow obstruction[76, 77]. These studies used spirometry and CT to measure lung function and emphysema progression in over 1000 participants[76, 77].

Furthermore, recent studies identified that COPD and lung cancer share a common aetiology[12, 13, 75]. Common biological mechanisms in these diseases include inflammation, EMT, oxidative stress, matrix degeneration, cell proliferation, anti-apoptosis, abnormal wound repair, angiogenesis and other pathways[12, 13, 75].

Potential genetic changes that may increase the susceptibility to both lung cancer and COPD include SNPs in genomic DNA (nicotine receptor polymorphisms), copy number variations (CNVs), epigenetic alterations, mRNAs and microRNAs[13]. A 16 study analysis of previous

emphysema and lung cancer produced a relative risk of 2.44 (95% CI: 1.64-3.62). Stratified analysis produced a relative risk of 2.21 (95% CI: 1.00-4.90) and 2.25 (95% CI: 1.50 -3.37) for never and ever smokers, respectively[18].

1.4.4 Socioeconomic Status

Socioeconomic status is inversely associated with lung cancer risk[78]. People from poorer and less educated areas of society are more susceptible to lung cancer than those from richer areas[78]. A Canadian study showed that education, income and higher social class analysed separately in male and female cancer cases, after adjusting for smoking, are inversely associated with lung cancer risk[78]. Similar results were found in a Netherlands-based study, which described a reciprocal relationship between highest level of education and lung cancer after adjusting for smoking, age, and dietary intake of vitamin C, beta- carotene and retinol[79]. Study conducted between 1998-2003 using English cancer registries identified lung cancer incidence to be associated with poor patients[80]. The postcode at diagnosis was used to allocate socioeconomic status calculated using the income domain of the Index of Multiple Deprivation 2004[80].

1.5 Molecular Genetics of Lung Cancer

Carcinogens bring about multiple genetic changes, mostly through DNA adduct formation[81]. Genomic integrity is disrupted if proper cell cycle regulation and checkpoints are not maintained, resulting in genomic instability and ultimately initiating cancer development[81].

Tumour suppressor genes (TSGs) and oncogenes are two groups of functional genes implicated in lung cancer[82]. Multiple TSGs are inactivated during carcinogenesis and cancer progression. Inactivation is usually by a two-step process known as "Knudson's two-hit hypothesis"[83], involving the loss of function of both copies of a TSG in the same cell. Inactivation can occur through deletion of large chromosomal segments and smaller mutational changes, or epigenetic aberrations[81]. Loss of heterozygosity (LOH) studies has helped to locate tumour suppressor genes by identifying the sites of chromosomal deletions[81]. Many TSGs and oncogenes mutated in lung cancer, control cell cycle processes. Structural and numerical changes, together with genetic and epigenetic changes in the genome, are also observed in lung cancer[81].

Tumour suppressor genes identified in lung cancer as being inactivated via chromosomal deletions include *SMAD2, SMAD4, PTEN, FHIT, PPP2R1B, p53, RB* and *p16*[84].

1.5.1 Proto Oncogenes

1.5.1.1 *EGFR* (Epidermal Growth Factor Receptor)

The *EGFR* family of proto-oncogenes comprises the *EGFR*, *HER2*, *HER3* and *HER4* genes that encode tyrosine kinase receptors[85]. Over-expression of *EGFR* and *HER2* is seen in 70% and 30% of NSCLC, respectively, while in SCLC, they are less often expressed[85]. *EGFR* mutations are also prevalent in adenocarcinomas, East Asians, women and never smokers, and activate P13K/AKT and STAT2/STAT5 pathways[17]. Mutations in these genes are seen mainly at the TK receptor domain and are largely confined to the first four exons[85]. Various drugs that specifically target *EGFR* or *HER2* are now available, including the monoclonal antibodies gefitinib, erlotinib, trastuzumab and cetuximab[85].

Activation of mutant *EGFR* in adenocarcinomas activates *STAT3* through increase of IL6 which is responsible for activation of *STAT3, MAPK* and *PI3K* via activation of *JAK* family tyrosine kinases[17]. Resistance to *EGFR* tyrosine kinase inhibitors (TKIs) (such as erlotinib and gefitinib) is linked to *EGFR* exon 20 insertion and T790M mutation, *KRAS* mutation or amplification of the *MET* proto oncogene. Inhibition of *MET* signalling can restore sensitivity to TKIs[17].

1.5.1.2 *RAS*

The *RAS* family of proto-oncogenes includes *HRAS*, *KRAS* and *NRAS*, which are involved in normal cellular differentiation, proliferation and survival, and regulate important signalling pathways[85]. It is one of the first oncogenes identified, and encodes a protein of molecular weight, 21 kDa[86]. Although *RAS* mutations are seen in human cancers from diverse tissues, including the lung, they are mostly absent in SCLCs, but are detected in 10-15% of NSCLC, especially adenocarcinomas[85].

The *KRAS* oncogene is mutated in 30% of adenocarcinomas, predominantly by G:T transversions in codon 12[86]. *KRAS* mutations that activate signalling pathways occur at three specific codons in 20% of lung cancers, mainly adenocarcinomas[17]. Ninety percent of lung cancer mutations in *KRAS* occur at three specific codons (80% in codon 12; 10% in codons 13 and 61)[17] and *KRAS* accounts for 90% of mutations in lung cancers[85]. Most resectable lung cancer cases show *RAS* mutations and expression associated with decreased survival[86].

Active *RAS* results in the activation of downstream signalling pathways such as PI3K and MAPK[17]. In a normal functioning cell, the intrinsic GTPase activity of *RAS* converts it from the active GTP–bound state to the inactive GDP–bound state[81]. A mutation in codon 12, 13 or 61 causes *RAS* to exist in a constitutively active GTP-bound state; this activated *RAS* leads to permanent activation of the RAF1/MAPK pathway[81].

RAF kinase and *MAPK* kinase (*MEK1*) are two entities targeted in drug trials as they are downstream effectors of *RAS* signalling[17]. Inhibition of *RAF* kinase included mRNA *RAF1* degradation and kinase activity inhibition (sorafenib)[17]. Tipifarnib and Ionafarnib therapies

have entered the phase II trial[17]. Small interfering RNAs (siRNAs) have also been identified that specifically target MAPK mutated human lung cancer cells in vitro[17].

1.5.1.3 *MYC*

The *MYC* oncogenes; *c-MYC* (cellular), *N-MYC* and *L-MYC*, encode transcriptional regulating nuclear DNA binding proteins[86]. Structurally, *MYC* contains a transactivating domain at its N terminus and nuclear localization signal, helix-loop-helix domain and leucine zipper at the C terminal[86]. It regulates transcription by forming homodimers or heterodimers (via the helix loop helix and leucine zipper domain) with proteins such as MAX, MAD and MX11[86]. MAX represses transcriptional activation when bound to *MYC*, while MAD and MX11 promote transcriptional activation by releasing *MYC* from the MAX bound state[86]. NSCLCs rarely show amplification of *c-MYC*, however in SCLCs this amplification adversely affects survival[86].

In a normal functioning cell, *MYC* controls cellular proliferation via regulation of downstream signalling pathways such as cell proliferation[17]. The activation of *MYC* is often through gene amplification[17]. *MYC* is usually activated in NSCLCs while all three members are activated in SCLCs[17].

1.5.1.4 *BCL-2*

The *BCL-2* gene inhibits apoptosis, increases survival of non-cycling cells and regulates cell death[86]. In a normally functioning cell, *BCL-2* forms complexes with *BAX*, and exists as a heterodimer to regulate apoptosis[86]. Factors such as radiation decrease the transcription of *BCL-2* and induce transcription of *BAX,* resulting in *BAX* homodimer formation and leading to apoptosis[86].

1.5.2 Tumour Suppressor Genes

1.5.2.1 *TP53*

TP53 is a tumour suppressor gene on chromosome 17p13.1, which functions by permitting DNA repair and initiating apoptosis or cell cycle arrest in response to cellular stress such as DNA damage[85]. It is 20kb long, has 11 exons, and encodes a 53 kDa nuclear protein of 393 amino acids in length[86]. An altered *TP53* gene is the most common genetic variation involving deletions, point mutations and over expression, associated with cancer[86]. *TP53* is inactivated in 90% and 50% of SCLCs and NSCLCs, respectively[85].

TP53 is a transcription factor for DNA damage and induces expression of *p21, MDM2* and *BAX,* which regulate the cell cycle and apoptosis[84]. *TP53* maintains integrity of the genome by playing a role in cell cycle checkpoints[84]. Its inactivation leads to accumulation of mutations,

chromosomal rearrangements and abnormal chromosomal segregation[84]. *TP53* is mutated early in lung carcinogenesis, demonstrating its role in the progression of malignancies[84].

In a normally functioning cell, the anti-apoptosis proteins BCL-2 and BAX are in homeostasis[86]. When the cell is damaged, p53 levels rise and bind to the BAX promoter, increasing its transcription, thus leading to cell apoptosis[86]. Mutant p53 can lose its tumour suppressor properties, promote cell multiplication and prevent apoptosis[86].

1.5.2.2 Deletions in 3p Region

More than 90% of SCLC cases and approximately 70% of NSCLC cases display allelic loss of the 3p chromosomal region[81]. Such losses have been reported in the 3p25-p26, 3p21-22, 3p14 and 3p12 chromosomal regions[81]. LOH has been noted in at least eight distinct sites[81]. These chromosomal regions contain several genes with tumour suppressing properties, and thus the disease progresses by inactivating the expression of these crucial tumour suppressors[85].

Normal or partially abnormal tissue of lung cancer patients and healthy smokers shows 3p allelic loss[81]. Furthermore, the frequency and intensity of these changes correlates with the severity of histopathological preneoplastic/preinvasive changes[81].

Loss of one copy of chromosome 3p is observed in 96% and 78% of lung tumours and lung preneoplastic lesions[17]. Studying this loss has helped identify several TSGs including *FHIT* (3p14.2), *RASSF1A, TUSC2, SEMA3, SEMA3F* (all at 3p21.3) and *RARβ* (3p24)[17].

Some TSGs located in this region include *FHIT* at 3p14.2; this gene has been shown to inhibit tumour growth by apoptosis and cell cycle arrests in lung cancer cell lines[81]. All SCLCs, and more squamous cell carcinomas than adenocarcinomas, demonstrate a loss of expression, also reported to be associated with smoking history in lung cancers[81].

Decreased expression due to epigenetic hypermethylation is seen in genes such as *FHIT, RASSF1A, SEMA3B* and *RARβ*[17]. *FHIT* induces apoptosis while *RASSF1A* alters cell cycle regulation[17]. *TUSC2* exerts its effect by inhibiting protein kinases (*EGFR, PDGFR, cAb1, c-kit* and *AKT*) and degradation of p53 through *MDM2* mediated inhibition[17]. *SEMA3B* decreases cell proliferation and induces apoptosis, *SEMA3F* inhibits vascularisation and tumourigenesis while *RARβ* functions by reducing cell growth and differentiation[17].

1.5.2.3 *RB*

RB gene is located at 13q14.11, and was discovered to be a tumour suppressor in a familial inheritance study of retinoblastoma[86]. It encodes a nuclear protein of106 kDa, which is crucial in cell cycle regulation during G0/G1 phase[86]. In its hypophosphorylated state, *RB* is active, and bound to the E2F family of transcription factors[86]. When *RB* is phosphorylated by the cyclin dependent kinase CDK4, E2F is released; this leads to cell cycle progression through the G1/S checkpoint[86, 87].

There is absence of RB protein expression in 70% of SCLC cell lines, as a result of structural alterations in *RB* gene or abnormal mRNA expression, and 10% of NSCLC cell lines show

absence of, or abnormal, RB mRNA, while 30% of tumours show absence of or abnormal, RB protein[86].

1.5.2.4 *p16^{INK4A}*

The *CDKN2A* gene, which encodes the *p16* protein, is located on chromosome 9p21, an area that is deleted in some lung cancers[86]. Absence of *p16^{INK4A}* inhibitor is noted in some lung cancers[86]. Variation of *p16^{INK4A}* is thought to be a late event in lung cancer progression[86]. *p16* prevents the transition from G1 to S phase by inhibiting *CDK4* in the cell cycle[86]. Inhibition of *CDK4* by *p16* keeps the tumour suppressor gene *RB* in its active, hypophosphorylated, state, thus inhibiting progression through the cell cycle[86].

1.5.3 Genetic Susceptibility to Lung Cancer

Following are some candidate genes whose variations may be associated with lung cancer susceptibility.

Xenobiotic Metabolising Enzymes: Candidate gene association studies for genes involved in tobacco smoke carcinogen metabolism have been extensively studied[14]. Cytochrome P450 (CYP)- related enzymes and Glutathione-S-transferases (GSTs) are metabolic enzymes involved in phase I and phase II metabolism, respectively[14]. Inherited variations in the genes encoding these enzymes can affect an individual's capacity to activate and detoxify foreign compounds

(including carcinogens) and hence impact upon susceptibility to a variety of cancers[14]. SNP rs1048943 (Mspl (T3801C)) in the *CYP1A1* gene was tested using 17 studies, comprising 1759 cases. This produced an OR of 2.36 (95% CI: 1.16-4.81) for presence versus absence of the Mspl site[88]. *GSTM1* (presence/null) was tested using 130 studies totalling 23,452 cases and 30,397 controls; this produced an OR of 1.18 (95% CI: 1.14-1.23) in an allelic model[89]. *GSTT1* (presence/null) was tested using 8 studies including 1364 cases; this produced an OR of 1.28 (95% CI: 1.10-1.49) in a recessive model[88].

DNA Repair Genes: SNPs in DNA repair genes studied include rs1800975 (G-23A) in the *XPA* gene[88]. This was analysed using 7 studies comprising of 1913 cases, and produced an OR of 0.73 (95% CI: 0.61-0.89) for the heterozygous versus the non-variant homozygous genotype[88]. The SNP rs2228001 (Lys939Gln) in *XPC* was analysed using 6 studies comprising 2580 cases, and produced an OR of 1.30 (95% CI: 1.11-1.53) in the recessive model[88]. rs1052550 (Lys751Gln) in *XPD* was analysed using 15 studies made up of 5004 cases, and produced an OR of 1.30 (95% CI: 1.13-1.49) for variant homozygous Gln versus other homozygous, Lys[88]. rs25487 (Arg399Gln) in *XRCC1* was studied using 6 studies containing 1702 cases, and produced an OR of 1.34 (95% CI: 1.16-1.54) for variant homozygous Gln versus other homozygous, Arg[88]. rs1052133 (Ser326Cys) in *OGG1* produced an OR of 1.32 (95% CI: 1.04-1.67) for non-smokers tested using a dominant model by using 17 studies comprising of 6375 cases and 6406 controls[90].

Cell cycle genes: SNPs in genes involved in cell cycle regulation such as rs1042522 (Arg72Pro) in *TP53* and rs2279744 (T309G) in *MDM2* were tested in a meta-analysis[91, 92]. rs2279744 (T309G)

in the *MDM2* gene was tested using 7 studies comprising of 4276 cases and 5318 controls; this produced an OR of 1.27 (95% CI: 1.12-1.44) when the variant homozygote GG was tested against the wild type homozygote, TT[91]. For rs1042522 (Arg72Pro) in *TP53*, a meta-analysis carried on 32 studies comprising of 9046 cases and 10127 controls, was performed using the genotypic model[92]. When comparing heterozygotes versus the Arg homozygotes; this produced an OR of 1.21 (95% CI: 1.01-1.23); for variant Pro homozygotes versus wild type Arg homozygotes an OR of 1.20 (95% CI: 1.03-1.39) was produced[92]. For the dominant and recessive model the OR was 1.14 (95% CI: 1.03-1.25) and 1.06 (95% CI: 1.01-1.12), respectively[92].

1.5.4 Epigenetics

Epigenetic mechanisms include DNA methylation and post translational modifications of histones[17]. These alterations are somatically heritable, and cause gene silencing without altering the DNA sequence itself[12, 75], they are therefore reversible[17]. An important epigenetic change associated with lung cancer is hypermethylation of cytosine residues within CpG dinucleotide islands in certain transcriptional promoter regions[12, 75, 85]. Other important epigenetic changes associated with lung cancer include global DNA hypomethylation, post translational modification of histones and miRNA silencing by DNA hypermethylation[75].

Genes including TSGs are inactivated by epigenetic changes occurring early in lung tumourigenesis[17]. Other genes include those involved in tissue invasion, DNA repair, detoxification of tobacco carcinogens and differentiation[17]. Lung cancer-associated promoter

hypermethylation has been detected in almost 80 genes including *RARB, TIMP3, p16INK4a, RASSF1A, MGMT, FHIT, DAPK, ECAD*, and *GSTP1* [12, 75, 85]. Genes affected by epigenetic changes have utility as biomarkers for early detection research and prognosis[17].

1.5.5 Micro RNA

Micro RNAs (miRNAs) are non-coding RNA sequences that may exert a negative regulatory influence on mRNA's stability or expression[93]. They play a role in many cellular processes including cell proliferation, differentiation and apoptosis[75]. They regulate gene expression at the post-transcriptional level and play a role in various developmental processes including neurogenesis, insulin secretion, cholesterol metabolism and the immune response[75, 93, 94]. MicroRNAs regulate many biological processes and may play a role in the pathogenesis of most human cancers (by modulating the expression or function of oncogenes and TSGs); those miRNAs that function as tumour suppressors or oncogenes, are referred to as oncomirs[93]. A recent study identified seven miRNAs in an analysis using paired tumour-normal tissues from 20 patients[95]. These miRNAs were optimised in a study of 36 cases and 36 controls: four miRNAs (miR-21, miR-486, miR-375 and miR-200b) capable of distinguishing lung adenocarcinoma patients from controls were identified, with a sensitivity and specificity of 80.6% and 91.7%, respectively[95]. Five miRNAs (hsa-let-7a, hsa-miR-221, hsa-miR-137, hsa-miR-372, and hsa-miR-182*) were identified in a study carried out on 112 NSCLC patients to predict relapse and survival[96]. MiRNAs could be potential tools influencing prevention, and diagnosis, prognosis of lung cancer, and therapy of lung cancer patients[97].

1.6 Early Detection Research

The overall five-year survival rate for those diagnosed with lung cancer is 16% in the USA while in the UK it is as low as 7.8% and 9.1% for men and women, respectively[98, 99]. This is largely due to late presentation of symptoms[100]. If lung cancer were to be detected at an earlier stage, it might be possible to improve the overall survival figures[100]. Early detection research has helped to identify changes that occur before the development of clinically evident lung cancer[100]. The following strategies have shown promise in early detection research.

1.6.1 Sputum

Sputum can be used in a variety of ways in early detection research, by means of DNA, RNA and protein analysis, routine cytological examinations and nuclear image analysis[101]. DNA methylation and *KRAS* mutations are frequently reported in sputum[101].

In previous reviews, average sensitivity for cytological detection of lung cancer was reported to be 65%[101]. To improve the accuracy, other techniques such as fluorescent in situ hybridisation (FISH), and studies of promoter hypermethylation and genetic mutations were investigated[102]. A combination of methodologies (sputum cytology with FISH) produced a sensitivity of 76% and specificity of 92%[102].

1.6.2 Computed Tomography

Several trials are underway to study the utilisation of CT screening for early detection of lung cancer[102]. This technology is advantageous due to its feasibility, speed, resolution, ability to reconstruct multiple series from a single data acquisition and detect small peripheral lesions[102].

European screening trials such as the UK Lung cancer Screening trial (UKLS)[103], Danish randomised CT screening trial, NELSON, ITALUNG and LUSI are still acquiring data before mortality figures can be calculated[102]. In 2002, the randomised National Lung Screening Trial (NLST) was set up to compare lung cancer mortality in high risk individuals screened with either low dose CT or chest radiography[104]. When the collected data were analysed in 2010, a 20% reduction in mortality was observed, thus supporting the use of low dose CT as a screening strategy[104].

Whilst CT screening predominantly detects adenocarcinomas, it is unable to detect preinvasive lesions and most importantly lesions in the central airway, high grade dysplasia and early stage squamous cell carcinomas[102].

1.6.3 Bronchoscopy

The most widely used technique for early diagnosis is light induced fluorescence endoscopy (LIFE)[105]. Other existing bronchoscopic imaging techniques include autofluorescence bronchoscopy (AFB), high magnification bronchovideoscope, and narrow band imaging (NBI);

and more precise techniques like endobronchial ultrasound (EBUS) and optical coherence tomography (OCT)[105].

Each technique has its own pros and cons, for instance, AFB is an improvement over LIFE and generates a red:green light ratio; NBI provides a higher specificity of 80% without compromising its sensitivity while OCT provides a highly resolved cross sectional image of the mucosa[102]. While high cost is the main drawback of this technique, based on real time imaging, it can differentiate between inflammation and premalignant lesions[102].

Bronchoscopy, together with CT scanning, was applied to smokers and former smokers with mild or moderate COPD, if positive for sputum cytology/cytometry, in a multicentre randomised controlled trial called Lung-SEARCH, in the UK[106]. The aim of this trial was to identify lung cancer patients at an earlier stage in the disease, compared to the unscreened group[106].

1.6.4 Breath Test

Exhaled breath analysis is a non-invasive approach to identify inflammatory and oxidative stress markers potentially involved in various respiratory conditions[102]. A technique called the exhaled breath condensate (EBC) is still in its initial stages of experimentation and the markers evaluated include 3p microsatellite signature, DNA methylation, angiogenic markers, *COX-2*, endothelin and survivin[102].

1.7 Novel Technologies in Lung Cancer Research

Completion of the human genome project[107] and the development of genomic technologies[108] have led to discoveries that will help improve clinical care for cancer patients[109]. Some of these developments are outlined below.

1.7.1 Gene Expression Profiling Using Microarray

cDNA and oligonucleotide expression microarrays are widely available to study differential gene expression in lung cancer, and can be applied to classify cancer types, identify new oncogenic markers, and predict prognosis and response to drug treatment[85].

The fundamental principle behind microarray technology is the complementary hybridisation of cRNA or cDNA to the sample containing the gene of interest[110]. Unlike genome wide association analysis, this is a "closed" gene expression technology where prior knowledge about the sequence of genes under study is required[110]. Available technologies include spotted microarrays that use customised product embedded on a glass slide, and the Affymetrix Genechip system that utilises a prefabricated oligonucleotide microarray[110]. Overall, microarray technology is a robust tool, finding its use in tumour classification, prognosis and patients' response to therapy[85].

1.7.2 Genome Wide Association Analysis

Genome wide association studies (GWAS) are gaining importance in recent years as they have enabled identification of genetic variants associated with human diseases[111-114]. Though such studies involve large sample sizes and extensive genotyping, they have the advantage of not being based on any prior assumptions about the functional significance of the typed variant[115].

In lung cancer research, three studies recently reported the identification of genetic variants on chromosomal regions 15q24-25.1, 5p15.33 and 6p21[111-114]. These include the IARC study[113] on 1989 lung cancer cases and 2625 controls, a study by the MD Anderson clinic[111] including 1154 cases and 1137 controls and a study from DeCode, Iceland[112] including 665 cases and more than 10,000 controls[115].

Interestingly, all 3 studies pointed towards a susceptibility region located at 15q25.1 with a more or less consistent OR, ranging from 1.30-1.32 across all studies[112-114].

Another SNP located at 6p21 was reported by the IARC study[113]. Additionally, the IARC conducted a GWAS scan by pooling together a larger number of cases (N=3259) and controls (N=4159); this study reported another susceptibility locus on chromosome 5p which contains the *TERT* and *CLPTM1L* gene[113]. The above result was also replicated in another meta-analysis involving more than 5000 cases and 5000 controls[115].

1.7.3 Next Generation Sequencing

Next generation sequencing (NGS) approaches have the ability to sequence a number of cancer genes in one attempt as well as simultaneously detect gene alterations (base substitutions, CNVs, insertions and deletions) making it an improvement over traditional sequencing methods[108]. NGS technology not only provides a thorough and in depth sequencing, but also produce more than 1 billion sequences in a 4 day run per instrument at a far cheaper cost per base compared to the traditional dye terminator technique[108].

Available technologies for NGS include 454 Pyrosequencing, Ion Torrent, Illumina, SOLiD (Supported Oligonucleotide Ligation and Detection) and Helicos[108]. Though these platforms are costly there is a potential for cost reduction in the near future. Some platforms such as 454 Pyrosequencing and Ion Torrent are faster than Illumina and SOLiD, but limited in their capability to carry out parallel deep sequencing[108]. Some technologies such as Helicos, SOLiD and Illumina are better suited for whole genome sequencing, while others, such as 454 Pyrosequencing and Ion Torrent, are for targeted sequencing[108].

All these technologies generate a large amount of data and therefore there are complexities regarding data processing, storage and analysis[108].

1.9 Risk Models

Risk models have been used in various diseases in deciding whether to opt for a particular invasive diagnostic test, and to predict the likelihood of the disease progressing to a later stage and the outcome of specific therapies[116]. Examples are the Gail model for breast cancer that computes the lifetime risk for an individual[117]; prediction via a scoring system from questionnaire data for colorectal cancer [118] and the Prostate Cancer Prevention Trial (PCPT) model that predicts the probability of prostate cancer[119]. Additionally, various nomograms are available for predicting disease progression in prostate cancer[120] and models that use biomarkers, like the ovarian cancer model that uses ROCA[121] and prostate cancer recurrence prediction calculator that utilises the prostate specific antigen (PSA) measurement after radiation therapy[120]. Furthermore, advances have been made in utilising expression data of 21 genes in breast cancer (Oncotype DX) to predict the recurrence risk[122].

Risk models for lung cancer could potentially have utility in targeting, screening and resources towards high risk populations and individuals[48, 50, 123]. Three prominent risk models used in lung cancer research include the Bach model[124], Spitz model[48] and the LLP risk prediction model[50]. The Bach model, devised to compute the risk for 10 years, was based on a cohort comprising of 18,172 current or former smokers[124]. Cox proportional hazard modelling was used to design the model including age, sex, prior history of asbestos exposure, smoking duration, average amount smoked per day for current smokers, and duration of abstinence from smoking for former smokers as significant covariates[124]. The Spitz model was based on a case control study with 1851 cases and 2001 controls, frequency matched for age, sex, ethnicity and smoking status[48]. Multivariate logistic regression was used to develop this one year absolute risk model

with age, sex, smoking history variables, environmental tobacco smoke, family history of lung cancer, emphysema, exposure to dust and asbestos as significant covariates[48].

The five year LLP risk model was also based on a case-control study comprising of 579 cases and 1157 controls, matched for age, sex and smoking status[50]. Multivariate conditional logistic regression was used to design the model, with age, sex, smoking duration, asbestos exposure, prior diagnosis of pneumonia, malignancy, and family history of lung cancer being significant covariates[50].

A five year absolute risk model for African-Americans was developed using 491 African-Americans with lung cancer and 497 matched African-American controls[123]. The existing models were not suitable for risk prediction in these minority populations, as they were developed and based on Caucasian populations; since there is variation between the risks for different ethnic groups, the significant covariates for the models differ[123].

CHAPTER 2

INFLUENCE OF COMORBIDITY ON THE INCIDENCE OF LUNG CANCER AND THE DEVELOPMENT OF AN INCIDENCE MODEL

2.1 Aim

The primary aim of this research was to analyse the effect of medical conditions represented in the form of the Charlson comorbidity index (CCI) and the Elixhauser comorbidity index (ECI) on the incidence of lung cancer.

The same dataset was used to develop a sex specific risk model to predict the risk of developing lung cancer for a definite time period with variables that can be easily collected either through questionnaire or a clinical practice using a Cox proportional hazard model. The risk score produced using this model is a cost effective way of identifying and referring high risk individuals for further clinical examinations.

2.2 Introduction

Lung cancer was the most common and the second most common cancer in males and females respectively, and the leading and second most leading cause of cancer death in males and females, respectively, in 2008, worldwide[38]. Female lung cancer death comprise of 11% of the total female cancer mortality while the male lung cancer death rate is decreasing in the western countries and increasing in countries such as China and other countries in Asia and Africa[38].

In the same year, lung cancer comprised 12.7% (1.61 million) of the total incident cancer cases (12.7 million) diagnosed in the world. Of these incident lung cancer cases, 16.5% (0.266 million) were males and 8.5% (0.137 million), females[1]. In 2009, the age standardised rate (ASR) for

incidence of lung cancer in England was 56.3 per 100,000 for males and 37.5 per 100,000 for females (Office of National Statistics)[125].

Feinstein defines comorbidity as any medical condition that pre or co exists with the disease under study[126]. Comorbidity affects patient care and is a major factor in long term survival of cancer patients[127]. Cancer specific research evaluating important comorbidities in the clinical trajectory of a patient can shed light on the diagnosis, treatment and long term monitoring of patients with comorbidities[127]. For example, the decreased chances of survival for patients with severe chronic obstructive pulmonary disease (COPD) due to unsuitability for lung malignancy resection or the presence of a congestive heart failure making the patient unfit for treatment[127]. It is therefore important to evaluate comorbidities in cancer patients as they govern the decisions involving prognosis, treatment and care[128].

Comorbidities associated with lung cancer include respiratory conditions such as pneumonia, tuberculosis (TB) and COPD (section 1.4.4), asbestosis (section 1.4.3.3), diabetes mellitus (DM)[129], body weight[130] and cardiovascular diseases[131, 132]. COPD is closely linked to lung cancer as they not only have a common environmental risk factor; cigarette smoke exposure, but are considered to have shared genetic and epigenetic mechanisms[131]. The risk of developing lung cancer is 4.5 times higher in patients with COPD[131]. Other diseases linked to COPD also include congestive heart failure and ischaemic heart disease[132]. This link between cardiovascular diseases and respiratory impairment may be due to shared risks such as cigarette smoking, severe infections and inflammation. Furthermore, respiratory defects cause the development and recurrence of cardiovascular disease imminent[132]. Subjects with respiratory condition and normal lung function have an increased risk of cerebrovascular disease equal to GOLD stage 3

or 4 for COPD, in magnitude[132]. Hypertension may be the link between respiratory and cardiovascular diseases[132].

Obesity was evaluated as a risk factor for lung cancer in current, former and never smokers, in 448732 individuals aged between 50-71 years[130]. The analysis was conducted by using body mass index (BMI) as a surrogate for obesity and individuals were recruited between 1995-1996 from the National Institutes of Health-AARP Diet and Health Study. BMI (≥35 vs 22.5–24.99 kg/m^2) produced a hazard ratio (HR) of 0.81 (95% CI: 0.70 - 0.94) and 0.73 (95% CI: 0.61 -0.87) associated with the risk of lung cancer for men and women, respectively, after adjustment for age at study entry, detailed smoking status and dose, cigar/pipe smoking, race/ethnicity, education level, history of emphysema, physical activity, and alcohol intake[130].

A study conducted on 1226 lung cancer cases with DM from a total sample of 61777 and 4281 lung cancer cases without DM produced an odds ratio (OR) of 1.296 (95% CI: 1.214-1.384) after adjusting for sex and age. Study subjects were recruited using Taiwan's National Health Research Institute database from 2000-2008[133]. Another study investigated the incidence of lung cancer in post-menopausal women aged between 50-79 years[134]. An age adjusted model indicate a HR of 1.26 (95% CI: 1.02-1.56) for individuals with treated diabetes[134]. Furthermore, overall survival study conducted on 1111 lung cancer cases produced an HR of 1.44 (95% CI: 1.15-1.80) for diabetes in a multivariable Cox regression model adjusted for gender, age, other cancers, TB, COPD, hypertension, and stage, indicating decreased survival[129].

Though comorbidities govern various decisions in a patient's trajectory, evaluating its effect is complex[127]. There are various sources from which comorbidity data can be obtained. Data from epidemiological studies can be obtained from clinical trials, cohorts (retrospective/prospective) or admission databases[127]. No data is perfect but clinical trial databases are considered to be

the best for survival related studies including outcome, disease progression and relapse as the study design avoids selection bias and collects patient information in detail while administrative databases are the weakest[127]. Administrative data cover large populations therefore allowing for generalizability[127]. Results obtained from cohort (prospective) are also generalizable, though they are relatively expensive[127]. Hospital Episode Statistics (HES) is an administrative database that contains patient and clinical details for admitted individuals through the National Health Service (NHS) hospitals in England[135]. Inpatient data from HES was used to study the incidence of community acquired lower respiratory tract infections and community acquired pneumonia in UK adults between April 1997 and March 2011[136]. It has also been used to assemble the datasets used to study the risk of emergency admissions to hospitals, as general practices were linked to HES[137]. HES was also integrated with the general practice research database (GPRD) and Office for National Statistics (ONS) mortality database for studying the incidence of cancer by emergency hospital admissions[138].

The reliability of HES was tested by comparing the data derived for vascular disease in women with general practice records in England from 1 April 1997 to 31 March 2005[139]. Ninety three percent of the women recorded to have vascular disease in GP records were also recorded to have a vascular disease in HES while 97% of the women with no record of vascular disease in GP records had no record in HES for vascular diseases, concluding that HES is a reliable source of information[139]. Another study utilised HES together with other clinical databases to determine the risk of postoperative death in hospital[140]. They concluded that the prediction model using HES had similar discriminatory power as clinical databases[140]. When the quality of HES was evaluated 2.3% admissions in 2003 had missing or invalid data on age, sex, admission method and discharge or admission date[140]. Furthermore, no secondary diagnosis was present for 47.9% and 41.6% of admissions in 1996 and 2003, respectively[140].

There are various concepts to be considered when using comorbidities in evaluating or considering it in clinical decision making[127]. Different comorbidities will have different effects on different cancers with the effect varying during the progression through different cancer stages[127]. For instance, COPD would affect surgical intervention for early stage lung cancer while patients having chemotherapy would be affected by renal conditions[127]. Some comorbidity measures do not include a measure for severity of the condition[127]. Including the extent of severity is important because, for instance, mild COPD is not uncommon in lung cancer patient whereas severe COPD will make the patient unfit for surgical resection[127].

Comorbidities will only be reported if a patient is medically examined[127]. Chronic illnesses that would initiate regular medical check-ups would indirectly increase the chances of detecting cancer[127]. Certain comorbidities would increase the risk of developing cancer while the same condition would not have any impact on the outcome of that cancer[127]. In other words, cancer diagnosis and prognosis is affected by different comorbidities[127]. Other factors such as age, sex, ethnicities and socioeconomic status will affect individuals' comorbidities in more than one way[127].

Comorbidity studies could identify illnesses that could increase the risk of cancer and therefore be screened for, as a preventive measure[128]. This practice could decrease the incidence of cancer with certain severe comorbidities encouraging regular and routine screening[128]. Comorbidity profiling could be a cost effective measure as detecting cancer earlier and treating it would reduce the cost of patient care[128].

Cancer is a heterogeneous disease and survival of lung cancer patients is not only influenced by histology of tumour, stage and age of diagnosis but also comorbidities[129]. Age has a major effect on the diagnosis of cancer and also increased incidence is seen with older age[141, 142].

Cancer may be complicated with age related ailments and long lasting conditions including diabetes, heart disease, hypertension and arthritis[142]. Comorbidities govern the treatment plan of cancer patients leading to a less aggressive treatment choice for cancer patients[143]. Pulmonary and cardiovascular diseases together with diabetes influence cancer survival, identified by a lower resection rate[143]. Furthermore, the morbidity and mortality of resected NSCLC cases is associated with poor pulmonary function and cardiovascular disease[143], with higher prevalence in older cases, while other publications indicated that prevalence of cardiovascular conditions, COPD and DM decrease the resection rate[144]. A comparison of NSCLC survival study indicated a 2-fold risk of death for patients with comorbidity[129]. Distribution of comorbidity indicated that 88.3% have ≥1 comorbidity, 54.3% have ≥3 comorbidities and 22.1% have ≥5 comorbidities in lung cancer patients[129].

It is necessary to evaluate comorbidities in cancer patients because increased number of comorbidities is associated with lower survival[145]. Studies have also reported that prediction of overall survival was linked to comorbidity[145]. All comorbidities in lung cancer are shown to be associated with increased toxicity and total dose reduction of chemotherapy[145]. Furthermore, increased toxicities could affect a patients' prognosis[145].

Comorbidities can therefore affect the diagnosis and prognosis of cancer[128]. It can lead to diagnosis at an earlier stage or affect diagnosis by affecting the presentation of illness and increase complexity during the course of the disease[128]. It can not only reduce the survival of cancer patients but influence all-cause mortality[128]. It can massively affect cancer treatment and vice versa[128]. Milder treatment is an option for those with increased comorbidity[128].

Adjusting for comorbidities, to allow for general applicability without increasing the risk of the patient is required in cancer therapy[146]. Furthermore, since certain comorbidities prevent the

optimal effect of therapy due to toxicity from treatment, they have to be considered in treatment management[146].

2.2.1 Comorbidity Index

Comorbidities can be studied individually or as a summary measure, which can be generalised and used across disease populations or be disease specific[127]. Disease specific comorbidity measures have been created for breast cancer[147] and lung cancer[148] which were developed and tested using a cancer specific patient cohort[127].

Disease specific models are considered to be better than general models as they would explain the outcome of interest better than the generalised model that assumes same impact for various diseases[127]. Furthermore, the choice of comorbidities in a disease specific comorbidity measure would consider that certain comorbidities are not independent of the disease i.e. manifested as a result of the disease[127]. For example, the existence of anaemia, weight loss, pneumonia and electrolyte disorders before cancer development[127].

Many indices have been formulated to collate comorbidities into a useful measure including the BOD index, CIRS, Cornoni-Huntley index, Hallstrom index, Hurwitz index, Incalzi index but the most widely used are the Charlson comorbidity index (CCI) and the Elixhauser comorbidity index (ECI)[126, 149]. Comorbidities present us with an opportunity to further refine prognoses and improve prediction[126]. In cancer research, the ECI has been compared to CCI to evaluate survival of colorectal cancer patients. The result showed that the ECI can form a superior risk adjustment model for predicting survival[150].

Charlson Comorbidity Index: In 1987, Mary Charlson and colleagues designed the Charlson comorbidity index (CCI) by studying the mortality at 1 year as a function of various comorbidities by using the internal medicine inpatient service data[151]. Any medical condition or disease that resulted in a relative risk of death greater than 1.2 was included in the index[151]. This resulted in a list of 17 comorbid conditions with different weights including myocardial infarction (MI), congestive heart failure (CHF), peripheral vascular disease (PVD), cerebrovascular disease (CVD), dementia, chronic pulmonary disease (CPD), connective tissue disease (CTD), peptic ulcer disease, mild liver disease and diabetes: 1, Hemiplegia, moderate or severe renal disease, diabetes with end organ damage, any tumour excluding lung cancer, leukaemia, lymphoma: 2 , moderate or severe liver disease: 3,metastatic solid tumour and AIDS: 6[151]. Depending on the type of disease studied, a slight modification in various publications is seen[141]. The CCI has found its use in various cancers to predict prognosis, survival and treatment, mainly because of its simplicity and ease of applicability[152-154].

Elixhauser Comorbidity Index: In 1998, Anne Elixhauser, developed a set of comorbidity measures to effectively handle and utilise administrative inpatient data[149]. A detailed set of 30 comorbidities were developed that could find its use in grouping individuals based on comorbidities and utilise them as binary indicators for discrete conditions or convert them into a score or index for handling multiple conditions[149]. These comorbidities were determined using inpatient data from 438 hospitals comprising a total of 1,779,167 patients, utilising length of stay, hospital charges and in hospital death as outcome variables[149]. ECI is defined by the presence of congestive heart failure, cardiac arrhythmias, valvular disease, pulmonary

circulation disorders, peripheral vascular disorders, hypertension (uncomplicated), hypertension (complicated), paralysis, neurodegenerative disorders, chronic pulmonary disease, diabetes (uncomplicated), diabetes(complicated), hypothyroidism, renal failure, liver disease, peptic ulcer disease, AIDS/HIV, lymphoma, metastatic cancer, solid tumour without metastasis, rheumatoid arthritis/collagen, coagulopathy, obesity, weight loss, fluid and electrolyte disorders, blood loss anaemia, deficiency anaemia, alcohol abuse, drug abuse, psychosis and depression[149].

Additionally, the CCI[151] has been used to study many cancers (Table 2.1) including survival in early stage lung cancer patients after surgical resection[154], proton therapy[155], and evaluate long term survival[152] and survival after surgery and radiotherapy[156] in NSCLC patients. Additionally, also used to analyse survival in bronchial cell carcinoma[157], head and neck squamous cell carcinoma[158, 159], chronic myeloid leukaemia[160], colorectal cancer[150], colon cancer[161], bladder cancer[162-165], renal cell carcinoma[166], ovarian cancer[167] and prostate cancer patients[168-170].

Table 2.1: Use of Charlson comorbidity index (CCI) in various cancers.

AUTHOR	AIM	METHOD	COMORBIDITY SCORE-CHARLSON COMORBIDITIES	RESULTS AND SUMMARY
Ganti *et al*., 2011[146]	[α]To evaluate the correlation between CCI and survival in **lung cancer**.	A retrospective study on 617 lung cancer patients using Cox proportional hazard model to evaluate the relationship between survival and CCI adjusting for sex, smoking history/pack years, family history of lung cancer, histopathological classification, stage of disease at diagnosis and type of initial treatment.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes **2**- Hemiplegia, Moderate to severe renal disease, Diabetes with end organ damage **3**- Moderate to severe liver disease **6**- AIDS **1**-For each decade over 40 years.	Multivariate analyses depicted that CCI was not associated with the death risk and for CCI≥5, the HR = 1.37 (95% CI: 0.52-3.62; p=0.54). The p value for Charlson index with and without age for every grading was not significant, therefore the prediction of lung cancer survival using CCI was not valid leading to a need of better prognostic models.
Do *et al*., 2010[155]	[α]To study the influence of comorbidity on survival in early stage **lung cancer** patients treated with proton radiotherapy using Charlson comorbidity index.	54 NSCLC patients treated prospectively in a phase II clinical trial with hypofractionated proton therapy were analysed for comorbidities using the Charlson comorbidity index.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	Chronic pulmonary disease was the most prevalent comorbid condition. The predicted survival and the observed comorbidity specific survival (CSS) correlated well, with the 3 year predicted survival based on CCI being 62% and the observed 3 year CSS being 57% with no statistical significance between them. Furthermore, correlation was seen between the mortality predicted by CCI and the observed mortality.

Pujol *et al*., 2008[171]	[αβ] To validate the simplified comorbidity score in a population of **non-small cell lung cancer** patients.	Therapeutic and clinical data were available on 301 non-small cell lung cancer patients	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	In the univariate analysis, shorter survival was predicted by poor PS, advanced stage, weight loss, anaemia, hyperleukocytosis, lymphopenia, high platelet count, high CYFRA 21-1, high NSE, hypoprotidemia, hypoalbuminemia, high LDH(lactate dehydrogenase), high alkaline phosphatases, hyponatremia, hypercalcaemia, high fibrinogen, SCS (simplified comorbidity score) >9, CCI ‡3 and patient LCSS (lung cancer symptom scale) score >22.2 and in the multivariable analyses low survival was predicted by stage grouping(HR= 4.03, 95% CI: 2.40–6.77); CYFRA 21-1 level (HR= 2.30,95% CI: 1.52–3.49); low QoL (quality of life)(HR = 2.20,95% CI: 1.48–3.27); SCS (HR=1.78, 95% CI: 1.21–2.63); anaemia (HR= 1.88, 95% CI: 1.16–3.07); high NSE level (HR= 1.66, 95% CI: 1.12–2.46); low sodium level (HR= 1.99, 95% CI: 1.04–3.77) and high alkaline phosphatases level (HR = 1.53, 95% CI: 1.01–2.32). SCS is more informative than CCI in predicting NSCLC.

Wang *et al*., 2007[154]	[θη]To determine the better predictor of prognosis in patients with stage I **non-small cell lung cancer** resection.	Medical records on 426 patients were used to calculate the KPI (Kaplan—Feinstein index) and the Charlson comorbidity index (CCI) using both the univariate and multivariate analyses.	**1**-CAD, CHF, CPD, Peptic ulcer disease, PVD, Mild liver disease, CVD, CTD, diabetes, dementia **2**- Hemiplegia, Moderate to severe renal disease, Any prior tumour, Leukaemia, Lymphoma **3**-Moderate to severe liver disease **6**-Metastatic solid tumour; AIDS	In univariate analyses, male gender (p= 0.016), patients aged ≥65 years(p=0.002), smokers(p=0.023), CCI score ≥2 (p=0.003), extensive resection and pathological stage IB cancer (p=0.007) had poorer 5-year survival. In multivariate logistic regression analysis, age ≥65 years (HR= 1.4, 95% CI: 1.02-1.93) pneumonectomy (HR=2.42, 95% CI: 1.2-3) CCI score ≥2 (HR= 1.74,95% CI: 1.25-2.42) and stage IB cancer (HR= 1.49, 95% CI: 1.12-1.98) were independent prognostic factors. Patients with CCI ≥2 had higher perioperative mortality and non-cancerous death after resection as compared with patients with CCI<2 whereas KFI had no impact on mortality.
Birim *et al*., 2005[152]	[αβθη]To validate the influence of Charlson comorbidity index on long term survival in operated **non-small cell lung**	Kaplan Meier was used to obtain survival curves and risk factors were determined using univariate and multivariate Cox regression model	**1**-CAD, CHF, CPD, Peptic ulcer disease, PVD, Mild liver disease, CVD, CTD, diabetes, dementia **2**- Hemiplegia, Moderate to severe renal disease, Any prior tumour, Leukaemia, Lymphoma **3**-Moderate to severe liver disease	Univariate analysis depicted that age, male gender, congestive heart failure, chronic pulmonary disease, Charlson comorbidity index, clinical and pathological stage, and type of resection were significantly associated with decreased survival. In the

	cancer patients and determine its efficiency as a predictor of long term survival compared to individual risk factors.		**6**-Metastatic solid tumour; AIDS	multivariate analysis age (RR= 1.02; 95% CI: 1.01–1.03), Charlson comorbidity grade 1–2 (RR= 1.4; 95% CI: 1.0–1.8), Charlson comorbidity grade ≥3 (RR= 2.2; 95% CI: 1.5–3.1), bilobectomy (RR= 1.7; 95% CI: 1.2–2.5), pneumonectomy (RR= 1.5; 95% CI: 1.1–2.0), pathological stage IB (RR= 1.5; 95% CI: 1.1–2.2), IIB (RR= 1.9; 95% CI: 1.2–3.0), IIIA (RR= 1.9; 95% CI: 1.1–3.1), IIIB (RR= 2.8; 95% CI: 1.2–6.8), and IV (RR=12.4; 95% CI: 3.2–48.2), were associated with an impaired survival. Charlson comorbidity index is a better predictor than individual risk factors.
Moro-Sibilot *et al*., 2005[172]	[α]To determine the impact of comorbidity on survival after stage I **non-small cell lung cancer** surgery.	588 patients underwent resection for stage 1 NSCLC. Comorbidity was assessed using the Charlson Index of comorbidity (CCI). Cox proportional hazards model, Kaplan-Meier and the log rank test were used for survival and forward stepwise logistic regression was used with survival as response variable.	**1**-CAD, CHF, CPD, Peptic ulcer disease, PVD, Mild liver disease, CVD, CTD, diabetes, dementia **2**- Hemiplegia, Moderate to severe renal disease, Any prior tumour(within 5 years of diagnosis), Leukaemia, Lymphoma **3**-Moderate to severe liver disease **6**-Metastatic solid tumour; AIDS(not only HIV positive)	No survival differences were seen between CCI=0 and CCI grade 1-2 (p= 0.37), and CCI 3-4 and CCI≥5 (p=0.96) but significant survival differences were detected between CCI 1-2 and CCI 3-4 (p=0.002). Comorbidities after surgical resection have an important impact on survival in stage 1 non small cell lung cancer. The use of CCI is recommended.

Birim *et al*., 2003[173]	[α]To study the influence of Charlson comorbidity index in patients with operated primary **non-small cell lung cancer**.	205 patients who underwent resection for primary non-small cell lung cancer were evaluated ; univariate and multivariate logistic regression was used to determine individual risk factors	**1**-CAD, CHF, CPD, Peptic ulcer disease, PVD, Mild liver disease, CVD, CTD, Diabetes, Dementia **2**-Hemiplegia, Moderate to severe renal disease, Diabetes with end organ damage, Any prior tumour, Leukaemia, Lymphoma **3**-Moderate to severe liver disease **6**-Metastasis solid tumour, AIDS	Univariate analyses showed that gender, prior tumour within 5 years, CCI grade 3-4 and chronic pulmonary disease were significant however in the multivariate analysis only CCI grade 3-4 was significant (OR= 9.8; 95% CI: 2.1-45.9). CCI is the strong predictor of major complications of surgery in NSCLC and is better than the individual risk factors.
Firat *et al*., 2002[156]	[θ]To determine the role of comorbidity in prognosis of **non-small cell lung cancer** for patients treated with surgery and radiotherapy.	Data on 163 patients with stage I non-small cell lung cancer was used to estimate the overall survival and comorbidity. 113 patients underwent surgery and 50 received radiotherapy. Charlson comorbidity index, Cumulative illness rating scale for geriatrics (CIRS-G) and the Karnofsky performance score (KPS) were used to rate comorbidity.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	Significant covariates upon univariate analyses, associated reduced survival include squamous cell histological type ($p < 0.001$), clinical Stage T2 ($p = 0.062$), tumour size > 4 cm ($p = 0.065$), > 40 pack-year tobacco use ($p < 0.001$), presence of a CIRS-G score of 4 ($p < 0.001$), severity index of >2 ($p < 0.001$), Charlson score >2 ($p = 0.004$), KPS <70 ($p < 0.001$),and treatment with RT ($p < 0.001$). Multivariate analyses of all patients with histological features, clinical T stage, age, tobacco use, KPS, CIRS-G and treatment group showed that SCC histological type (RR= 2.3, 95% CI: 1.5-3.5), >40

				pack year tobacco use (RR=2.1,95% CI: 1.3–3.4),KPS <70 (RR=2.7, 95% CI: 1.7–4.2) and presence of CIRS-G (RR=3.4, 95% CI: 2.1–5.3) were independently associated with reduced overall survival.
Sanchez *et al*., 2006[157]	[αη]To study the role of comorbidities in the treatment of **bronchial carcinoma.**	Comorbidity measurement on 305 bronchial carcinoma cases was used. The Torrington-Henderson scale and the Charlson scale were used for patient categorisation on risk or death.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	Logistic regression revealed that the Charlson score (p = 0.001) and BMI (p=0.003) score significantly correlated with complications. In the multivariable analyses FEV_1(p = 0.001) and prolonged air leak(p < 0.001) determined respiratory complications. Charlson score of 3 or 4 and the Torrington-Henderson score of 3 were associated with a greater number of postoperative complications in patients with bronchial carcinoma.
Liu *et al*., 2010[158]	[α]To study the impact of comorbidity on survival in patients with **head and neck squamous cell carcinoma.**	CCI was calculated for 241 patients treated with radiotherapy or radiotherapy and chemotherapy. The overall survival and disease specific survival were calculated.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma.	Higher CCI was associated with older age, fewer years of education and no CT (p<0.05). CCI score ≥ 2(HR= 2.7, 95% CI: 1.7-4.2), stage IV disease (HR = 2.3, 95% CI: 1.2-4.7), a RT dose < 70 Gy (HR = 1.5, 95% CI: 1.1-2.1), and no CT (HR = 1.8, 95% CI: 1.3-2.6) were significant predictors of

			3 - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	poorer overall survival and CCI score ≥ 2 (HR= 2.4, 95% CI: 1.5-3.8), stage IV disease (HR = 2.2,95% CI: 1.1-4.4), a RT dose < 70 Gy (HR =1.5, 95% CI: 1.1-2.1) and no CT (HR =1.9, 95% CI: 1.3-2.7) disease specific survival in multivariate analyses. Comorbidity has a significant impact on survival of patients with NSCLC treated by radiotherapy or radiotherapy and chemotherapy.
Singh *et al*., 1997[159]	[β]To validate the Charlson comorbidity index in **head and neck cancer** patients.	Study conducted on 88 patients. Cox proportional hazard model was used to determine the relative risk of individual risk factors and survival	**1**-MI, CPD, CHF, Ulcer, PVD, Mild liver disease, cerebrovascular accident, Diabetes, Dementia **2**-Hemiplegia, Moderate to severe renal disease, Diabetes with end organ damage, Any tumour, Leukaemia, Lymphoma **3**-Moderate to severe liver disease **6**-Metastatic solid tumour, AIDS	Patients with advanced comorbidity had a RR =2.35 (95% CI: 1.23-4.46; p=0.009) times greater relative risk for cancer related death than low grade comorbidity. Charlson index produced 100% applicability compared to the Kaplan Feinstein index (80%) (p<0.0001). CCI is a valid prognostic indicator of head and neck cancer and is better suited for retrospective study.
Breccia *et al*., 2011[160]	[θ]To utilise Charlson comorbid index to predict the development of pleural	125 elderly patients (>60) with chronic myeloid leukaemia who received dasatinib after imatinib resistance or intolerance were retrospectively evaluated	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate	Significant association between Charlson index and drug reduction or suspension was seen. During dasatinib treatment 49% of score 0 patients saw a reduced dose compared to 63%

	effusions in elderly **chronic myeloid leukemia** patients treated with dosatinib after resistance or intolerance with inatinib.	using the Charlson comorbidity index and adult comorbidity evaluation 27(ACE27).	or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	of patients with score 1, 74% of patients with score 2 and 100% of patients with score 3 or 4 (p=0.0001). Association between the Charlson index and development of pleural effusions were seen. Stratification by use of Charlson index may allow identification of patients with high rate of having major toxicities.
Lieffers *et al*., 2010[150]	[αη]To compare the Charlson and Elixhauser comorbidity measures in **colorectal cancer.**	574 colorectal patients on whom administrative data for cancer, comorbidity and survival (2 and 3 year survival) was available were used and analyses were conducted using robust Poisson regression to analyse survival for both indices.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	A baseline model with age, sex and stage showed a significant discrimination for the 2 and 3 year survival analyses (C statistics, >0.8). Adding the Charlson comorbidities to this baseline model did not show any improvement (2-year survival, p =0.14; 3-year survival, p = 0.17) however adding the Elixhauser comorbidities to the baseline model showed discrimination (2-year survival, p = 0.0051; 3-year survival, p = 0.0017). For survival prediction, Elixhauser method is a better comorbidity risk adjustment model for colorectal cancer.
Hines *et al*., 2009[161]	[αη]To check for association	496 patients underwent surgery for colon cancer.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease,	All three methods produced similar results (ACE-27: HR = 1.63;

	between comorbidities and mortality after **colon cancer** surgery using three different methodologies.	Overall and colon cancer specific mortality was evaluated using the Cox proportional hazard for the three methods namely Adult Comorbidity Evaluation-27 (ACE-27), the National Institute on Aging (NIA) and National Cancer Institute (NCI) Comorbidity Index, and the Charlson Comorbidity Index (CCI).	Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	95% CI: 1.24 - 2.15); (NIA/NCI, HR = 1.83; 95% CI: 1.29 - 2.61); (CCI: HR =1.46; 95% CI: 1.14 - 1.88) Shorter survival after colon cancer surgery was significantly predicted by all three methods
Gore *et al*., 2010[164]	[α]To compare the survival outcomes of patients with **bladder cancer.**	3262 patients over the age of 66 years at diagnosis with stage II muscle invasive bladder cancer were recruited. Use of radical cystectomy studied.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	21% of study subjects underwent radical cystectomy. The overall survival for patients who underwent cystectomy was better than chemotherapy and/or radiation (HR = 1.5, 95% CI: 1.3-1.8) and surveillance (HR = 1.9; 95%CI: 1.6-2.3). The 5 year adjusted survival was 42.2% (95% CI: 39.1%-45.4%), for cystectomy it was 20.7% (95% CI: 18.7%-22.8%), for chemotherapy and/or radiation it was 14.5% (95% CI: 13% -16.2%). The overall survival was better for patients that had cystectomy compared with those who underwent alternative therapy, concluding that many bladder cancer patients might

				benefit from surgery are receiving alternative, less salubrious treatments.
Fisher *et al*., 2009[165]	[α]To study the risk factors for postoperative cardiac complications after cystectomy for **bladder cancer.**	A retrospective review on 283 patients who underwent cystectomy was carried out by considering 12 preoperative risk factors including age, CCI, type of urinary diversion and previous cardiac history. Analysis was carried out using univariate and multivariate analysis	**1**-MI, CPD, CHF, Ulcer, PVD, Mild liver disease, Cerebrovascular accident, Diabetes, Dementia **2**-Hemiplegia, Moderate to severe renal disease, Diabetes with end organ damage, Any tumor, Leukemia, Lymphoma **3**-Moderate to severe liver disease **6**-Metastatic solid tumour, AIDS	POCC risk was associated with ileal conduit urinary diversion (OR = 5.58, 95% CI: 1.23-25.36, p= .026) and Charlson index score (OR =1.28, 95% CI: 1.024-1.60, p= .030) on multivariate analysis. Therefore, patients with a prior cardiac history should be counselled about the increased risk of postoperative cardiac complications.
Koppie *et al*., 2008[162]	[α]To study the survival after cystectomy in **bladder cancer** patients. Age adjusted Charlson comorbidity index (ACCI) was used to characterise the impact of comorbidity and age on disease progression and analyse its association with	1121 patients underwent radical cystectomy for bladder cancer. Logistic regression was used to study the associations of various clinical features. Multivariate logistic regression model was used for overall and progression free survival and Cox proportional hazard model was used for endpoint overall survival analysis.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes **2**- Hemiplegia, Moderate to severe renal disease, Diabetes with end organ damage **3**- Moderate to severe liver disease **6**- AIDS **1** – for each decade over 40 years.	For overall survival the patients with moderate score had a HR= 1.46 (95% CI: 1.20–1.78) compared to the patients low ACCI score and for the patients with high ACCI score the HR= 2.66 (95% CI: 2.00–3.55).There was a significant association between ACCI score and the disease free progression survival (p= 0.03). Emphasises the importance of age and comorbidity in treatment selection and survival and therefore its importance in treatment.

	clinicopathologic and treatment characteristics.			
Miller *et al*., 2003[163]	[α]To study the influence of comorbidities on control and survival of cancer after radical cystectomy for **bladder cancer**.	106 patients with localised disease underwent radical cystectomy. Charlson index was used to assess the preoperative co morbidity. Logistic regression was used to ascertain the relationship between the Charlson index and pathological stage while Cox regression was used to for the 2 survival end points (disease specific and overall).	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour (last 5 years), Leukaemia, Lymphoma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	Bivariate analyses depicted a decreased association of disease specific (HR=1.26; p= 0.049) and overall survival (HR= 1.26; p = 0.016) with Charlson Index. In the multivariate analyses, decreased cancer survival (HR= 1.257; 95% CI: 1.001-1.578; p =0.049) and increased extravesical disease (OR= 0.659; 95% CI: 0.449-0.968; p=0.033) was associated with Charlson index.
Gettman *et al*., 2003[166]	[θ]To study outcome prediction after **renal cell carcinoma** surgery using the Charlson comorbidity index.	303 patients underwent surgical resection. Kaplan Meier was used for survival analyses and multivariate Cox proportional hazard analyses were carried out using Charlson index, sex, age, tumour level, TNM stage, grade, perinephric fat invasion, completeness of resection and surgical era.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour (last 5 years), Leukaemia, Lymphoma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	Significant univariate predictors were age at surgery (p = 0.03), lymph node status (p = 0.005), metastasis (p =0.0001), grade (p = 0.0001), perinephric fat involvement (p = 0.005) and tumour levels 0 versus I through IV (p = 0.056). The final model revealed metastasis (p = 0.0001), grade (p = 0.0001), perinephric fat involvement (p = 0.02) and tumor levels 0 versus I through IV (p = 0.048) as multivariate predictors of cause specific survival.

Tetsche *et al*., 2008[167]	[α]To study the prevalence of co morbidities with respect to stage of the **ovarian cancer** and to evaluate the impact of age and comorbidity on survival by stage.	5213 patients with ovarian cancer on whom comorbid data was available were used in this study. Kaplan Meier survival curves were constructed for every level of Charlson index and staging of the cancer and hazard ratios were computed using the Cox proportional hazard regression method	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	One and five year survival was high with patients without comorbidities. For patients with Charlson mortality score of 1-2 and 3+, the one year MMR (mortality rate ratio) declined from 1.8 to 1.4 and from 2.7 to 2, respectively, after adjusting for age and calendar time and furthermore, declined to 1.3 and 1.8, respectively, after adjusting for stage. Similar decline was seen for the five year survival rate. Mortality was observed in patients with prevalence of co morbidities and severe comorbidities were associated with advanced stage of ovarian cancer.
Wahlgren *et al*., 2010[168]	[θ]To study the impact of pre-treatment comorbidity and post treatment (radiotherapy) health related quality of life score (HRQoL) for **prostate cancer**.	158 patients 5 years after the completion of therapy were used. The association between CCI and the HRQoL was analysed using ANCOVA and multivariate regression was used with tumour stage, tumour grade, diabetes status, and cardiovascular status, CCIs were included as fixed factors, whereas age at treatment, pretreatment PSA, and	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour,	For various aspects of HRQoL, a statistically significant inverse relationship was observed between global health (QL) and CCI ($p<0.01$) and between physical function (PF) and CCI ($p<0.01$). The Charlson score was associated with global health status (QL) ($p=0.0002$), physical function (PF) ($p=0.015$) and emotional function (EF) ($p=0.04$) in the univariate analysis. CCI was

		neoadjuvant hormonal treatment were included as covariates.	AIDS	valid but mainly useful in long term predictive studies. QL, PF and EF were negatively significantly associated with diabetes (p=0.009). Also, diabetes had a stronger impact than cardiovascular status. Upon multivariate analyses, the Charlson CCI score and diabetes remained statistically significant.
Alibhai *et al*., 2008[170]	[η]To define an optimal co morbidity index for **prostate cancer** (the Charlson Index, the Diagnosis Count, the Index of Coexistent Disease (ICED), and the number of medications).	345 men with newly diagnosed prostate cancer cases with information about their comorbidity and treatment were available. The performance of the 4 indices was compared by using it to predict the overall survival and receipt of curative treatment.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	For the receipt of curative treatment, the Gleason score and the PSA level predict the receipt of curative therapy and all the 4 indices depict an association in the univariate analyses (c statistics; $p < 0.05$) however for multivariable models adjusted for age, Gleason score and PSA level only Charlson score appeared significant. For the survival analyses, age, local stage of disease, Gleason score, PSA level, and receipt of curative therapy were associated with survival. All the 4 models showed an association with survival in the univariate analyses and multivariate analyses. The optimal comorbidity index for prostate cancer for curative and

				overall survival analyses still remains to be elucidated.
Kastner *et al*., 2006[169]	[θ]To study the application of Charlson Score in planning the treatment of patients with **prostate cancer**.	1043 patients were used (37 with localised prostate cancer patients). 10 year survival was calculated using the Kaplan Meier method for each Charlson index group. Cox regression analysis was used to check for significance.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	Charlson index was a significant predictor of survival following radical treatment of localised prostate cancer (p=0.005). Confirms the reliability and practicality of the Charlson score in prostate cancer patients under the age of 75 and recommends its use in treatment options for localised prostate cancer patients.
Froehner *et al*., 2003[174]	[α]To compare the American society of anaesthesiologists physical status (ASA) classification with the Charlson score for prediction of survival after radical prostatectomy for **prostate cancer**.	444 patients participated in the study. The ASA categorisation was obtained from the anaesthesia charts and the Charlson index was based on the conditions reported during preoperative risk assessment for cardiopulmonary conditions. Kaplan Meier time event survival curve and Mantel Haenszel HR was calculated for comorbid and overall survival.	**1** - MI, CHF, PVD, CVD, Dementia, CPD, CTD, Peptic ulcer disease, Mild liver disease, Diabetes (without end organ damage). **2** - Diabetes with end organ damage, Hemiplegia, Moderate or severe renal disease, 2nd solid tumour, Leukaemia, Lymphoma, Multiple myeloma. **3** - Moderate or severe liver disease **6** - 2nd metastatic solid tumour, AIDS	There was no significant difference in the comorbid and overall mortality with respect to age. However, ASA 3 (Comorbid: HR= 17.68; 95% CI: 4.13–75.80; p <0.01) and Overall: HR = 7.21; 95% CI: 2.39–21.76, p <0.01) and Charlson score 2+ (Comorbid: HR= 17.68, 95% CI: 5.08–61.52; p <0.01) and Overall: HR = 2.83; 95% CI: 1.18–6.82; p=0.02) was significant in both the computed mortalities. When analysed with respect to various age groups, no significant difference was seen however, increased mortality was

				seen for a second cancer when the age group over 70 years was compared to 60-69 years (HR = 11.54, 95% CI: 1.78-74.95; p=0.01)

MI-Myocardial infarction; CAD – Coronary artery disease; CHF – Congestive heart failure; PVD – Peripheral vascular disease; CVD – Cerebrovascular disease; CPD – Chronic pulmonary disease; CTD – Connective tissue disease; α –Survival/risk factor; β –Validation; η - Methodology comparison; θ – Prediction of prognosis/treatment.

Cox Proportional Hazard Model:

The Cox Proportional Hazard model is given by

$$\lambda(p) = \lambda_o(p) \exp(\sum \beta a) \quad [175]$$

λ(p) is the event rate at time p expressed as the function of risk variables, λ_0(p) is the baseline event rate and $\exp(\sum \beta a)$ is the proportionality constant indicator for the risk factors[175]. Since $\lambda_o(p)$ is unspecified, the model is semi parametric and is used widely as the effect can be estimated without the knowledge of $\lambda_o(p)$. The robustness of this model makes it popular because it fit the data well[175].

The Cox model survival function is given by

$$S(p) = S_o(p)^{\exp \sum_{m=1}^{t} \beta_m a_m} \quad [175]$$

The hazard ratio (HR) is defined as the ratio of hazard for one individual to the hazard for another individual[175].

$$\text{HR} = \frac{\lambda(p)^*}{\lambda(p)}$$

$$= \frac{\hat{\lambda}_o(p) \exp(\sum \beta a^*)}{\lambda_o(p) \exp(\sum \beta a)}$$

$$= e^{\sum_{k=1}^{t} \beta_k (a_k^* - a_k)} \quad [175]$$

2.3 Incidence Analysis

2.3.1 Material and Methods

The study population were residents of Liverpool area recruited through the Liverpool lung project (LLP). The HES database records every hospital admission in England. Such information was available for the 10,808 individuals in the LLP cohort admitted through the inpatient, outpatient and accident and emergency. The diagnosis in the HES database were recorded using the International Classification of Disease (version 10) introduced by the World Health Organisation[176]. The diagnosis codes were used to calculate the CCI and ECI[177].

The distribution of population characteristics between cases and controls were evaluated using the Pearson's X^2-test and Fisher's exact test was used for cells with values less than 5. The Cox proportional hazard model was used to evaluate the effect of previous medical conditions on the development of lung cancer in individuals that were free of lung cancer at the start of the study[178]. The study period spanned from 01 January 1999 to 31 March 2010 and individuals with the reported ICD-10 code "C34" and "C780" were classified as cases. The time variable for the Cox proportional hazard regression was the time spent by each individual at risk of developing lung cancer for the duration of the study, unless death occurred before the study ended.

The CCI score (CCIS) was calculated as the sum of the weight of comorbidities reported by an individual as defined by the CCI while the ECI score (ECIS) was calculated as the total number of comorbidities reported for an individual, defined by the ECI. For both indices,

the distribution of comorbidities between cases and controls were evaluated. The indices were computed for every individual and were grouped into three categories, 0, 1-2 and ≥3.The index scores were tested to determine the risk posed by each of them in a univariate and multivariate model, after adjusting for age (as on 01 January 1999), smoking pack years and gender. All analyses were conducted using Stata version 12[179].

2.3.2 Results

The study population comprised of 9533 individuals of which 1389 (14.57%) developed lung cancer (cases) and 8144 (85.43%) did not develop lung cancer (controls). The majority of the study population (50.2%) were female. The mean age of the cohort at the start of the study was 59.30 (standard deviation (SD) = 7.93) and the mean smoking pack years was 20.81 (SD = 23.83). Two thousand eight hundred and two (32.2%) of individuals were non-smokers, of these 94 (8.6 %) were cases and 2708 (35.5%) were controls. Table 2.2 represents the distribution of various covariates, results depict that the distribution of gender and groups of smoking pack years, ECIS and CCIS, is significantly different in cases and controls. Of the total individuals, 5001 (52.5%) individuals had a CCI score of 0, 2940 (30.8%) had a CCI score of 1-2, 2784 (8.2%) had a CCI score of 3-4 and 808 (8.5%) individuals had a CCI score of ≥5 while 3759 (39.4%) individuals had an ECI score of 0, 3468 (36.4%) had an ECI score of 1-2 , 1540 (16.2%) had an ECI score of 3-4 and 766 (8%) had an ECS score of ≥5.

Table 2.3 represents the distribution of Charlson comorbidities with chronic pulmonary condition (22.7%), any malignancy, including lymphoma and leukaemia, except malignant neoplasm of skin and lung cancer (11%) and diabetes (without complications) (10.3%) as

the three most reported conditions while Table 2.4 represents the distribution of ECI with hypertension (uncomplicated) (33.3%), chronic pulmonary condition (31.2%), and cardiac arrhythmias (11.6%) as the three most reported conditions, respectively.

Table 2.2: Patient characteristics for LLP cohort.

Covariates	**Cases (N%)**	**Controls (N%)**	**Total (N%)**	**p-value**
Gender				
Female	606 (43.6)	4176 (51.3)	4782 (50.2)	
Male	783 (56.4)	3968 (48.7)	4751 (49.8)	
Total	1389 (100)	8144 (100)	9533 (100)	<0.0001
Age (start of study)				
≤60	530 (38.2)	4651 (57.1)	5181 (54.3)	
>60	859 (61.8)	3493 (42.9)	4352 (45.7)	
Total	1389 (100)	8144 (100)	9533 (100)	<0.0001
Smoking duration (years)				
0	94 (8.6)	2708 (35.5)	2802 (32.2)	
1-19	141 (13)	2003 (26.3)	2144 (24.6)	
20-39	350 (32.2)	1698 (22.3)	2048 (23.5)	
40-59	318 (29.2)	847 (11.1)	1165 (13.4)	
60+	185 (17)	369 (4.8)	554 (6.4)	
Total	1088 (100)	7625 (100)	8713 (100)	<0.0001
CCI score				
0	502 (36.1)	4499 (55.2)	5001 (52.5)	
1-2	450 (32.4)	2490 (30.6)	2940 (30.8)	
3-4	83 (6)	701 (8.6)	784 (8.2)	
≥5	354 (25.5)	454 (5.6)	808 (8.5)	
Total	1389 (100)	8144 (100)	9533 (100)	<0.0001
ECI score				
0	329 (23.7)	3430 (42.1)	3759 (39.4)	
1-2	691 (49.7)	2777 (34.1)	3468 (36.4)	
3-4	285 (20.5)	1255 (15.4)	1540 (16.2)	
≥5	84 (6)	682 (8.4)	766 (8)	
Total	1389 (100)	8144 (100)	9533 (100)	<0.0001

Table 2.3: Frequency distribution of Charlson comorbidities.

CCI comorbidities	Cases (N%)	Controls (N%)	p-value
Myocardial Infarction	81 (5.8)	639 (7.8)	0.009
Congestive Heart Failure	38 (2.7)	507 (6.2)	<0.0001
Peripheral Vascular Disease	123 (8.9)	463 (5.7)	<0.0001
Cerebrovascular Disease	63 (4.5)	436 (5.4)	0.206
Dementia*	3 (0.2)	100 (1.2)	<0.0001
Chronic Pulmonary Disease	434 (31.2)	1725 (21.2)	<0.0001
Connective Tissue Disease	37 (2.7)	208 (2.6)	0.811
Peptic Ulcer Disease	31 (2.2)	172 (2.1)	0.775
Mild Liver Disease	21 (1.5)	97 (1.2)	0.318
Diabetes (without complications)	139 (10)	845 (10.4)	0.676
Diabetes (with end organ damage)	16 (1.2)	119 (1.5)	0.367
Hemiplegia	13 (0.9)	100 (1.2)	0.353
Moderate or severe renal disease	25 (1.8)	263 (3.2)	0.004
Any malignancy, including lymphoma and leukaemia, except malignant neoplasm of skin and lung cancer	169 (12.2)	873 (10.7)	0.11
Moderate or severe liver disease	5 (0.4)	21 (0.3)	0.5
Metastatic solid tumour	334 (24)	187 (2.3)	<0.0001
AIDS	_	_	_

p values were derived using the chi-square test unless otherwise stated.*Fisher's exact test was used

Table 2.4: Frequency distribution of Elixhauser comorbidities.

ECI comorbidities	Cases (N%)	Controls (N%)	X^2 p-value
Congestive heart failure	38 (2.7)	507 (6.2)	<0.0001
Cardiac arrhythmias	161 (11.6)	1163 (14.3)	0.007
Valvular disease	39 (2.8)	309 (3.8)	0.07
Pulmonary circulation disorders	16 (1.2)	155 (1.9)	0.051
Peripheral vascular disorders	123 (8.9)	463 (5.7)	<0.0001
Hypertension (uncomplicated)	462 (33.3)	2837 (34.8)	0.254
Hypertension(complicated)	10 (0.7)	164 (2)	0.001
Paralysis	13 (0.9)	100 (1.2)	0.353
Neurodegenerative disorders	32 (2.3)	239 (2.9)	0.191
Chronic pulmonary disease	434 (31.2)	1725 (21.2)	<0.0001
Diabetes (uncomplicated)	139 (10)	839 (10.3)	0.738
Diabetes(complicated)	16 (1.2)	129 (1.6)	0.224
Hypothyroidism	52 (3.7)	347 (4.3)	0.374
Renal failure	25 (1.8)	263 (3.2)	0.004
Liver disease	23 (1.7)	108 (1.3)	0.329
Peptic ulcer disease	26 (1.9)	151 (1.9)	0.964
AIDS/HIV	_	_	_
Lymphoma*	4 (0.3)	70 (0.9)	0.020
Metastatic cancer	334 (24)	187 (2.3)	<0.0001
Solid tumour without metastasis	160 (11.5)	799 (9.8)	0.05
Rheumatoid arthritis/collagen	48 (3.5)	283 (3.5)	0.971
Coagulopathy	7 (0.5)	40 (0.5)	0.95
Obesity	29 (2.1)	288 (3.5)	0.005
Weight loss	87 (6.3)	264 (3.2)	<0.0001
Fluid and electrolyte disorders	43 (3.1)	335 (4.1)	0.072
Blood loss anaemia	_	_	_
Deficiency anaemia	37 (2.7)	248 (3)	0.44
Alcohol abuse	45 (3.2)	248 (3)	0.698
Drug abuse*	4 (0.3)	9 (0.1)	0.108
Psychosis	6 (0.4)	28 (0.3)	0.61
Depression	34 (2.4)	262 (3.2)	0.127

*Fisher's exact test was used

The CCIS and ECIS (Table 2.5) were evaluated in a univariate and multivariate analysis after adjusting for age, sex and smoking pack years in a Cox proportional hazard regression analysis. The hazard ratio (HR) for the univariate analysis was 1.63 (95% CI: 1.43-1.85) for CCIS 1-2 and 3.71 (95% CI: 3.26 – 4.22) for CCIS ≥3, and in the multivariate analysis the HR

produced were 1.32 (95% CI: 1.14 – 1.53) for CCIS 1-2 and 2.46 (95% CI: 2.11 – 2.88) for CCIS ≥3. Similarly, hazard ratio was 2.51 (95% CI: 2.20 – 2.86) for ECIS 1-2 and 2.15 (95% CI: 1.85 – 2.49) for ECIS ≥3 in the univariate and 2.09 (95% CI: 1.79 – 2.43) for ECIS 1-2 and 1.41 (95% CI: 1.18- 1.68) for ECIS ≥ 3 in the multivariate analysis.

Table 2.5: Regression analysis of Charlson and Elixhauser comorbidity index.

Index score	Univariate analysis HR (95% CI)	Multivariate analysis* HR (95% CI)
CCI		
1-2	1.63 (1.43-1.85)	1.32 (1.14 – 1.53)
≥3	3.71 (3.26 – 4.22)	2.46 (2.11 – 2.88)
ECI		
1-2	2.51 (2.20 – 2.86)	2.09 (1.79 – 2.43)
≥3	2.15 (1.85 – 2.49)	1.41 (1.18- 1.68)

*Multivariate analysis adjusted for age at study start, sex and smoking pack years

2.3.3 Discussion

The distribution of various comorbidities forming the CCI and ECI have been reported and their effect studied using the univariate and multivariate Cox proportional hazard regression analysis adjusted for age at study start, sex and smoking pack years.

Chronic pulmonary condition was the most reported comorbidity among all the CCI comorbidities, which has been previously suggested in lung cancer susceptibility[18]. Although the pathogenesis of lung cancer is yet to be elucidated, it has been hypothesized that chronic airway inflammation induced by respiratory infections may contribute to the alterations in the bronchial epithelium and lung environment, thus provoking a milieu conducive to lung carcinogenesis[11].

The CCI results depict an increased risk of developing lung cancer with increase in score. Although to the best of our knowledge no study has investigated the impact of CCI and ECI in predicting the incidence of lung cancer, the association of lung cancer and CCI has been previously studies[145]. The study was conducted on 1719 cases and 6876 controls using logistic regression that produced an OR of 2.07 (95% CI: 1.78 -2.40) for CCI score 1-2, and an OR of 2.12 (95% CI: 1.67 – 2.68) for CCI score ≥ 3, when compared to the baseline of CCI score 0[145]. A similar study was also published by Ording *et al*. (2012)[180] which was a nested case-control study that evaluated the impact of using the CCI on the incidence of breast cancer. Their study included 46,324 cases and 463,240 population controls of Danish women aged 45-85. They concluded that there was no substantial association between comorbidity measured with the CCI and breast cancer risk[180].

An increased risk of developing lung cancer with ECI scores 1-2 and ≥ 3 was observed. But with the increase in the ECI score the HR for contracting lung cancer decreased. Although the ECI identified 30 comorbidities, only six of the identified comorbidities had frequencies ≥ to 10% among individuals that developed lung cancer.

The strengths of this study include the population-based design, the large sample size, the long follow-up period and the use of HES data, minimising the chances of missing information on comorbidities. In addition, detailed information about potential risk factors in the LLP was collected using standardised questionnaires.

In conclusion, CCI and ECI produced significant results in the subgroup analysis indicating their use in lung cancer incidence studies. CCI was better than ECI as increased hazards were seen as the scores increased for CCI but not for ECI. However, a validation study using another comorbidity dataset derived from a different source, for instance clinical, would help judge the reliability of the dataset for future use.

2.4 Risk Model Development

2.4.1 Introduction

Patient care can be improved if cancer can be diagnosed and treated at an earlier and curable stage[178]. Although smoking is a major risk factor for lung cancer, smoking status and history alone cannot predict the risk of developing lung cancer because not every individual who smokes or has had a smoking history, develop lung cancer[178]. Furthermore, with the decreasing number of smokers, the incidence of lung cancer is still on the increase supporting the fact that other risk factors such as environmental tobacco smoke (ETS), asbestos exposure and genetic predisposition may play important role in the pathogenesis of lung cancer[178]. Therefore, a composite measurement or risk estimation using risk models that include covariates contributing to a persons' risk of developing lung cancer is warranted[178].

Risk prediction models may find their use in clinical settings to identify individuals at high risk or to select individuals that would really benefit from and improve the outcome of clinical trials[178]. Risk prediction models have been developed for many cancers including colorectal, melanoma, ovarian, prostate and breast[117, 178]. For instance, the Gail model developed for breast cancer is used to advise women with a high risk score to undergo screening or genetic evaluation[117].

Current lung cancer risk prediction model include the Bach model[124], the Spitz model[48] and the LLP model[181], that differ from each other by the population used for development, covariates, statistical model and time period for which the predictive risk can be

estimated[178]. The Bach model was developed to predict lung cancer incidence and the probability of a non-lung cancer mortality using over 14,000 individuals enrolled in the β carotene and retinol efficacy trial (CARET)[124]. Cox proportional hazard regression was used to develop a one year risk estimation including age, gender, number of cigarettes smoked per day, number of years smoked and exposure to asbestos[124]. Absolute risk using this model is calculated for smokers by running the incidence and mortality models recursively for the number of times corresponding to the years of risk estimation[124]. When validated internally, the model produced an area under receiver operative curve (AUC) value of 0.72 and an AUC of 0.69[182] and 0.66[183] when validated externally.

The Spitz model was developed using 1851 cases and 2001 hospital-matched controls from the University of Texas MD Anderson cancer centre[48]. Separate models were developed for former, current and never smokers using logistic regression to obtain an absolute risk of developing lung cancer[48]. Covariates in the model included smoking pack years, family history of cancer, asbestos and wood dust exposure, previous emphysema and previous hay fever[48]. Absolute risks over a predefined time period was developed using baseline relative risks together with age and smoking adjusted gender specific incidence rates[48]. The model was developed using and therefore applicable to Caucasians[48]. Internal validation of this model produced an AUC of 0.59, 0.63 and 0.65 for never, former and current smokers[48] while the external validation produced an AUC of 0.69 for the overall model[183].

Finally, the LLP model was developed using 579 lung cancer cases and 1157 controls recruited from Liverpool, UK[181]. Significant covariates included in this model are number of years smoked, family history of lung cancer, occupational exposure to asbestos, prior non-malignant tumour and prior pneumonia[181]. Relative risks obtained using logistic regression model together with population incidence rates for different combination of age and gender was used to estimate the 5-year absolute risk of developing lung cancer[181]. The

model produced an AUC of 0.70[181] when validated internally and 0.69[183] when validated externally, in the overall model. The model also produced an AUC of 0.76 (95% CI: 0.75-0.78) in the Harvard population and AUC of 0.82 (95% CI: 0.80-0.85) in the LLP population based prospective cohort (LLPC) study[184]. The model was developed in Caucasians and therefore can only be applied to Caucasians[181].

Another study included the prostate lung colorectal ovary (PLCO) screening trial to design a model for the general population (N=70,962) and for ever-smokers (N=38,254) using age, education, BMI, family history of lung cancer, COPD, recent chest X-rays, smoking status (never, former or ever), pack-years smoked and smoking duration[185]. For the smokers' only model, time of quitting smoking was included[185]. Logistic regression was used to develop the model[185]. The model for the general population produced an AUC of 0.57 and 0.841 for the internal and external validation, respectively while for the smokers only model, the AUC was 0.805 and 0.784 for the internal and external validation, respectively[185].

Lung cancer risk models have also found their use in CT screening trials[178]. Identifying individuals having high risk and considering them for a CT trial would reduce the incidence of lung cancer and aid early detection and treatment of lung cancer[178]. Application of CT screening in lung cancer would be possible only if the current CT trials are designed appropriately[178]. That includes enrolling individuals with high risk in the trial with the hope of obtaining positive results, increasing efficiency, improving healthcare by reducing lung cancer morbidity and mortality[178]. Risk models can also be used by clinicians to decide on interventions and encourage high risk individuals to adopt healthier habits[178].

2.4.2 Material and Methods

Sex specific Cox proportional hazard regression models were used to design the incidence model using data collected over a period of 11.25 years from 01 January 1999 to 31 March 2010. Covariates used were age at study start, chronic pulmonary disease and smoking pack years. Comorbidity information was extracted using the HES database and participants were filtered to keep individuals aged between 45-79 and filter out individuals that had less than 5 years of incidence time. Participants were confirmed as cases if a recorded ICD code of "C34" or "C780" was reported while a case of chronic pulmonary condition was identified under the ICD code I27.8, I27.9, J40–J47, J60–J67, J68.4, J70.1 and J70.3[177].

Model and Point System Development:

Covariates were tested in a univariate Cox proportional hazard model and only the significant covariates ($p < 0.05$) were included in the multivariate Cox proportional hazard model. The method described in Sullivan *et al*. (2004)[186] was used in the developing this model. There are various steps involved in the development of a point based system of risk prediction[186]. The point based system is also supplemented with corresponding risk estimates, to extract the risk associated with the presence of a particular comorbidity, for being of a particular age and gender, and for smoking, adjusted as smoking pack years[186]. The process begins with selecting covariates that would be included in the sex specific risk prediction model. Covariates significant in the univariate analysis were selected to be included in the risk model[186]. This is followed by the determination of categories and selecting baseline value for the base category for each covariate[186]. The age inclusion

criteria for participating in LLP are between 45-79 years. Therefore, individuals between 45 and 79, inclusive, on the first day of the study were included. The age categories were developed using an interval of 4.99 years. The reference age for this covariate was 47.5, the mid-point of the base category. The reference age for each of the remaining categories was considered to be the mid-point of the category, calculated as the average of the extreme values of the range representing that category[186].

For smoking pack years, the base category was non-smokers. The categories for smoking pack years include, minimum value to 20.99, 21-40.99; 41-60.99, and 61 to the maximum value. Smoking pack years categories were treated as a factor in the prediction model using non-smokers as baseline, producing a regression coefficient for the remaining categories.

Chronic pulmonary condition is an important comorbidity in lung cancer incidence[48, 124, 181]. Retaining individuals with more than 5-year worth of comorbid information in the study ensures that the chronic pulmonary condition was diagnosed within a minimum of 5 years of lung cancer diagnosis and that the analysis does not suffer due to the lack of information. Every category is then presented in terms of baseline values[186]. If β is the regression coefficient, W_c is the reference value for one category and W_f the reference value from the base category, the above process is carried out by subtracting W_f from W_c and multiplying this difference by β. i.e. β (W_c - W_f)[186].To determine the risk of developing lung cancer in 5 years, a constant C is required[186]. It is calculated by multiplying the regression coefficient for age by 5 i.e. C = 5 * (regression coefficient for age)[186]. Dividing β (W_c - W_f) by C gives the point associated with each category[186]. For each risk profile available, the risk score is calculated by summing the points associated with each covariate[186]. To estimate the probability of developing lung cancer associated with the point total, the following formula is used[186].

$$p = 1 - S_0(t)^{exp(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta \bar{X}_i)}$$

where $S_0(t)$ is the baseline hazard function at time 11.25 years calculated using the mean value of the risk factors by the "survival"[274,275] library in R[229]. $\sum_{i=1}^{p} \beta \bar{X}_i$ is the sum of the product of regression coefficients and means or proportions of the covariates and $\sum_{i=1}^{p} \beta_i X_i$ is calculated by multiplying the regression coefficient for age with the reference age of the baseline category and adding the product of C with the point total[186]. The model was developed in R[229].

Model Performance:

The sex specific Cox proportional hazard models were tested for its ability to discriminate between cases and controls by using the concordance statistics (C-statistics)[187, 188]. The test demonstrates the probability that, if two observations are picked at random, the one with the shortest survival will have the largest risk. The measure is similar to area under the receiver operating characteristic curve (AUC) for logistic regression model[187, 188]. Since the discrimination statistics are performed on the same dataset, a 10 fold cross validation was conducted using the original incidence dataset to obtain the C-statistics[189] as described below.

The lung cancer incidence indicator was used to subset the data into training and testing dataset. The lung cancer cases and controls were randomly sampled and divided into 10 equal parts. A part of cases and controls were combined to form the testing set while the rest was used as the training set. This is repeated 10 times with each group of cases and controls appearing exactly once for validation. R[229] package "cvAUC"[357] was used for obtaining area under the receiver operating curve (AUC) estimate for both the sex specific risk predictor models.

2.4.2 Results

The study developed sex specific Cox proportional hazard regression models to predict the risk of developing lung cancer for 5 years using males 4112 and 4306 females. Table 2.6 describes the distribution of covariates for males and females for various covariate categories. The distribution was statistically significant between cases and controls for various categorical groups of smoking pack years, chronic pulmonary condition and age.

Table2.6: Distribution of population characteristics for both genders

Covariates	**cases**	**controls**	**X^2 p-value**
Males			
Age			
≤60	200	2008	
> 60	277	1627	<0.0001
Smoking pack year			
Non-smoker	39	1063	
0.05-20.99	62	1053	
21-40.99	139	814	
41-60.99	145	438	
61-114.86	82	246	<0.0001
Chronic pulmonary disease			
Absent	325	2932	
Present	152	703	<0.0001
Females			
Age			
≤60	169	2362	
> 60	235	1540	<0.0001
Smoking pack year			
Non-smoker	47	1624	
0.05-20.99	69	1031	
21-40.99	146	831	
41-60.99	102	335	
61-114.86	36	81	<0.0001
Chronic pulmonary disease			
Absent	270	3144	
Present	134	758	<0.0001

The sex specific Cox proportional hazard models produced significant results for age, smoking pack years and chronic pulmonary condition for the univariate analysis in both genders and only in men, for the multivariate analysis (Table 2.7). For men, the 11.25-year baseline survival was 0.89 and 0.93 for women. The univariate and multivariate HR for age was 1.05 (95% CI: 1.04-1.06) and 1.05 (95% CI: 1.04-1.06), smoking pack years was 1.02 (95% CI: 1.02-1.02) and 1.02 (95% CI: 1.01-1.02) and chronic pulmonary condition was 1.95 (95% CI: 1.61-2.36) and 1.27 (95% CI: 1.03-1.56) in men while in women the univariate and multivariate HR for age was 1.06 (95% CI: 1.04-1.07)and 1.06 (95% CI: 1.04-1.07), smoking pack years was 1.03 (95% CI: 1.03-1.04) and 1.03 (95% CI: 1.03-1.04) and chronic pulmonary condition was 2.05 (95% CI: 1.67-2.53) and 1.22 (95% CI: 0.98-1.52), respectively.

Table2.7: Sex specific Cox proportional hazard regression model

Covariates	**HR(95% CI)$^{\$}$**	**HR(95% CI)***
Men [$S_0(11.25) = 0.89$]		
Age	1.05 (1.04-1.06)	1.05 (1.04-1.06)
Smoking pack years	1.02 (1.02-1.02)	1.02 (1.01-1.02)
Chronic pulmonary disease	1.95 (1.61-2.36)	1.27 (1.03-1.56)
Women [$S_0(11.25) = 0.93$]		
Age	1.06 (1.04-1.07)	1.06 (1.04-1.07)
Smoking pack years	1.03 (1.03-1.04)	1.03 (1.03-1.04)
Chronic pulmonary disease	2.05 (1.67-2.53)	1.22 (0.98-1.52)

$^{\$}$univariate; * multivariate

Table 2.8 display the proportion or means of covariates used in the risk prediction model. These are needed later when the point based system is compared to the Cox proportion model risk estimate, for two case studies.

Table2.8: Mean/Proportion distribution of model covariates for males and females

Covariates		Means or proportions	
		Female	**Male**
Age		58.45	59.38
Smoking pack years	Non-smoker	0.39	0.27
	Minimum -20.99	0.26	0.27
	21-40.99	0.23	0.23
	41-60.99	0.1	0.14
	≥61	0.03	0.09
Chronic pulmonary disease	Present	0.21	0.21
	Absent	0.79	0.79

Table 2.9 shows the male Cox proportional hazard model β coefficient and the point value for covariate subgroups for men. The points ranged from 0-6 for various age categories with 47.5 years as the baseline value, 0 – 8.78 for various smoking pack years categories and 0 – 0.81 for categories of chronic pulmonary condition.

Table2.9: Male Cox proportional hazard model beta coefficient and point value for covariate subgroups

Covariate	Categories	Reference	Beta	Beta (Reference - W_f)	Points
Age					
	45-49.99	47.5	0.047	0	0
	50-54.99	52.5		0.235	1
	55-59.99	57.5		0.47	2
	60-64.99	62.49		0.70453	3
	65-69.99	67.5		0.94	4
	70-74.99	72.5		1.175	5
	75-79.99	77.5		1.41	6
Smoking pack years					
Non-smoker	No	0	Baseline	0	0
	Yes	1			
0.05-20.99	No	0			
	Yes	1	0.392	0.392	1.67
21-40.99	No	0			
	Yes	1	1.404	1.404	5.97
41-60.99	No	0			
	Yes	1	2.009	2.009	8.55
61-256	No	0			
	Yes	1	2.064	2.064	8.78
Chronic pulmonary disease					
	Absent	0		0	0
	Present	1	0.19	0.19	0.81

Table 2.10 shows the Cox proportional hazard model beta coefficient and point value for covariate subgroups in women. The points ranged from 0-6 for various age categories with 47.5 years as the baseline value, 0 – 9.32 for various smoking pack years categories and 0 – 0.73 for categories of chronic pulmonary condition for men and for women the points ranged from 0 - 6 for various age categories, 0 - 8.5 for various smoking pack years categories and 0-0.73 for categories of chronic pulmonary condition (Table 2.10).

Table2.10: Female Cox proportional hazard model beta coefficient and point value for covariate subgroups

Covariate	Categories	Reference	Beta	Beta (Reference - W_f)	Points
Age					
	45-49.99	47.5	0.058	0	0
	50-54.99	52.5		0.29	1
	55-59.99	57.5		0.58	2
	60-64.99	62.49		0.86942	3
	65-69.99	67.5		1.16	4
	70-74.99	72.5		1.45	5
	75-79.99	77.5		1.74	6
Smoking pack years					
Non-smoker	No	0	Baseline	0	0
	Yes	1			
0.05-20.99	No	0	0.834		
	Yes	1		0.834	2.88
21-40.99	No	0	1.763		
	Yes	1		1.763	6.08
41-60.99	No	0	2.208		
	Yes	1		2.208	7.61
61-136	No	0	2.703		
	Yes	1		2.703	9.32
Chronic pulmonary disease					
	Absent	0	0.212	0	0
	Present	1		0.212	0.73

These models were tested for its discriminatory ability using C statistics in a 10 fold cross validation. The c-statistics was 0.77 (95% CI: 0.74-0.79; standard error = 0.012) for women and 0.73 (95% CI: 0.71 - 0.75, standard error = 0.0113) for men.

2.4.3 Discussion

Given that the lung cancer cases are diagnosed at an advanced stage where no treatment is available to revert the condition, early detection or prevention is the only option[190]. Previous lung cancer risk prediction models used logistic regression to develop risk models[50, 123, 124, 185] for different populations and validated using other population[184]. Bach *et al*. (2003)[124] and Park *et al*. (2013)[191] are the only incidence model that used a Cox proportional hazard regression model. Bach[124] developed a 10 year risk prediction model (refer above) using individuals of a carotene trial while Park *et al*. (2013)[191] used 1,324,804 Korean men to develop a model using smoking exposure, age at smoking initiation, BMI, physical activity and fasting glucose level, and produced a performance statistics of 0.871 (95% CI: 0.867-0.876). The drawback of the Bach model was that it was developed using all smokers from a high risk population in a trial with mixed ethnicities[124]. Furthermore, this model was developed for males and females, separately, while Bach developed a general model. The 5 year sex specific incidence model was developed using data collected over a period of 11.25 years and important covariates like age and smoking pack years, implicated in lung cancer. The C statistics for the sex specific models suggests that the model has a good discriminatory power.

The developed model was converted into a point based system where given the measurement for a particular covariate, it can be converted into points and the corresponding risk estimate can be obtained (Table 2.11). For instance (Appendix) a 60 year old male with a reported chronic pulmonary condition and a smoking pack year value of 24 has a point based risk estimate of 19.87% while the Cox proportional regression model produces a risk estimate of 20.68%. For a 50 year old female with no reported history of chronic pulmonary condition and a smoking pack year value of 34 produced a point based

risk estimate of 10.12% and a Cox proportional hazard risk estimate of 9.48%. This is the first lung cancer risk prediction model that has developed a point based system and the close risk values by the two system further validates that the point based system is a valid tool applicable in a clinical practice.

Table 2.11: Lung cancer points with corresponding risk estimate

Male		**Female**	
Points	Risk (%)	Points	Risk (%)
0	2.1	0	1.36
1	2.65	1	1.81
2	3.35	2	2.41
3	4.21	3	3.22
4	5.3	4	4.27
5	6.66	5	5.67
6	8.34	6	7.5
6.97	10.37	7.08	10.12
7.97	12.93	8.08	13.29
8.97	16.06	9.08	17.35
9.97	19.87	10.08	22.48
10.97	24.43	11.05	28.64
11.97	29.84	12.05	36.3
13.36	38.82	13.05	45.26
14.36	46.28	14.05	55.31
15.36	54.44	15.06	65.91
		16.05	76.27

Although the information about chronic pulmonary condition was collected using HES dataset, the information about smoking pack years is subjected to recall bias. This model is specifically designed using Caucasian populations, though it still needs to be validated, it should also be tested for its applicability in other populations. In conclusion, this is a first

sex specific 5-year Cox proportional absolute risk prediction model in lung cancer designed using both smokers and non-smokers forming a part of the general population.

2.5 Appendix

The following are 2 cases depicting how the risk calculated using the point system relates to the actual risk using the Cox proportional hazard regression model.

Case 1: risk factor for a LLP male

Risk factor	Value	Point
Age	60	3
Smoking pack years	24	5.97
Chronic pulmonary condition	1	0.81

$$p = 1 - S_0(t)^{exp(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta \bar{X}_i)}$$

$$= 1 - (0.89)^{1.987664}$$

$$= 20.67605$$

The risk estimated by the point based system (Table 2.11) for the above male case is 19.87% and by the Cox regression model is 20.67605%.

Case 2: risk factor for a LLP female

Risk factors	Value	Point
Age	50	1
Smoking pack years	34	6.08
Chronic pulmonary condition	0	0

$$p = 1 - S_0(t)^{exp(\sum_{i=1}^{p} \beta_i X_i - \sum_{i=1}^{p} \beta \bar{X}_i)}$$

$$= 1 - (0.93)^{1.371821}$$

$$= 1 - 0.905241$$

$$= 0.094759$$

The risk estimated by the point based system for the above female case is 10.12% and by the Cox regression model is 9.48%.

CHAPTER 3

GENOMEWIDE CASE CONTROL ASSOCIATION ANALYSIS

3.1 Aim

In 2008, lung cancer was the leading cause of death due to cancer in males and the second highest cause of death in females, worldwide[38] and therefore evaluating factors leading to its causation is important in its prevention[178]. Tobacco smoking is a crucial environmental factor responsible for 75% of lung cancer cases, and therefore studying genes involved in tobacco-induced lung cancer, such as those involved in carcinogen metabolism and repairing the damage caused by, carcinogens in tobacco smoke have been under extensive evaluation[14].

With the availability of commercial SNP arrays and a dense human genome reference map[107], conducting a genome wide association study (GWAS) to evaluate complex polygenic diseases, including cancers, where several genes with modest effect sizes may have a contributing role, is feasible[192].

5p15.33, 6p21 and 15q24-25.1 chromosomal regions in smokers and 6q23-25 and 13q31.3 chromosomal regions in non-smokers were identified in a GWAS of lung cancer[14, 47, 112, 113, 193], however, the contribution of these loci towards lung cancer susceptibility is moderate. Furthermore, lung cancer is not only affected by genetic changes but also geographical differences[14]. Therefore, the aim of the present study is to identify SNPs associated with lung cancer susceptibility in Liverpool using cases from the Liverpool Lung Project (LLP) and the 1958 Birth Cohort[194] as controls for whom, a genome-wide single nucleotide polymorphism (SNP) dataset is available. Furthermore, information of significant SNPs from this study will be integrated with functional annotations for associated genes.

3.2 Introduction

Tobacco smoking is the major environmental factor for lung cancer causation and therefore genes influenced by cigarette smoke induced genetic changes have been extensively studied by candidate gene association studies[14]. The only direct evidence of familial aggregation of lung cancer is related to the rare Mendelian cancer syndromes such as retinoblastoma gene mutation carriers[195], and xeroderma pigmentosum[196], Bloom's[43] and Werner's[44] syndrome patients and in constitutional carriers of *TP53*[197].

There is a two-fold increase in risk of lung cancer for individuals with a family history of lung cancer[45]. This risk of lung cancer for an individual is associated with the relatives' early age of onset and number of family members affected[45]. However, these studies may be confounded by unadjusted environmental factors such as smoking[45].

Studies conducted on never smokers to eradicate the complications due to common familial smoking habits indicate that genetic or environmental factors may affect the familial aggregation of lung cancer[45].

Studies have also been conducted on monozygotic and dizygotic twins to dissect the genetic and environmental factors influencing familial aggregation of lung cancer[45]. Twin studies conducted on a female population, where the incidence of lung cancer is low, are supportive of a genetic predisposition to lung cancer with lung cancer incidence being more in monozygotic than dizygotic twins[45]. However, other twin studies suggest that environmental factors such as smoking, may be confounded by genetic factors (i.e cancer in twins may be due to genetic factor and not influenced by smoking)[45].

Familial linkage analysis identified a gene *RGS17*[47] following an initial identification of 6q25-23 chromosomal region in a study by the Genetic Epidemiology of Lung Cancer Consortium (GELCC) on 52 families with at least three or more, lung or larynx cancer cases[46].

Dependence on nicotine, a component of cigarette smoke, and susceptibility to lung cancer differ from individual to individual and this forms the basis of lung cancer studies[14]. The interindividual differences in nicotine dependence is quantitatively associated with tobacco smoke carcinogen intake and the interindividual discrepancies in the metabolism of tobacco smoke carcinogens as well as DNA repair activity (which counteracts the mutagenic changes) play a vital role in smoking related cancer[14].

Lung cancer is divided into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), which is further divided into adenocarcinoma (ADC), squamous cell carcinoma (SQC) and large cell carcinoma (LCC)[14]. LCC is more heterogeneous than the other cancer types, with limited information[14]. Basal cells in bronchi, neuroendocrine cells in bronchi and Clara cells in bronchioles and/or type II pneumocytes in the alveoli are the precursor cells for SQC, SCLC and ADC, respectively[14]. SQC and SCLC are more strongly associated with smoking compared to ADC[14]. A unique property of lung cancer cells caused by tobacco smoking is the excess G:C to T:A transversions and therefore the main focus of genetic susceptibility study has been the metabolic enzyme and DNA repair activity that causes or prevents these transversions[14]. For instance, benzopyrene-diol-epoxide, a metabolite of tobacco smoke component, benzo[a]pyrene[198], causes DNA adducts and induces the above transversions in the *TP53* gene[199]. An Arg72Pro SNP also located in the *TP53* DNA repair gene improves the survival of DNA damaged cells for the 72Pro allele than the 72Arg allele[92].

Another example is the Ile462Val SNP located in the *CYP1A1* gene[14, 92, 200]. *CYP1A1* encodes a protein that bioactivates polyaromatic hydrocarbons (PAHs) and this SNP causes a difference in enzyme activity, with individuals having the 462Val allele showing a higher risk of lung cancer than those with the 462Ile allele[92, 200]. With the feasibility of GWAS, the research focus of lung cancer susceptibility studies has shifted from a candidate gene to a whole genome approach[201, 202]. Several GWAS have been conducted in lung cancer research[113, 193, 203], but due to disease complexity, it is anticipated that many markers are yet to be uncovered.

3.2.1 Genetic Variation in Lung Cancer

Lung cancer results from a complex interplay of genetic changes that occur in a sequence involving multiple biological pathways[87]. Some of the important developments in lung cancer are the loss of heterozygosity, activation of proto-oncogenes, inactivation of tumour suppressor genes (TSGs) and epigenetic modification[87].

DNA amplification and loss of heterozygosity are indicators of the modified function of oncogenes and TSGs, respectively[82]. Oncogenes identified in lung cancer research include *KRAS*, *MYC*, *Cyclin D1* and *EGFR*[82]. Copy number amplifications on chromosomal region 8q activate the proto-oncogene *MYC*[87] while amplification of 14q13.3 lead to the identification of *NKX2-1* proto-oncogene[82].

Two mutations are required for the inactivation of some TSGs, this process is called the "two hit hypothesis"[83, 204]. Loss of heterozygosity, one of the two hits for TSG inactivation, has been reported for chromosomal region 17p, 13q14 and multiple loci on 3p,

corresponding to the genes *TP53*, *RB* and *FHIT*, respectively[87]. The *TP53* gene located on chromosome 17p13 is an important protein for maintaining the integrity of the human genome and is frequently mutated in human cancers[87]. This mutation may be related to tobacco smoking which usually causes G-T transversions by the tobacco smoke carcinogens[87].

A major lung cancer-associated epigenetic modification is the hypermethylation of CpG islands found at the 5' promoter region of many genes[87]. Genes silenced in lung cancer by methylation include *DAP kinase*, *GSTP1*, *MGMT* and *CDH13*[87]. Aberrant promoter hypermethylation can be detected in lung cancer patients, in lung tissue devoid of cancer and in high risk individuals and are a promising candidate for use in early detection biomarker research[87].

Changes involved in lung cancer include 9p allele loss, 8p allele loss, 17p allele loss and *TP53* mutations in squamous cell carcinoma and 5q, 9p, 11q and 13q deletions frequent in adenocarcinomas while frequent deletions in 17p, 18q and 22q are observed during lung cancer progression[87]. These genetic changes are seen in the 'field cancerisation' effect where repeated exposure to a carcinogen (eg, in tobacco smoke) can cause neoplasia in the aerodigestive region[87].

3.2.2 Single Nucleotide Polymorphisms

The true extent of interindividual and interpopulation genomic variability has only been revealed since the publication of the initial draft human genome in 2001[202, 205]. Subsequent

to these publications, projects such as HapMap[107] and the 1000 genomes[206] have documented genomic variability both within and between diverse populations.

The structural and sequence variations observed range from large chromosomal regions, such as segmental duplications, inversions and translocations, to smaller regions, such as microsatellites and minisatellites, and finally to the most extensively studied source of variability, single nucleotide polymorphisms (SNPs) and indels[207]. SNPs, which we define here to include both coding and non-coding single base variations, and indels, defined as a single or short string of nucleotides that may be inserted or deleted at a given position[207]. Consequently, a minority of the SNPs and indels in the genome may have coding effects[207]. It is estimated that the genomes from any given pair of individuals from a population may differ from each other by < 0.5%[207].

The SNP is the most common variation in the germline genome[207, 208]. An estimated 11 million SNPs exists in the genome for a given human[208]. Of these, around 7 million occur at a minor allele frequency (MAF) greater than 5% while the rest occur between 1%-5%[208], where for a SNP at a given locus, we define the MAF to be the frequency of the second most common of the biallele in the population of interest[207]. The allelic distribution, for a given site, may differ from population to population due to various evolutionary forces such as the effects of natural selection[209]. A selective sweep may result in extreme frequencies of SNPs located near the selected sites while negative selection would result in a low frequency of the SNP[209]. Common SNPs taken to be those with a MAF > 5%, typically occur in all populations but their distribution may vary across populations with less than 10% of all SNPs being specific to a particular population[207].

3.2.3 Types of Associations

Association studies aim to identify polymorphisms associated with a continuous or discrete trait or disease[210]. Alleles identified through GWAS, may be associated with a disease-causative locus for all the individuals comprising the study group while in linkage studies different loci may be associated with the same trait in different families[210]. Association studies are feasible because humans share common ancestry and although these approaches provide a powerful method to detect smaller effects, association studies do require more markers and a larger study population than are required in linkage analysis[210].

Associations where the polymorphism under study can be considered a disease-causative variant are referred to as "direct associations"[210, 211]. Such an allele may be expected to be a non-synonymous variant in an exon of some gene, or may affect the expression, regulation, splicing and/or function of the associated gene[210]. For example, many putative disease-causative SNPs have been localized to non-coding regions[210]. For instance, SNP rs2522833 in the intron of gene *PLCO* was associated with major depressive disorder suggesting *PLCO* as a causal factor[212].Ten to fifteen thousand non-synonymous SNPs with a MAF of >1% in Europeans, that can be screened in GWAS, have been identified[210].

Conversely, an indirect association can occur when a SNP is strongly associated with a disease, but the disease-causing effect is mediated by a neighbouring causal-SNP[210, 211]. In such a case the indirectly associated SNP is referred to as the "tag SNP" [210, 211]. Identifying the causal variant in indirect association studies requires identifying other SNPs in linkage disequilibrium (LD) with the tag SNP, for example, by densely genotyping the surrounding area of the tag SNP[210, 211].

Identifying SNPs in indirect association would require genotyping several SNPs in the vicinity of the tag SNP, thus making it a less powerful technique than direct association since the exact SNPs to be genotyped is not known[210]. Furthermore, there exists the possibility that indirect association studies may not identify a genuine causal variant, even when one exists[210]. Therefore, such studies should be supplemented by candidate gene studies allowing for finer genotyping of the region, leading to the identification of the true causal variant[210].

3.2.4 Association versus Linkage Studies

Detection of genetic regions associated with human diseases can be broadly classified into linkage and association analysis[192].

SNPs can be directly or indirectly (by being in LD with the causal SNP and co-segregating) associated with disease variant (section 3.2.3). This same principle of LD is used in linkage studies where familial aggregation of markers can be detected over a large genetic distance if the number of familial generations (and thus the number of possible recombinational events) is limited and consequently, increase in familial generations can destroy the intermarker association (via increased number of recombinations over time) leading to short distance LD[213].

Family based linkage studies have been successful in identifying genes associated with Mendelian disorders[192]. In linkage analysis, a pedigree of individuals with multiple members affected with a disease is used to identify chromosomal regions that are common to diseased individuals[192]. Such individuals will share a high proportion of the markers from

along specific regions of the genome[192]. In lung cancer, linkage analysis conducted on 52 high risk pedigrees led to the identification of the 6q23-25 lung cancer susceptibility locus[46]. Fine mapping of this region identified the gene *RGS17,* which is overexpressed in lung tumours and induces cell proliferation in lung tumours[47]. The gene was identified using 24 6q linked cases and 72 unrelated controls in an association analysis, genotyped using Affymetrix 500K chipset, while the validation dataset was made up of 226 familial cases and 313 controls from the GELCC and 154 familial cases and 325 controls from the Mayo clinic[47].

Unfortunately, for complex traits, linkage type analysis is of limited use as the polymorphisms that bring about the disease may only be slightly higher in frequency when compared with unaffected controls[213]. The extensive familial nature of complex diseases, indicate a strong genetic component[213]. However, this heritability may be the result from many genes with small effects, and genetic heterogeneity, where the same phenotype arises as a result of the combinations of different genetic variations[213].

The linkage method is suited for high penetrance familial aggregation of genetic variants but for complex diseases that are caused by multiple genes with small effects, this method is unsuitable[192]. For such complex traits and diseases, a case-control study design is best suited[192].

Association studies aim to detect a relationship between a genotypic polymorphism and a phenotype[210]. Ideally, such a phenotype would be quantitative, such as a trait measurement, or qualitative, such as a disease status[210]. Such studies are based on comparing allelic distributions between cases and controls to identify markers that are significantly more common in cases than would be expected based on their frequency in control subjects[214].

Unlike the linkage approach, which is based on the assumption of a familial aggregation of a disease-causing variant, an association study is conducted on unrelated individuals[207, 210]. Hence, recruitment of large numbers of individuals for analysis is more feasible in the GWAS setting[207].

Association studies had been previously carried out on selected number of candidate genes that were assumed to be involved in the biological processes of lung cancer[192]. However, with the commercial availability of high throughput genotyping technology and the reference map of the human genome[107], genome wide analysis is now possible[192].

In studies where the two approaches have been combined, linkage typically precedes association studies[46, 47]. In the latter, finer details about a trait-associated locus can be revealed than is possible in linkage studies, which have relatively low resolution[210]. Candidate gene association studies have identified disease-associated regions that span over several megabases of DNA but association studies conducted on these regions can identify markers associated with disease, thus identifying the marker of interest[192].

3.2.5 Genome Wide Association Studies in Lung Cancer

Three prominent GWAS have been carried out in lung cancer to identify SNPs associated with lung cancer in Caucasians[112, 113, 193]. Hung *et al*. (2008) conducted a GWA study on 1989 lung cancer cases and 2625 controls in the discovery phase and 2513 lung cancer cases and 4752 controls in the replication phase[113]. The Illumina Sentrix HumanHap 300 bead chip array, containing 317,139 SNPs, was used to genotype the above individuals[113]. SNPs were

analysed in a multivariate unconditional logistic regression model standardised using age, sex and country to identify significant SNPs[113].

The study identified chromosomal region 15q25 as conferring an association with the risk of lung cancer[113]. This region contains the *CHRNA5*, *CHRNA3* and *CHRNB4* genes[113]. Two SNPs were strongly associated with the disease rs1051730 ($p = 3x10^{-9}$) and rs8034191 ($p = 9 x 10^{-10}$)[113]. For rs8034191, the odds ratio (OR) was 1.27 (95% CI: 1.11-1.44) for carrying 1 copy and for carrying 2 copies it was 1.80 (95% CI: 1.49-2.18); and for the allelic model it was 1.32 (95% CI: 1.21-1.45) in the central European population of 1922 cases and 2520 controls[113]. While in the combined analysis of 4435 cases and 7272 controls an OR of 1.21 (95% CI: 1.11-1.31) for one copy; 1.77 (95% CI: 1.58-2.00) for two copies and 1.30 (95% CI: 1.23-1.37) for the allelic model standardised using age, sex and country was observed[113].

Thorgeirsson *et al*. (2008) carried out a study to identify variants associated with smoking dependence (using smoking related measurements)[112]. SNPs identified using the 10,995 Icelandic smokers in a genome wide study was used to study their risk of both lung cancer and peripheral arterial disease[112]. Lung cancer analysis was carried out using 1024 lung cancer cases and 32,244 controls, whereas peripheral arterial disease analysis was performed using 2738 cases and 29,964 controls[112].

Individuals were genotyped using the Illumina Human Hap300 and Human Hap300-duo+ Bead arrays[112]. The T allele of the SNP rs1051730, within the *CHRNA3* gene located on chromosome 15q24, was associated with smoking quantity ($\beta = 0.095$; 95% CI: 0.075–0.115; $p=6x10^{-20}$), nicotine dependence (OR= 1.40; 95% CI: 1.29–1.52, $p=7x10^{-15}$), lung cancer risk (OR =1.31; 95% CI: 1.19–1.44, $p=1.5x10^{-8}$) and peripheral arterial disease (OR = 1.19; 95% CI: 1.12–1.27, $p= 1.4x10^{-7}$)[112].

Finally, McKay *et al*. (2008), conducted a GWAS on 2971 lung cancer cases and 3746 controls and discovered two significant SNPs, rs402710 and rs2736100, with OR of 1.22 (95% CI: 1.13-1.32) and 1.18 (95% CI: 1.10-1.26)[193]. These SNPs were replicated in 2899 lung cancer cases and 5573 controls and produced an OR of 1.15 (95% CI: 1.07-1.24) and 1.09 (95% CI: 1.02-1.17) for rs402710 and rs2736100, respectively[193]. For the combined analysis of 5870 cases and 9319 controls, the allelic model and the genotypic model for carrying one and two copies of the minor allele of rs402710 produced an OR of 1.18 (95% CI: 1.12-1.24), 1.18 (95% CI: 1.05-1.33) and 1.40 (95% CI: 1.24-1.57), respectively[193]. For SNP rs2736100, the OR for allelic and genotypic model for carrying one and two copies of the minor allele was 1.14 (95% CI: 1.08-1.20), 1.07 (95% CI: 0.98-1.17) and 1.29 (95% CI: 1.17-1.43), respectively[193]. The statistical model was adjusted for age, sex and country and the individuals were genotyped using the Illumina Sentrix HumanHap300 BeadChip[193]. The susceptibility region on 5p15.33 carries two potential candidate genes *TERT* and *CLPTM1L* [193].

Genome wide association studies revealed three chromosomal regions associated with lung cancer namely, 15q24-25.1, 5p15.33 and 6p21 in European, Americans and Asians[14]. The 15q24-25.1 region harbours the gene, *CHRNA3* and *CHRNA5*, whose function has been associated with nicotine dependence[14]. The 5p15.33 region contains the *TERT* gene involved in telomerase replication, maintenance and cell proliferation[14] and the *CLPTM1L* gene, which is associated with apoptosis[193, 215]. The 6p21 MHC[216] region contains the *BAT3* gene whose product binds to p300 and MSH5 which are, respectively, involved in the DNA damage response and mismatch repair[14].

3.2.6 Factors Affecting Genome Wide Association Studies

3.2.6.1 Power

The power of a statistical test is defined as follows,

$$\text{Power} = 1 - \beta$$

where β, the type II error rate, is the probability of accepting a false null hypothesis[217]. The power of detecting a true significant association depends on the effect size and the sample size of the genome wide association study[218]. Making sure that the studies are well powered is a crucial point to reduce the number of false positive associations, also called the ”False Positive Report Probability” (FPRP), determined by the magnitude of the p value and the proportion of the tested hypotheses that are true[219].

3.2.6.2 Hardy Weinberg Equilibrium

In a large randomly mating population devoid of selection, mutation, or migration, the expected frequency of diploid genotypes at a locus can be predicted as a simple function of the allele frequency[220, 221]. This phenomenon was independently described by Hardy & Weinberg in 1908, and hence called the “Hardy Weinberg Equilibrium” [220, 221]. For instance, given a biallelic locus, with genotypes AA, Aa and aa, the expected frequency of these

genotypes within the population are respectively, $(1-q)^2$, $2q(1-q)$, and q^2, with respect to the minor allele frequency q[220, 221].

Deviations from the Hardy Weinberg equilibrium amongst a GWAS study population may indicate population stratification, admixture, cryptic relatedness, inbreeding, selection, genotyping error and, where observed, this may alter the association of a genetic marker with the disease[221].

Inbreeding, mating between closely-related individuals leads to a decrease in heterozygosity across the genome, increasing the number of homozygotes[221]. Similarly, small population sizes increase homozygosity through genetic drift until the population is fixed for homozygotes[221].

A close kinship in a sample of unrelated individuals may lead to increased homozygosity, this is known as "Cryptic Relatedness" and the presence of such cryptic relatedness may increase false positive results in genetic association studies[221]. Similarly, an admixed population, made up of many sub populations, each of which may be in Hardy Weinberg equilibrium, may display Hardy Weinberg disequilibrium[221].

Other causes for deviation from the Hardy Weinberg equilibrium at the genotyping level may be, mutations in PCR primer sites, contaminated DNA leading to a wrong allelic call, low quality or quantity of DNA leading to uncalled alleles or to genotyping calling errors[221]. Such errors can lead to inflated type I and II error rates in association studies[221].

Detecting Hardy Weinberg proportions:

Hardy Weinberg proportions are tested on the null hypothesis that there is no significant difference between the observed and the expected genotypic counts, the alternative hypothesis being that there is a significant difference between the observed and the expected genotypic counts[221]. The common tests used for the Hardy Weinberg proportions are the Pearson's X^2goodness- of-fit test and the Monte Carlo Markov Chain (MCMC) exact test[221].

Pearson's X^2 goodness-of-fit:

Consider a sample with N individuals and the observed count for the genotypes AA, AB, BB at a single locus being n_{AA}, n_{AB} and n_{BB}, respectively than the chi square test statistic is given by

$$\chi^2 = \Sigma \frac{(Observed\ number\ of\ genotypes - Expected\ number\ of\ genotypes)^2}{Expected\ number\ of\ genotypes} \quad 221$$

$$= \frac{(n_{AA} - \mathrm{Np}_{\mathrm{AA}}^2)^2}{Np_{AA}^2} + \frac{[n_{AB} - 2Np_A(1-p_A)]^2}{2Np_A(1-p_A)} + \frac{\left[n_{AA} - \mathrm{N}(1-\mathrm{p}_{\mathrm{AA}})^2\right]^2}{\mathrm{N}(1-\mathrm{p}_{\mathrm{AA}})^2} \quad 221$$

Where p_A is the allele frequency given by, $p_A = \frac{2n_{AA} + n_{AB}}{2N}$ 221

Hardy Weinberg Exact test:

Consider a sample with N diploid individuals and the observed count for the genotypes AA, AB and BB at a single locus being n_{AA}, n_{AB} and n_{BB} [221]. The expected number of individuals with each genotype would be Np_A^2 , $2Np_A(1-p_A)$ and $\mathrm{N}(1-\mathrm{p_A})^2$ respectively, where p_A is the allele frequency of A[221]. Then the conditional probability can be expressed as the probability of heterozygous genotype n_{AB} determined by the observed count of A allele given as

$$\Pr(n_{AB}|N,n_A) = \frac{n!n_A!n_B!2^{n_{AB}}}{[(n_A-n_{AB})/2]!n_{AB}![N-(n_A+n_{AB})/2]!(2N)!} \quad [221]$$

Where $n_A = 2n_{AA} + n_{AB}$ is the observed count for allele A, N the sample size and n_B = 2N - n_A [221].

Complete enumeration is not practical when it comes to dealing with a large sample size of multiple alleles therefore a permutation or resampling based method has been developed[221]. This includes the Markov chain Monte Carlo (MCMC) method that was proposed by Guo and Thompson[222]. This method functions by generating a large number of independent genotypes depending on the observed allele count and sample size[221].

3.2.6.3 Linkage Disequilibrium

The non-random association of alleles is referred to as linkage disequilibrium (LD) and can be detected using two measurement, r^2 and D' [214, 223].

Considering that not all polymorphisms are genotyped, LD allows the detection of non-genotyped causal variants[214]. Consider two SNPs, A and B at a biallelic locus with frequencies AA/Aa/aa and BB/Bb/bb at different positions on a strand of a chromosome, then primarily, LD, is the measurement of the difference between the observed number of AB pairs and product of the expected frequency of A and B[224].

$$D' = D / D_{max} \quad [224]$$

where $D_{max} = \min(p(A)p(b),p(a)p(B))$ if $D \geq 0$,or, if $D < 0$, then

$$D_{max} = \min(p(A)p(B), p(a)p(b)) \quad [224]$$

$$\text{and } D = P_{AB}P_{ab}-P_{Ab}P_{aB}$$

where $P_{AB,}$ P_{ab}, P_{Ab} and P_{aB} are frequencies of haplotype AB, ab, Ab and aB, while p(A),p(a),p(B) and p(b) are frequencies of alleles A, a, B and b[224].

3.2.6.4 Population Stratification

Population stratification arises when the allele frequency between cases and controls are different due to ancestry rather than the disease in question[223, 225]. Therefore, a case control study should be designed to obtain cases and controls from populations with common ancestry[223, 225].

The effect of population stratification can be eliminated or decreased by identifying and excluding individuals of divergent ancestry, correcting the association statistics for genomic inflation and controlling by using Principal component Analysis (PCA) or Multidimensional Scaling (MDS)[223, 225].

3.2.6.4.1 Genomic Control

Population stratification can be detected and controlled for by calculating the genomic control parameter, λ, computed as the median X^2statistic divided by the constant, 0.456 [225]. It corrects by adjusting the deviation from the null hypothesis of homogeneity[225, 226]. If the genomic parameter is large, the association statistics can be corrected by dividing it with λ[225, 226].

3.2.6.4.2 Multidimensional Scaling

Implementing the multidimensional scaling (MDS) method requires constructing the pairwise identity-by-descent (IBD) matrix, which is calculated using identity-by-state (IBS) [223]. Given the genotypes for a pair of individuals, the pairwise IBS is calculated by summing up the total number of alleles they share in common at individual loci divided by the total number of non-missing, common SNPs under study[223].

The IBD matrix is calculated using a Hidden Markov Model (HMM) that utilizes the observed IBS sharing and the genome wide level of relatedness between pairs[227]. For any two individuals from a random mating homogenous population, a method of moments is used to infer the probability of sharing 0, 1, or 2 alleles by IBD[227].

IBD is used to detect sample and genotyping error like duplicates, contamination and mislabeling, or relatives. Samples contaminated with other DNA may result in false heterozygote calls which would inflate the IBD estimates[227].

3.2.7 Quality Control

Quality control in a GWAS is essential to eliminate the risk of false positive or negative results[223]. Several quality controls have to be carried out to identify and remove participants and/or individual markers that may result in a spurious association[223].

Discordant gender: Sex checks should be carried out to identify discrepancies between the reported sex of an individual versus those derived from the genotypes[223]. The X chromosome is used to determine the gender of an individual and this can be checked against the reported sex in the data[223]. Gender discrepancies can be ascertained by calculating the homozygosity rate across the X chromosome for every individual, where, if male, the homozygosity rate should be 1 and for females, it should be <0.2[223].

Missing genotypes: Poor quality samples or low concentrations of DNA can be reflected in the genotypic failure and heterozygosity rate[223]. Heterozygosity rate is calculated as, the difference between the number of non-missing genotypes and the observed number of homozygous genotypes divided by the number of non-missing genotypes[223]. Individuals with poor genotyping call rate (<95%) and low heterozygosity rate should be removed from a study[223].

Related individuals: Over representations of genotypes due to familial relations or duplicates within the sample may introduce a bias to a study, through over-representation of a specific family, relative to the study population[223]. To identify duplicate or related

individuals an IBS matrix is calculated[223]. To reduce computational complexity, regions of high LD may be excluded for IBS matrix calculation[223].

Recent shared ancestry is inferred from the Identity-by-descent (IBD) score for a pair of individuals[223]. This is determined from the genome wide IBS data. An IBD of 1 corresponds to duplicated individuals or monozygotic twins[223]. An IBD of 0.5, 0.25 and 0.125 corresponds to first, second and third degree relatives, respectively[223]. There is of course some degree of discrepancy around these theoretical values but essentially removing one individual from any pair that produces an IBD > 0.1875, a value between second and third degree relative, is considered an acceptable means to reduce this bias[223].

Ethnic outliers: Confounding due to population stratification may arise due to different origins or ancestries[223]. This can be controlled if the cases and controls are sampled from a common ancestral population or the analysis is adjusted for population substructure[223]. Population substructure arises as a result of different allele frequencies in the case and control population[223].

As discussed in Section 3.4.4, allele frequencies can differ considerably between subpopulations (for example, between geographic regions). Correspondingly, population stratification either within or between the arms of a case-control GWA study may confound the identification of disease-associated loci[223]. The simplest means to control for population stratification is to ensure that individuals are randomly sampled from a homogeneous population, although, even here, unexpected population substructure may be revealed at a molecular level[223]. Alternatively, outliers (individuals with a vastly different allele complement across multiple SNPs) may be identified and removed from a dataset[223].

SNP-level quality controls: Apart from individual-based quality control, SNP based tests are crucial to avoid any spurious associations[223]. These tests include testing for Hardy Weinberg equilibrium in control population, testing for missingness rate between cases and controls and excluding SNPs that have a minor allele frequency or a genotyping call rate below a specified threshold[223]. The latter thresholds are typically between 0.01 to 0.05% and <95% to <97%, respectively[223].

3.2.8 Genetic Models in Association Studies

3.2.8.1 Additive Model

Usually while detecting significant SNPs associated with a disease condition the mode of inheritance is usually not known [226]. In such cases the additive model is preferred [226]. This can be tested using the X^2 and the Cochran-Armitage Trend test [226]. Consider a case control set up of N biallelic loci and the genotype and allelic distribution for a locus is of the following form [226]

Table 3.1: Genotype and allele distribution for additive model

	3x2 Genotype Distribution				**2x2 Allele Distribution**		
	A_1 alleles						
	0	1	2	Total	A1	A2	Total
Case	a_0	a_1	a_2	A	a_1+2a_2	a_1+2a_0	2A
Control	b_0	b_1	b_2	B	b_1+2b_2	b_1+2b_0	2B
Total	c_0	c_1	c_2	C	$c_1+ 2c_2$	$c_1 +2c_0$	2C

The X^2 statistic with 1 degree of freedom is given by

$$\chi_A^2 = \frac{2C[2C(a_1+2a_2)-2A(c_1+2c_2)]^2}{(2A)2(C-A)[2C(c_1+2c_2)-(c_1+2c_2)^2]} \quad 226$$

and the Cochran-Armitage Trend statistic under the null hypothesis of no association, with one degree of freedom is given by

$$T = \frac{C[C(a_1+2a_2)-A(c_1+2c_2)]^2}{A(C-A)[C(c_1+4c_2)-(c_1+4c_2)^2]} \quad 226$$

The odds ratio for the additive models is calculated by the formula

$$\text{Odds Ratio} = \frac{(a_1+2a_2)/(b+2ab_2)}{(a_1+2a_0)/(b_1+2b_0)} \quad 221, 227$$

3.2.8.2 Dominant Model

The genotype and allele distribution for the dominant model is as follows[226, 227].

Table 3.2: Genotype and allele distribution for dominant model

	3x2 Genotype Distribution				**2x2 distribution for dominant model**		
	A_1 alleles						
	0	1	2	Total	A1	A2	Total
Case	a_0	a_1	a_2	A	a_1+a_2	a_0	A
Control	b_0	b_1	b_2	B	b_1+b_2	b_0	B
Total	c_0	c_1	c_2	C	c_1+c_2	c_0	C

The X^2 statistic and odds ratio under the dominant model is calculated using the following expressions:

$$\chi_A^2 = \frac{C[C(a_1+a_2)-A(c_1+c_2)]^2}{(A)(C-A)[C(c_1+c_2)-(c_1+c_2)^2]} \quad 226, 227$$

$$\text{Odds Ratio} = \frac{(a_1+a_2)/(\text{b}+b_2)}{a_0/b_0}$$ 221, 227

3.2.8.3 Recessive Model

The genotype and allele distribution for the recessive model is as follows[226, 227].

Table 3.3: Genotype and allele distribution for recessive model

	3x2 Genotype Distribution				**2x2 distribution for recessive model**		
	A_1 alleles						
	0	1	2	Total	A1	A2	Total
Case	a_0	a_1	a_2	A	a_2	a_1+a_0	A
Control	b_0	b_1	b_2	B	b_2	b_1+b_0	B
Total	c_0	c_1	c_2	C	c_2	c_1+c_0	C

The X^2 statistic and odds ratio for the recessive model is calculated using the following expressions:

$$\chi_A^2 = \frac{C[C(a_2)-A(c_2)]^2}{(A)(C-A)[C(c_2)-(c_2)^2]}$$ 226, 227

$$\text{Odds Ratio} = \frac{a_2/b_2}{(a_1+a_0)/(\text{b}+b_0)}$$ 221, 227

3.2.8.4 Genotypic Model

The genotype and allele distribution for the genotypic model is as follows[226, 227].

Table 3.4: Genotype and allele distribution for genotypic model

	3x2 Genotype Distribution			
	A_1 alleles			
	0	1	2	Total
Case	a_0	a_1	a_2	A
Control	b_0	b_1	b_2	B
Total	c_0	c_1	c_2	C

The X^2 statistic and odds ratio for the genotypic model is calculated using the following expressions.

$$\chi^2_A = \frac{(a_1 - (Aa_1/C))^2}{Aa_1/C} + \frac{(a_2 - (Aa_2/C))^2}{Aa_2/C} \quad [226, 227]$$

$$\text{Odds Ratio for carrying one copy of } A_1 \text{ allele} = \frac{a_1/b_1}{a_0/b_0} \quad [221, 227]$$

$$\text{Odds Ratio for carrying two copies of } A_1 \text{ allele} = \frac{a_2/b_2}{a_0/b_0} \quad [221, 227]$$

3.2.9 Multiple Testing

The testing of multiple SNPs for association studies requires the reduction of false positive associations through correction techniques like the Bonferroni correction[214]. Testing for n SNPs results in a Bonferroni correction level of α/n where α is the type I error rate which, if set at 0.05, the Bonferroni correction level for 1 million SNPs is (5×10^{-8}) [214].

3.3 Materials and Methods

3.3.1 DNA Extraction, Genotyping and Quality Control

Lung cancer cases from Liverpool were consented to participate in the study according to the LLP protocol[228]. They were histologically or cytologically confirmed. DNA from blood was extracted using Qiagen kits (Qiagen, Valencia, CA) using established protocols[113] and genotyped on the Illumina 300K bead chip array (http://www.illumina.com/downloads/HUMAN HAP300Datasheet.pdf).

Genotype data from the Illumina1.2M SNP platform for 3000 controls was downloaded from the Wellcome Trust Case-Control Consortium for the 1958 British birth cohort[194]. No other phenotypic information was available for the control dataset, hence limiting it to the univariate analyses. Due to imperfect overlap between the genotypes assayed on the Illumina 300K array and the Illumina1.2M SNPs arrays, analysis was restricted to those SNPs present on both arrays.

Quality control filters as available within the PLINK program[227] were applied to the dataset and are described presently. Individuals with discordant gender were removed from the dataset and not considered any further. For a pair of individuals with IBD > 0.1875, an individual was selected at random and removed from the dataset[223]. Any outliers in the study defined as individuals that outwith the sample heterozygosity mean ± 3sd and genotyping rate of < 0.95% were removed from further analysis[223]. Individuals that represented population outliers were identified using multidimensional scaling[223]. This was done by using the genotypes of SNPs that are not correlated ($r^2 < 0.2$) and the HapMap3

data for the European (Utah residents with Northern and Western European ancestry from the CEPH collection (CEU)), Asian (Japanese in Tokyo, Japan (JPT) and Han Chinese in Beijing, China (CHB)) and African (Yoruba in Ibadan, Nigeria (YRI)) populations. The case-control genotype data was merged with the HapMap3 data for which the MDS matrix was calculated using the uncorrelated SNPs[223]. Cluster plots were produced for the first two clusters or components and any outliers were removed from further analyses.

Also, SNPs with significant ($p<0.00001$) missing rate between cases and controls, depicting genotyping discrepancies, SNPs with MAF < 1%, genotype call rate of < 97% and a Hardy-Weinberg p-value < 0.001, were removed from the study.

3.3.2 Statistical Analysis

The allelic model compares the frequencies of the minor allele, the dominant model compares the presence or absence of the minor allele, the recessive model models the effect of carrying two copies of the minor allele while the genotypic model treats each, the homozygous and the heterozygous genotype containing the minor allele as a factor (Table 3.5)[226]. The models that were used to identify significant SNPs, were run using PLINK[227]. The odds ratio (OR) and 95% confidence interval (95% CI) for the allelic model were presented by PLINK[227], but for the dominant, recessive and genotypic model, the OR and 95% CI was calculated using STATAv12[179], after extracting the genotypes using PLINK[227]. For the genotypic model, the heterozygous and homozygous carrier of the minor allele were studied with the other homozygous allele as the baseline in a stratified analysis, while for the dominant, the carriers of the minor allele were compared to the other homozygote and

in the recessive model, the homozygote for the minor allele was compared to the other two forms of genotype using logistic regression.

Table 3.5: Different models tested in the genome wide association study

MODEL	AA	AB	BB	Degrees of freedom
Allelic	0	1	2	1
Dominant	0	1	1	1
Recessive	0	0	1	1
Genotypic	0	1	2	2

*B(minor allele)

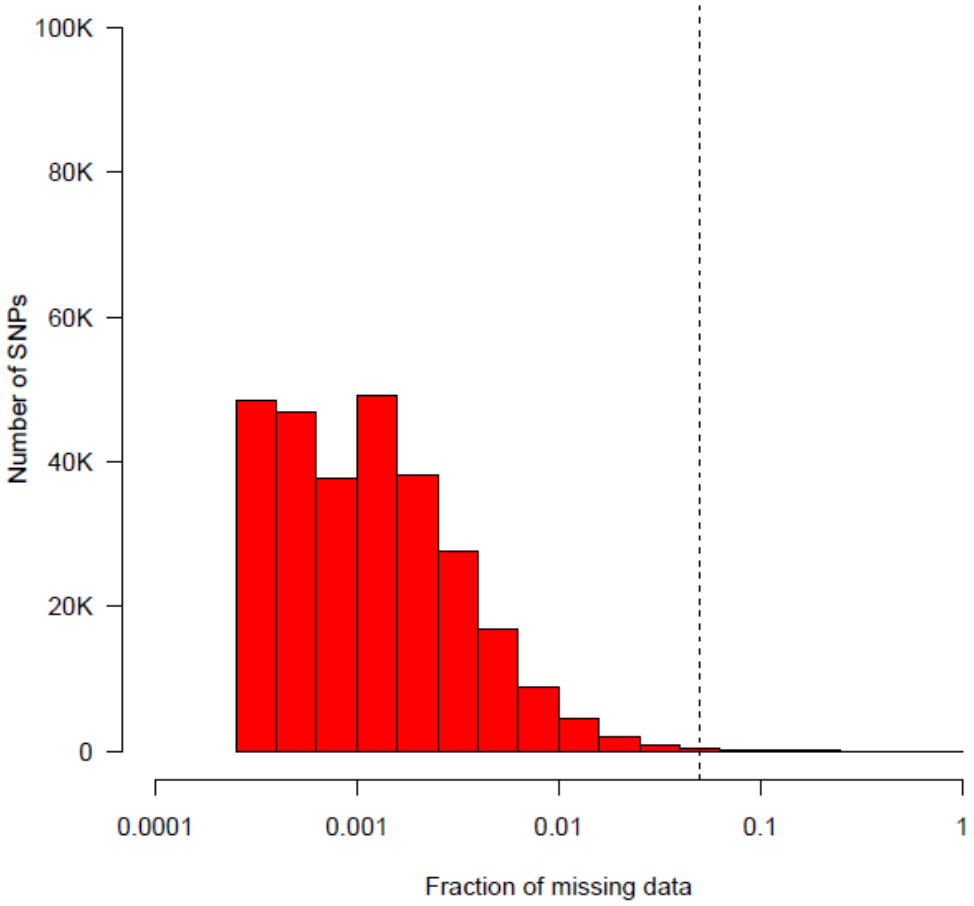
Manhattan plots were drawn for every model by using packages in R statistical software[229] such as "calibrate"[230] and significant SNPs obtained were queried for gene location, cytogenic position and function using package "NCBI2R"[231] in R[229]. The Bonferroni correction level was calculated for every model, by dividing 0.05 by the total number of tests conducted. The Venn diagrams were plotted in R[229] using package "Vennerable"[232].

3.4 Results

The study comprised of 570 LLP cases and 3000 1958 Birth Cohort controls. The case population comprised of 58.26% males and 41.74% females with a mean of 42.62 (standard deviation = 26.85) and 67.19 (standard deviation = 9.09) for smoking pack years and age at diagnosis, respectively. Each SNP was tested individually for the allelic, dominant, recessive and the genotypic model. The allelic model was applied to 277471 SNPs while the dominant, recessive and genotypic SNP was applied to 239757 SNPs. This is because PLINK[227] disregards SNPs that have a frequency of less than 5 in the 2x3 table for the

dominant, recessive and genotypic model but it conducts association analysis for all SNPs when using the allelic model. In the allelic model, 277461 were included, since 10 were dropped for having a value of less than 5, in the 2x2 table. The genomic inflation factor for the allelic association study generated by PLINK[227] was 1.17, indicating that there is no confounding due to population stratification[113, 226]. The genotypic rate for all of the individuals in the study was 99.88%.

Figure 3.1: Distribution of SNPs with missing genotypes. This was carried out in PLINK (Purcell, Neale *et al*., 2007) by steps provided by Anderson *et al*.,2010.



Most of the loci were genotyped in all individuals, with majority of the individuals having minimal missing SNP information (Figure 3.1). Individuals were excluded from further analysis if the displayed a missing call rate of > 3% or a rate of heterozygosity beyond ±3

standard deviations from the mean heterozygosity rate of the case-control population (Figure 3.2).

Figure 3.2: Plot of heterozygosity rate versus missing genotypes.

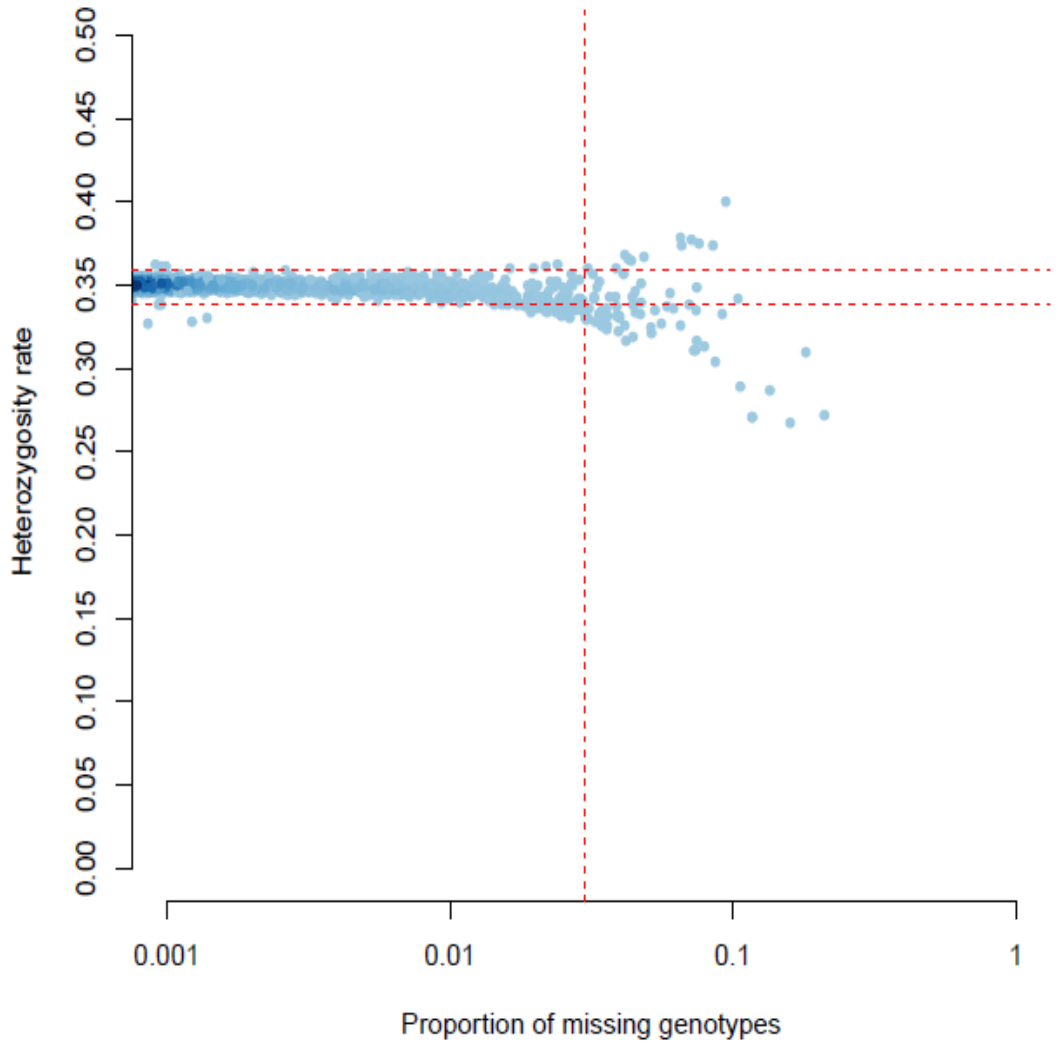


Figure 3.2 is a plot of heterozygosity rate against the proportion of missing genotypes. The vertical red broken line is plotted at a x axis value of 0.03 while the parallel red broken lines drawn horizontal to the x axis are at 3 standard deviations above and below the mean heterozygosity rate. Individuals, depicted as blue dots, contained in the enclosed area formed by the y axis and the red broken lined were included for further analyses.

Figure 3.3: Cluster plot of cases, controls and HapMap3 populations.

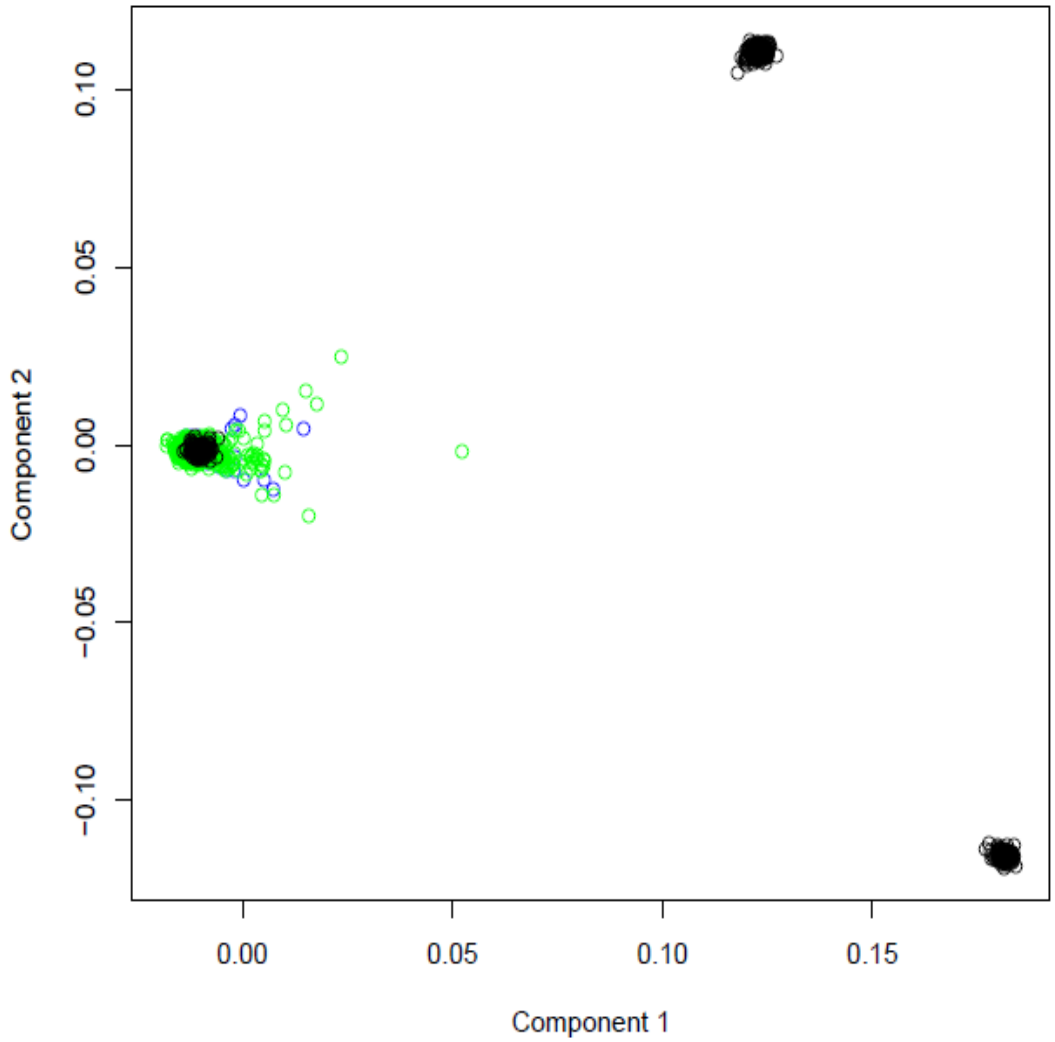
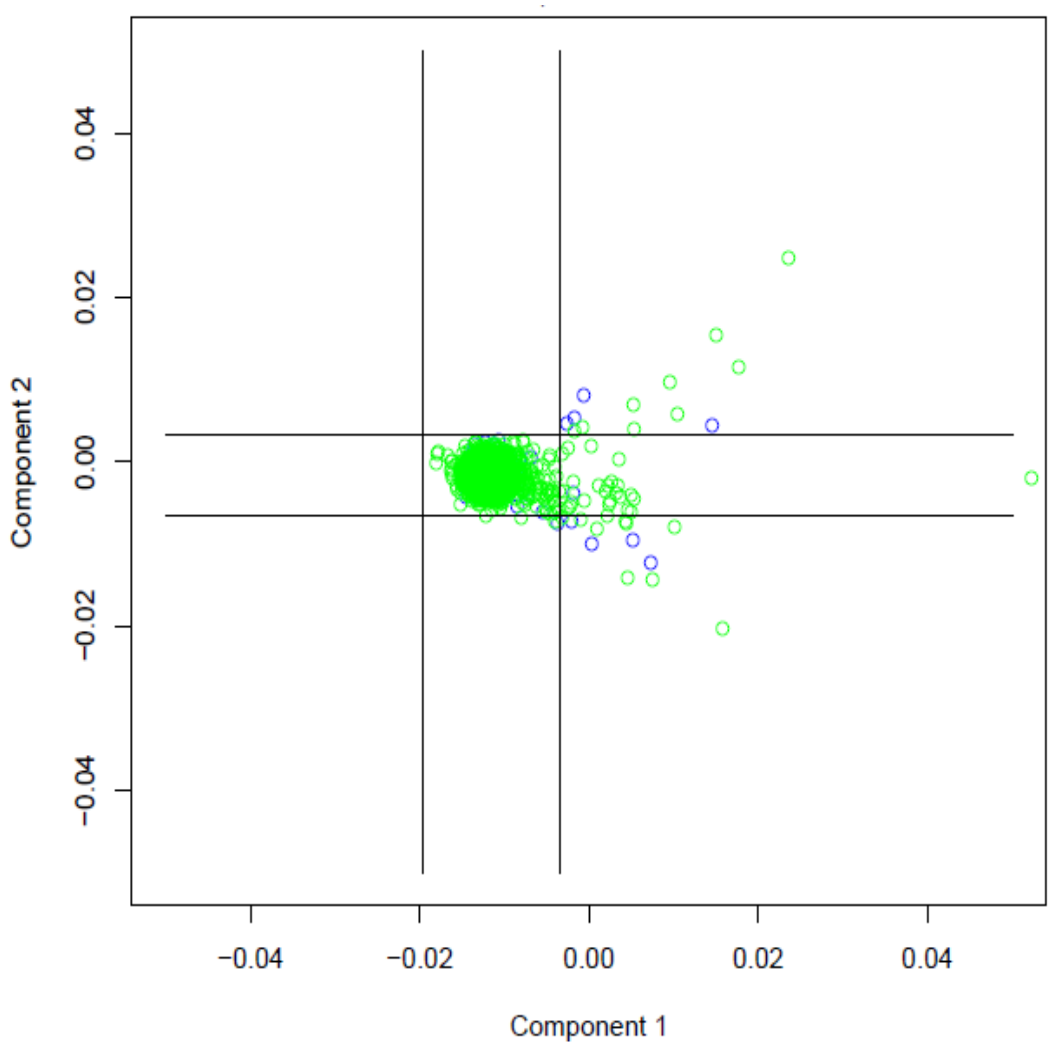


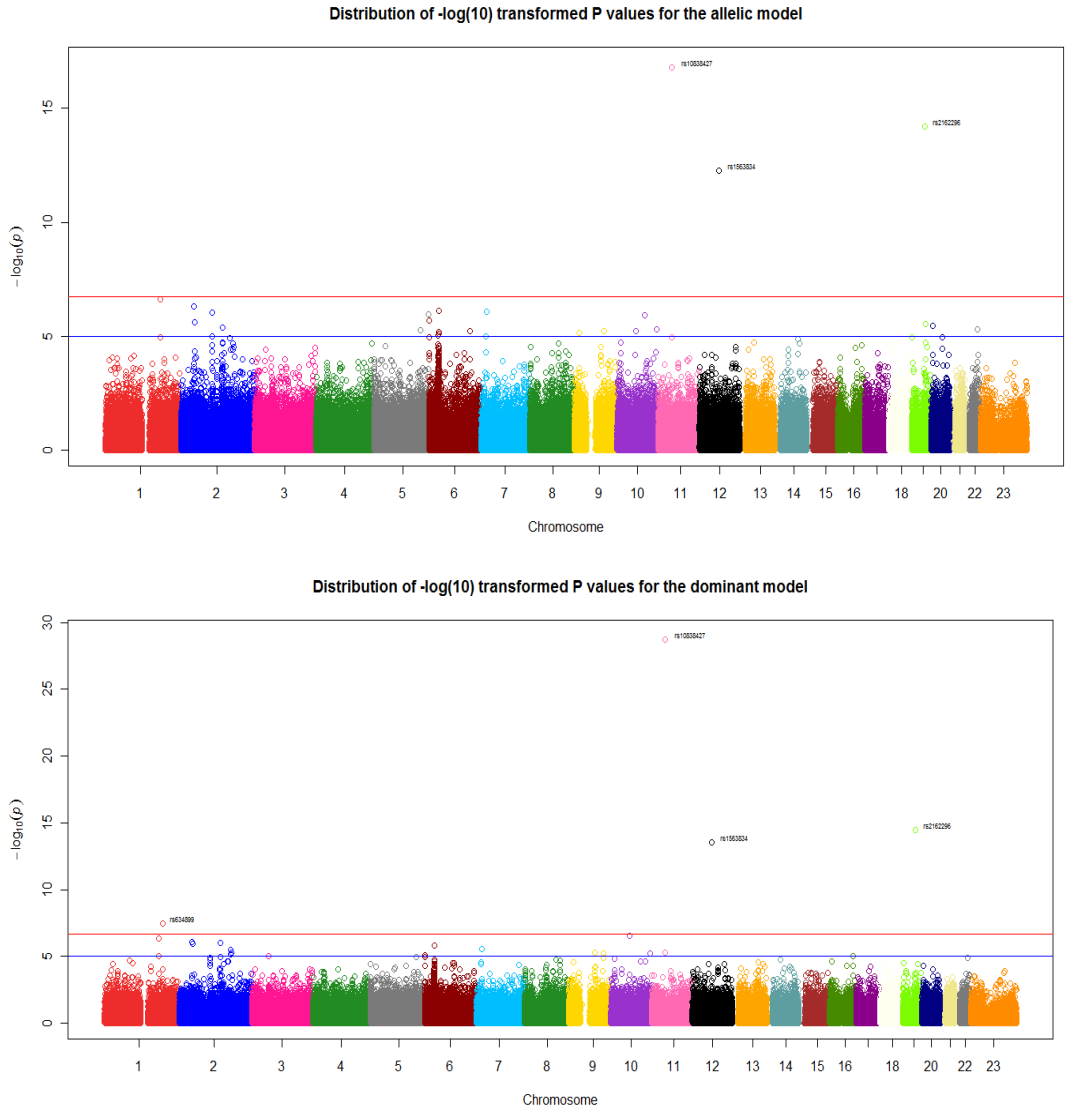
Figure 3.3 is the cluster plots for LLP cases, 1958 Birth Cohort controls and HapMap3 dataset made up of European, Asian and African population, for the first two components. The blue cluster, which is superimposed by the green and black cluster are the LLP cases. The green cluster is the 1958 Birth Cohort used as controls and the black cluster, on the green cluster is the HapMap3 European population. The other two black clusters are Asians and Africans from the HapMap3 population. The common European ancestry of the cases and controls is shown by them, clustering together with the Hapmap3 European population.

Figure 3.4: Cluster of cases and controls



Further enlarging on the European cluster and excluding the HapMap3 population from Figure 3.3, the rectangle formed by the parallel lines drawn from either axis in Figure 3.4 are individuals included for further analyses. These individuals were chosen because they form a tight cluster indicating common ancestry, while the rest are scattered.

Figure 3.5: Manhattan plots of -$\log_{10}$(p) versus base pair position, for the allelic, dominant, recessive and genotypic model.



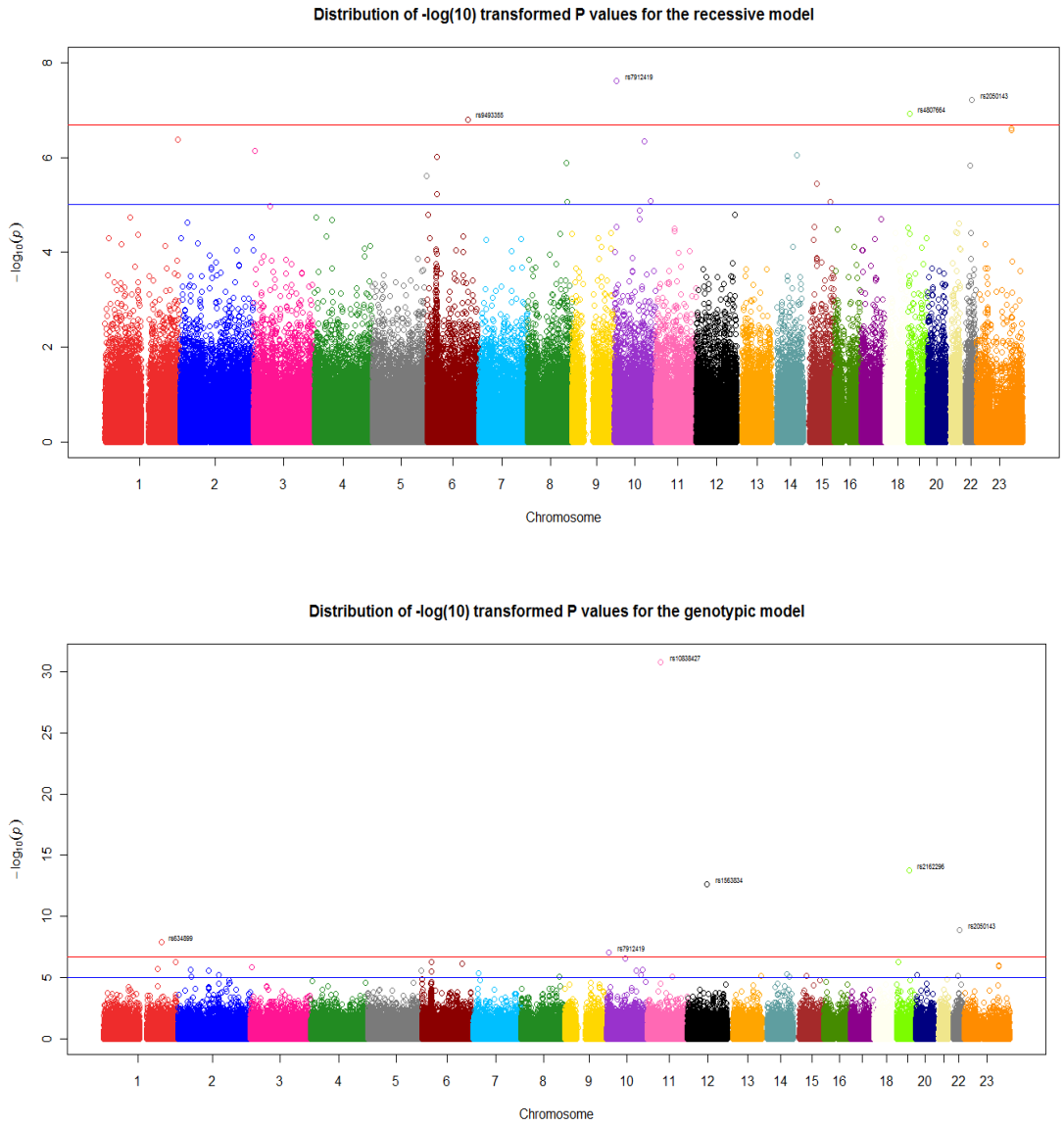


Figure 3.5 shows Manhattan plots for the various models tested using the case-control dataset. It is a plot of the negative of the log to the base of 10 of the p values versus the base pair position for every SNP within a chromosome, for every chromosome. The red horizontal line is the Bonferroni correction level while the blue line is the suggested 10^{-5} level.

Three SNPs reached the Bonferroni significance level of 1.80×10^{-07} in the allelic model while 6, 4 and 4 SNPs reached the Bonferroni significance level of 2.09×10^{-07} in the genotypic, dominant and recessive model, respectively.

Table 3.6: SNPs ($p \leq 10^{-5}$) showing a significant evidence of allelic association.

Chr	SNP	Minor allele	Case MAF	Control MAF	OR (95% CI)	p-value
11	rs10838427	A	0.21	0.35	0.51 (0.43 - 0.6)	1.56E-17
19	rs2162296	A	0.11	0.21	0.45 (0.36 - 0.55)	6.3E-15
12	rs1563834	A	0.06	0.15	0.4 (0.31 - 0.52)	5.48E-13
1	rs16857239	C	0.16	0.11	1.62 (1.35 - 1.95)	2.35E-07
2	rs2888881	A	0.3	0.23	1.45 (1.25 - 1.68)	4.96E-07
6	rs3130564	A	0.32	0.25	1.43 (1.24 - 1.65)	7.44E-07
6	rs3868542	G	0.28	0.36	0.69 (0.6 - 0.8)	7.62E-07
7	rs7787541	A	0.1	0.06	1.76 (1.4 -2.2)	8.54E-07
2	rs4851692	A	0.24	0.17	1.48 (1.27 - 1.74)	9.47E-07
5	rs4410655	A	0.53	0.44	1.39 (1.22 - 1.58)	1.06E-06
10	rs10509535	G	0.11	0.06	1.73 (1.38 - 2.16)	1.16E-06
6	rs9405681	A	0.16	0.23	0.66 (0.55 - 0.78)	2.04E-06
2	rs6735530	G	0.08	0.14	0.58 (0.46 - 0.73)	2.55E-06
19	rs2304214	A	0.4	0.32	1.38 (1.21 - 1.58)	3.02E-06
20	rs2232081	A	0.17	0.12	1.53 (1.28 - 1.84)	3.39E-06
2	rs6739713	G	0.13	0.18	0.63 (0.52 - 0.77)	4.22E-06
10	rs2096285	G	0.26	0.33	0.71 (0.61 - 0.82)	4.91E-06
22	rs4822112	C	0.21	0.15	1.47 (1.25 -1.74)	5.12E-06
5	rs7709656	A	0.12	0.17	0.63 (0.52 - 0.77)	5.37E-06
6	rs17062322	A	0.08	0.13	0.58 (0.46 - 0.74)	5.92E-06
9	rs953715	A	0.09	0.14	0.6 (0.47 - 0.75)	6.04E-06
10	rs10994443	A	0.06	0.11	0.55 (0.42 - 0.71)	6.13E-06
6	rs9366778	A	0.29	0.36	0.72 (0.62 - 0.83)	6.57E-06
9	rs12683609	C	0.38	0.46	0.73 (0.64 - 0.84)	7.18E-06
6	rs3130977	G	0.42	0.35	1.36 (1.19 - 1.55)	7.73E-06
6	rs4713175	A	0.06	0.11	0.55 (0.42 - 0.72)	9.18E-06

The above table lists the results obtained for the association analysis using allelic model. Twenty six significant SNPs were identified at $p \leq 10^{-5}$, of which 3 were significant at Bonferroni correction level. The description of genes that harbour or are in the vicinity of these SNPs are described in Table 3.10.

Table 3.7: SNPs ($p \leq 10^{-5}$) significant in the dominant model.

Chr	SNP	Minor allele	Case MAF	Control MAF	Odds ratio (95% CI)	p-value
11	rs10838427	A	0.21	0.35	0.33 (0.27 - 0.40)	1.86E-29
19	rs2162296	A	0.11	0.21	0.41 (0.33 - 0.51)	3.71E-15
12	rs1563834	A	0.06	0.15	0.35 (0.26 - 0.46)	2.96E-14
1	rs634899	A	0.29	0.35	0.59 (0.49 - 0.71)	3.38E-08
10	rs10994443	A	0.06	0.11	0.47 (0.35 - 0.63)	2.82E-07
1	rs16857239	C	0.16	0.11	1.71 (1.38 - 2.10)	4.39E-07
2	rs2888881	A	0.3	0.23	1.60 (1.32 - 1.93)	8.44E-07
2	rs932206	G	0.28	0.35	0.63 (0.52 - 0.76)	9.95E-07
2	rs6735530	G	0.08	0.14	0.54 (0.42 - 0.70)	1.21E-06
6	rs3868542	G	0.28	0.36	0.63 (0.53 - 0.76)	1.59E-06
7	rs7787541	A	0.1	0.06	1.79 (1.40 - 2.28)	2.64E-06
2	rs2056202	A	0.18	0.13	1.60 (1.31 - 1.96)	3.35E-06
9	rs1932649	G	0.21	0.16	1.57 (1.29 - 1.90)	4.98E-06
11	rs3802785	A	0.26	0.2	1.54 (1.28 - 1.86)	5.36E-06
2	rs6757680	A	0.18	0.13	1.59 (1.30 - 1.94)	5.6E-06
10	rs2096285	G	0.26	0.33	0.65 (0.54 - 0.78)	5.77E-06
9	rs960957	A	0.32	0.38	0.65 (0.54 - 0.78)	6.05E-06
18	rs1551821	C	0.12	0.17	0.60 (0.48 - 0.75)	6.68E-06
2	rs7580232	A	0.18	0.13	1.58 (1.29 - 1.93)	7.26E-06
6	rs9405681	A	0.16	0.23	0.63 (0.52 - 0.78)	7.83E-06
1	rs650635	C	0.19	0.14	1.56 (1.28 - 1.91)	9.41E-06
16	rs6564872	A	0.43	0.5	0.64 (0.52 - 0.78)	9.71E-06

Table 3.6 lists the results obtained for the association analysis using the allelic model. Twenty six significant SNPs were identified at $p \leq 10^{-5}$, of which 4 were significant at Bonferroni correction level. The description of genes that harbour or are in the vicinity of these SNPs are described in Table 3.10.

Table 3.8 : SNPs ($p \leq 10^{-5}$) significant in the recessive model.

Chr	SNP	Minor allele	Case MAF	Control MAF	Odds ratio (95% CI)	p-value
10	rs7912419	A	0.2	0.16	3.49 (2.19 - 5.56)	2.4E-08
22	rs2050143	G	0.32	0.3	2.09 (1.59 - 2.74)	6.07E-08
19	rs4807664	C	0.17	0.13	3.56 (2.16 - 5.87)	1.22E-07
6	rs9493355	G	0.24	0.19	2.55 (1.78 - 3.66)	1.58E-07
***23**	rs5910340	G	0.11	0.09	1.32 (0.92 - 1.88)	2.49E-07
***23**	rs5910338	G	0.12	0.09	1.33 (0.93 - 1.90)	2.7E-07
1	rs6697552	A	0.28	0.32	0.31 (0.19 - 0.50)	4.3E-07
10	rs4146727	A	0.15	0.12	4.19 (2.29 - 7.67)	4.65E-07
3	rs9828404	A	0.16	0.12	3.54 (2.08 - 6.02)	7.35E-07
14	rs2225271	A	0.12	0.11	4.18 (2.25 - 7.76)	8.94E-07
6	rs3130564	A	0.32	0.25	2.09 (1.55 - 2.82)	9.73E-07
8	rs6470588	C	0.51	0.45	1.68 (1.36 - 2.07)	1.32E-06
22	rs2858344	C	0.1	0.08	5.95 (2.61 - 13.56)	1.47E-06
5	rs4410655	A	0.53	0.44	1.65 (1.34 - 2.04)	2.41E-06
15	rs8023560	C	0.12	0.09	5.46 (2.44 - 12.23)	3.64E-06
6	rs204993	A	0.3	0.27	1.80 (1.39 - 2.33)	5.94E-06
10	rs3781564	G	0.1	0.06	6.98 (2.59 - 18.83)	8.22E-06
15	rs2007084	A	0.09	0.08	5.04 (2.29 - 11.10)	8.63E-06
8	rs1372452	A	0.15	0.13	3.35 (1.90 - 5.88)	8.7E-06

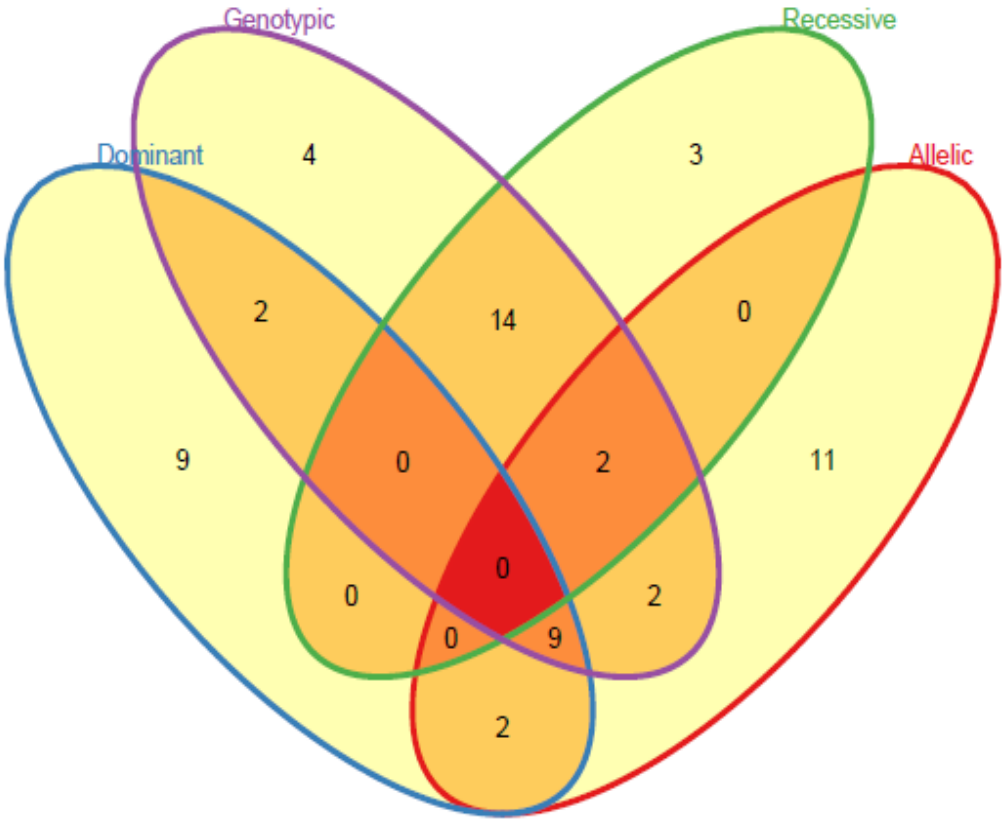
Table 3.7 lists the results obtained using dominant model. Twenty two significant SNPs were identified at $p \leq 10^{-5}$, of which 4 were significant at Bonferroni correction level. Genes associated with these SNPs are described in Table 3.10.

Table 3.9: SNPs ($p \leq 10^{-5}$) significant in the genotypic model.

Chr	SNP	Minor allele	Case MAF	Control MAF	Heterozygous OR (95% CI)	Homozygous OR (95% CI)	p-value
11	rs10838427	A	0.21	0.35	0.26 (0.21-0.33)	0.60 (0.44-0.81)	1.7E-31
19	rs2162296	A	0.11	0.21	0.43 (0.34-0.55)	0.22 (0.10-0.51)	1.7E-14
12	rs1563834	A	0.06	0.15	0.34 (0.25-0.45)	0.48 (0.22-1.05)	2.4E-13
22	rs2050143	G	0.32	0.3	0.69 (0.56-0.85)	1.79 (1.35-2.38)	1.4E-09
1	rs634899	A	0.29	0.35	0.53 (0.43-0.65)	0.80 (0.60-1.08)	1.3E-08
10	rs7912419	A	0.2	0.16	1.14 (0.92-1.40)	3.63 (2.27-5.81)	8.4E-08
10	rs10994443	A	0.06	0.11	0.42 (0.31-0.58)	1.08 (0.54-2.17)	2.5E-07
1	rs6697552	A	0.28	0.32	1.19 (0.98-1.45)	0.33 (0.20-0.55)	5.3E-07
19	rs4807664	C	0.17	0.13	1.12 (0.89-1.39)	3.65 (2.21-6.05)	5.3E-07
6	rs3130564	A	0.32	0.25	1.26 (1.03-1.54)	2.31 (1.68-3.15)	5.4E-07
6	rs9493355	G	0.24	0.19	1.09 (0.89-1.34)	2.62 (1.82-3.79)	7.8E-07
***23**	rs5910340	G	0.11	0.09	0.67 (0.46-1.00)	1.28 (0.90-1.83)	9.4E-07
***23**	rs5910338	G	0.12	0.09	0.70 (0.48-1.04)	1.30 (0.91-1.85)	1.2E-06
3	rs9828404	A	0.16	0.12	1.20 (0.96-1.49)	3.69 (2.16-6.30)	1.4E-06
1	rs16857239	C	0.16	0.11	1.67 (1.35-2.07)	2.14 (1.15-3.96)	2E-06
2	rs2888881	A	0.3	0.23	1.55 (1.27-1.88)	1.97 (1.36-2.86)	2.1E-06
10	rs3781564	G	0.1	0.06	1.39 (1.07-1.79)	7.32 (2.71-19.75)	2.2E-06
10	rs4146727	A	0.15	0.12	1.06 (0.85-1.33)	4.25 (2.32-7.79)	2.7E-06
2	rs4851692	A	0.24	0.17	1.43 (1.17-1.74)	2.47 (1.60-3.82)	2.7E-06
5	rs4410655	A	0.53	0.44	1.25 (0.99-1.58)	1.91 (1.47-2.48)	2.8E-06
6	rs3868542	G	0.28	0.36	0.67 (0.55-0.82)	0.50 (0.35-0.70)	3E-06
7	rs7787541	A	0.1	0.06	1.73 (1.35-2.22)	3.57 (1.29-9.88)	4.3E-06
14	rs2225271	A	0.12	0.11	0.94 (0.74-1.20)	4.13 (2.22-7.69)	5.1E-06
10	rs4918735	A	0.29	0.33	0.61 (0.49-0.74)	0.93 (0.69-1.26)	5.6E-06
2	rs932206	G	0.28	0.35	0.62 (0.51-0.76)	0.66 (0.48-0.90)	6.1E-06
20	rs2232081	A	0.17	0.12	1.44 (1.16-1.79)	2.99 (1.68-5.33)	6.5E-06
15	rs8023560	C	0.12	0.09	1.21 (0.95-1.53)	5.66 (2.52-12.68)	6.7E-06
13	rs9522264	G	0.39	0.41	0.62 (0.50-0.76)	0.99 (0.76-1.28)	6.9E-06
22	rs2858344	C	0.1	0.08	1.10 (0.85-1.43)	6.04 (2.65-13.78)	7.1E-06
2	rs6735530	G	0.08	0.14	0.54 (0.42-0.70)	0.54 (0.23-1.25)	7.6E-06
8	rs6470588	C	0.51	0.45	1.03 (0.82-1.29)	1.71 (1.32-2.20)	8.1E-06
14	rs8022758	G	0.19	0.21	0.63 (0.51-0.78)	1.41 (0.95-2.08)	8.1E-06
11	rs10501590	G	0.14	0.17	0.63 (0.50-0.80)	1.79 (1.10-2.92)	8.7E-06

Table 3.9 lists the results obtained using the dominant model. Thirty three significant SNPs were identified at $p \leq 10^{-5}$, of which 6 were significant at Bonferroni correction level. Genes associated with these SNPs are described in Table 3.10.

Figure 3.6: Venn diagram representing SNPs ($p<10^{-5}$) common to the four models.



The Venn diagram represents the overlap of SNPs between various models. The allelic, recessive, genotypic and dominant models identified 11, 3, 4 and 9 SNPs, respectively, specific only to these models while the rest of the SNPs significant at $p \leq 10^{-5}$ were common to more than one genetic model. None of the SNPs were common to all models. The overlap may indicate a number of possibilities, for instance if a SNP is significant in both the allelic and dominant model, it would suggest that just carrying the minor allele, regardless of the genotype being homozygous or heterozygous, can increase the susceptibility. For a SNP significant in both the allelic and genotypic model, indicate that heterozygous and homozygous carriers of minor allele have different but significant risks, while SNPs that are significant in both the allelic and recessive model indicate that homozygous carriers of the minor allele have a higher risk than any of the other carriers.

The three most significant SNPs in the GWAS analysis were rs10838427, rs2162296 and rs1563834.

rs10838427, located in the *PRDM11* gene, was significant in the allelic (OR= 0.51; 95% CI: 0.43 – 0.6 ; p = 1.56 $\times 10^{-17}$), dominant model (OR= 0.33; 95% CI: 0.27 - 0.40 ; p = 1.86 $\times 10^{-29}$) and genotypic model ($OR_{het=}$ 0.26; 95% CI_{het}: 0.21-0.33 and OR_{hom} = 0.60; 95% CI_{hom} :0.44-0.81; p = 1.7 $\times 10^{-31}$). rs2162296, located in gene *ZNF382,* also appeared significant in the allelic (OR = 0.45; 95% CI: 0.36 –0.55; p = 6.30 $\times 10^{-15}$), dominant (OR = 0.41 ; 95% CI: 0.33 - 0.51; p = 3.71 $\times 10^{-15}$) and genotypic (OR_{het}= 0.43 ; 95% CI_{het}: 0.34-0.55 and OR_{hom}= 0.22 ; 95% CI_{hom}: 0.10-0.51; p = 1.7 $\times 10^{-14}$) models. SNP rs1563834 located in gene *HMGA2* also appeared significant in the allelic (OR = 0.4; 95% CI: 0.31 - 0.52; p = 5.48 $\times 10^{-13}$),dominant (OR = 0.35 ; 95% CI: 0.26 - 0.46; p = 2.96 $\times 10^{-14}$) and genotypic (OR_{het}= 0.34 ; 95% CI_{het}: 0.25-0.45 and OR_{hom}= 0.48 ; 95% CI_{hom}: 0.22-1.05; p = 2.4 $\times 10^{-13}$) models.

The most significant SNP in the recessive model after Bonferroni correction was rs7912419 (OR =3.49; 95% CI: 2.19 - 5.56; p = 2.4 $\times 10^{-8}$) located in gene *ITIH2*. This SNP was also significant in the genotypic (OR_{het}= 1.14; 95% CI_{het}: 0.92-1.40 and OR_{hom}= 3.63; 95% CI_{hom}: 2.27-5.81; p = 8.4 $\times 10^{-8}$) model. rs2050143 located near gene *PDGFB* was significant in the recessive (OR =2.09; 95% CI: 1.59 - 2.74; p = 6.07 $\times 10^{-8}$) as well as the genotypic (OR_{het} = 0.69 ; 95% CI_{het}: 0.56-0.85 and OR_{hom} = 1.79; 95% CI_{hom}: 1.35-2.3; p = 1.4 $\times 10^{-9}$) model.

SNPs that appeared significant in the dominant model after Bonferroni correction level included rs634899 (OR =0.59; 95% CI: 0.49 - 0.71; p =3.38 $\times 10^{-8}$), located in gene *RP11-563D10.1.1,* which also appeared significant in the genotypic model (OR_{het} = 0.53; 95% CI_{het}: 0.43-0.65 and OR_{hom} = 0.80; 95% CI_{hom}: 0.60-1.08; p =1.3 $\times 10^{-8}$).

Table 3.10: Description of genes harbouring or located near significant SNPs ($p \leq 10^{-5}$) extracted using the NCBI2R[231] package in R[229]. Models that identified significant SNPs and the least p value obtained by these models is shown.

Markers[Model/s] p-value	Type	Cytogenetic location	Gene	Gene summary
rs10501590[G] 8.70E-06	intronic	11q14.1	*DLG2*	Encodes a member of the membrane-associated guanylate kinase (MAGUK) family which may bind to a related family member and interact at postsynaptic sites to form a multimeric scaffold for the clustering of receptors, ion channels, and associated signaling proteins.
rs10507935[A] 2.60E-10	intergenic	13q31.3	*RNU6-67*	
rs10509535[A] 1.16E-06	intergenic	10q23.2	*RP11-380G5.4.1*	
rs10838427[AGD] 1.70E-31	intronic	11p11	*PRDM11*	
rs10994443[AGD] 2.50E-07	intronic	10q21	*ANK3*	Ankyrins are a family of proteins that play a role in cell motility, activation, proliferation, contact, and the maintenance of specialized membrane domains. **Developmental Biology*
rs12683609[A] 7.18E-06	intergenic	9p22.3	*C9ORF92*	
rs1372452[R] 8.70E-06	intergenic	8q24.21	*AC068570.1*	
rs1551821[D] 6.68E-06	intronic	18q12.1-q21.1	*SLC14A2*	Encodes a protein that belongs to the urea transporter family and plays an important role in the urinary concentration mechanism. **Transmembrane transport of small molecules*
rs1563834[AGD] 2.96E-14	intronic	12q15	*HMGA2*	Encodes a protein that belongs to the non-histone chromosomal high mobility group (HMG) which function as a architectural and transcriptional regulating factor, and contains structural DNA-binding domain. Gene may be involved in diet-induced obesity. *$ Transcriptional misregulation in cancer*
rs16857239[AGD] 2.35E-07	upstream (5kb)	1q25.3	*CACNA1E*	These calcium channels regulate the entry of calcium ions into excitable cells, and participate in a variety of calcium-dependent processes, including muscle contraction, hormone or neurotransmitter release, gene expression, cell motility, cell division, cell death and modulation of

				firing patterns of neurons important for information processing. *$MAPK signalling pathway, $Calcium signalling pathway, $Type II diabetes mellitus*
rs17062322[A] **5.92E-06**	intronic	6q23	*EYA4*	Encodes a member of the eyes absent (EYA) family of proteins. It finds its importance in eye development, and for continued function of the mature organ of Corti. Mutations in this gene are associated with postlingual, progressive, autosomal dominant hearing loss at the deafness, autosomal dominant nonsyndromic sensorineural 10 locus. Gene defect also causes dilated cardiomyopathy 1J.
rs1932649[D] **4.98E-06**	intergenic	9q22.2	*SLC28A3*	Regulates vascular tone, neurotransmission, adenosine concentration in the vicinity of cell surface receptors, and transport and metabolism of nucleoside drugs. **Transmembrane transport of small molecules*
rs2007084[R] **8.63E-06**	intronic	15q25-q26	*ANPEP*	Plays a role in the final digestion of peptides generated from hydrolysis of proteins by gastric and pancreatic proteases. Defects lead to various types of leukemia or lymphoma. *Hematopoietic cell lineage, $ Glutathione metabolism, $ Renin-angiotensin system, $ Metabolic pathways*
rs204993[R] **5.94E-06**	intronic	6p21.3	*PBX2*	Encodes a ubiquitously expressed member of the TALE/PBX homeobox family, the protein of which binds to the TLX1 promoter and is a transcriptional activator.
rs2050143[GR] **1.40E-09**	intergenic	22q13.1	*PDGFB*	Encodes a member of the platelet-derived growth factor family. Polymorphisms are associated with meningioma. *$ Prostate cancer, $ Glioma, $ Melanoma, $ Renal cell carcinoma, $ Transcriptional misregulation in cancer, $ Pathways in cancer, $ Regulation of actin cytoskeleton, $ HTLV-I infection, $ Focal adhesion, $ Gap junction, $ Cytokine-cytokine receptor interaction, $ MAPK signalling pathway, *Hemostasis, *Signal Transduction*
rs2056202[D] **3.35E-06**	intronic	2q24	*SLC25A12*	Encodes a calcium-binding mitochondrial carrier protein. Polymorphisms may be associated with autism, and global cerebral

				hypomyelination. **Metabolism, *Metabolism of proteins*
rs2096285[AD] **4.91E-06**	intronic	10q26	*PTPRE*	Encodes a member of the protein tyrosine phosphatase (PTP) family that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation.
rs2162296[AGD] **3.71E-15**	intronic	19q13.12	*ZNF382*	Encodes a KRAB domain zinc finger transcription factor (KZNF) that may play a critical role in the regulation of many cellular processes including differentiation, proliferation and apoptosis, inhibition of activating protein 1 (AP-1) and nuclear factor kappa-B (NF-kB) signalling and may function as a tumour suppressor in multiple carcinomas. **Gene Expression*
rs2225271[GR] **8.94E-07**	intergenic	14q31.2	*RP11-22K10.1.1*	
rs2232081[AG] **3.39E-06**	intronic	20p12.3	*FERMT1*	Encodes a protein belonging to the fermitin family which is involved in integrin signalling and linkage of the actin cytoskeleton to the extracellular matrix.
rs2304214[A] **3.02E-06**	synonymous	19q13	*DLL3*	Encodes a protein belonging to the delta protein ligand family that functions as Notch ligands. *$ Notch signalling pathway*
rs2858344[GR] **1.47E-06**	intronic	22q12.3	*SYN3*	Belongs to the synapsin gene family and may have a role in several neuropsychiatric diseases like schizophrenia. Another gene, TIMP3 is located within an intron of this gene and is transcribed in the opposite direction. **Neuronal System*
rs2888881[AGD] **4.96E-07**	intronic	2p21	*PLEKHH2*	
rs3130564[AGR] **5.40E-07**	intronic	6p21.3	*PSORS1C1*	Confers susceptibility to psoriasis and systemic sclerosis.
rs3130977[A] **7.73E-06**	upstream (2kb)	6p21.3	*C6orf15*	
rs3781564[GR] **2.20E-06**	intronic	10q25.3	*PNLIPRP1*	*$ Fat digestion and absorption, $ Pancreatic secretion, $ Metabolic pathways, $ Glycerolipid metabolism*
rs3802785[D] **5.36E-06**	upstream	11p11.2	*LOC221122*	
rs3868542[AGD] **7.62E-07**	upstream (2kb)	6p21.33	*PSORS1C3*	
rs4146727[GR] **4.65E-07**	intronic	10q24.1	*PIK3AP1*	*$ B cell receptor signalling pathway, *Immune System*
rs4410655[AGR] **1.06E-06**	Non coding transcript	5q35.3	*AACSP1*	
rs4713175[A] **9.18E-06**	intergenic	6	*TRNAA40*	

rs4807664[GR] **1.22E-07**	intronic	19p13.3	*UHRF1*	The protein binds to specific DNA sequences, recruits a histone deacetylase to regulate gene expression, functions in the p53-dependent DNA damage checkpoint and plays a major role in the G1/S transition and retinoblastoma gene expression.
rs4822112[A] **5.12E-06**	downstream	22q13.2	*NFAM1*	Encodes a type I membrane receptor that activates cytokine gene promoters and contains an immunoreceptor tyrosine-based activation motif (ITAM). It regulates the signaling and development of B-cells.
rs4851692[AG] **9.47E-07**	Non coding transcript	2q12.1	*LOC100287010*	
rs4918735[G] **5.60E-06**	intronic	10q25-q26	*TECTB*	Encodes the major non-collagenous proteins of the tectorial membrane of the cochlea.
rs5910338[GR] **2.70E-07**	intergenic	Xq24	*WDR44*	Alternate splicing results in multiple transcript variants and the encoded protein may play a role in endosome recycling.
rs5910340[GR] **2.49E-07**				
rs634899[GD] **1.30E-08**	intergenic	1q31.3	*RP11-563D10.1.1*	
rs6470588[GR] **1.32E-06**	intronic	8q24	*PVT1*	
rs650635[D] **9.41E-06**	intronic	1q25.3	*CACNA1E*	These calcium channels regulate the entry of calcium ions into excitable cells, and participate in a variety of calcium-dependent processes, including muscle contraction, hormone or neurotransmitter release, gene expression, cell motility, cell division, cell death and modulation of firing patterns of neurons important for information processing. *$MAPK signalling pathway, $Calcium signalling pathway, $Type II diabetes mellitus*
rs6564872[D] **9.71E-06**	intronic	16q24.1	*GAN*	Defects in this gene are a cause of giant axonal neuropathy (GAN). Encoded protein regulates neurofilament architecture and mediate the ubiquitination and degradation of some proteins. **Immune System*
rs6697552[GR] **4.30E-07**	intergenic	1q43	*RP11-331N16.1.1*	
rs6735530[AGD] **1.21E-06**	intronic	2p21	*CRIPT*	
rs6739713[A] **4.22E-06**	intergenic	2q21.3	*R3HDM1*	
rs6757680[D] **5.60E-06**	intronic	2q31.2-q33.1	*HAT1*	Encodes protein involved in the rapid acetylation of newly synthesized cytoplasmic histones which

				plays an important role in replication-dependent chromatin assembly. *$ Alcoholism*
rs7580232[D] **7.26E-06**	intronic	2q24	*SLC25A12*	Encodes a calcium-binding mitochondrial carrier protein. Mutations may cause autism and global cerebral hypomyelination. **Metabolism, *Metabolism of proteins*
rs7709656[A] **5.37E-06**	intronic	5q32	*GLRA1*	Defected gene causes startle disease (STHE), also known as hereditary hyperekplexia or congenital stiff-person syndrome. Encodes a protein that forms a part of the pentameric inhibitory glycine receptor that mediates postsynaptic inhibition in the central nervous system. *$ Neuroactive ligand-receptor interaction, *Transmembrane transport of small molecules*
rs7787541[AGD] **8.54E-07**	intronic	7p21.1	*AC007091.1.1*	
rs7912419[GR] **2.40E-08**	intronic	10p15	*ITIH2*	Associated with prevention of tumour metastasis and extracellular matrix stabilization.
rs8022758[G] **8.10E-06**	Non coding transcript, 3-prime-utr	14q21	*ATXN3*	Mutation causes an autosomal dominant neurologic disorder called the Machado-Joseph disease, also known as spinocerebellar ataxia-3. *$ Protein processing in endoplasmic reticulum*
rs8023560[GR] **3.64E-06**	regulatory	15q21.1	*TRIM69*	Encodes a member of the RING-B-box-coiled-coil (RBCC) family.
rs932206[GD] **9.95E-07**	intergenic	2q22.1	*AC068492.1.1*	
rs9366778[A] **6.57E-06**	intergenic	6p21.33	*WASF5P*	Is a pseudogene belonging to the family of genes encoding Wiskott-Aldrich syndrome (WAS) causing Wiskott-Aldrich syndrome (immune system). The proteins encoded bring about transmission of signals to the actin cytoskeleton.
rs9405681[AD] **2.04E-06**	intergenic	6p25.3	*EXOC2*	*Exocyst complex formation, Diabetes pathways, Insulin Synthesis and Processing*
rs9493355[GR] **1.58E-07**	intergenic	6q23.2	*RPL21P66*	
rs9522264[G] **6.90E-06**	intergenic	13q34	*RP11-65D24.2.1*	
rs953715[A] **6.04E-06**	intronic	9q22.33	*HIATL2*	
rs960957[D] **6.05E-06**	intergenic	9q32	*RP11-18B16.1.1*	
rs9828404[GR] **7.35E-07**	upstream (5kb)	3p26.1	*AC090955.3.1*	

pathways are italicised.*- Reactome Event;$ - KEGG pathway; Pathway names are italicised.

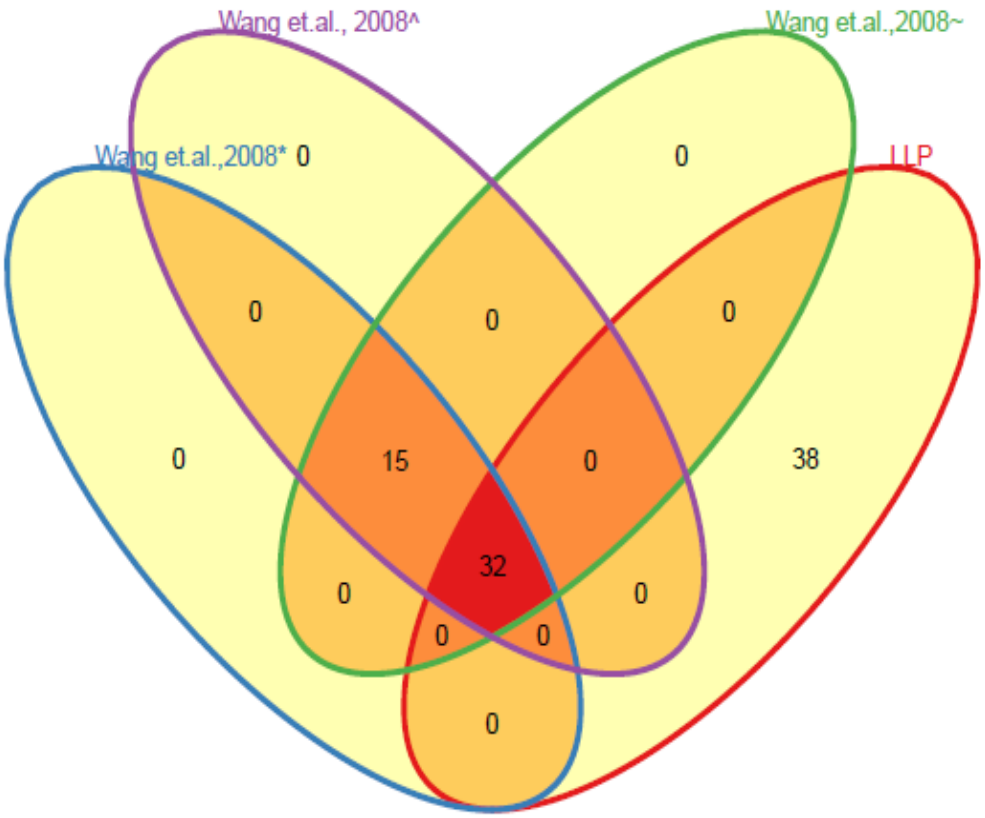
3.5 Discussion

This study identified SNPs associated with lung cancer using 526 cases and 2816 control individuals that passed quality control, using the allelic, dominant, recessive and genotypic model. Varying number of SNPs were identified in different models (Table 3.6-3.9) and an overlap of SNPs was seen between models (Figure 3.6).

The results of the allelic model were compared to other significant publications that conducted a GWAS using similar models. One of the reasons why significant SNPs in this study were not reported in these previous publications, may be due to the various significance cut off values chosen[113, 114, 193, 203, 233]. The most prominent SNPs in these publications include rs402710 on chromosome 5p15.33 and rs1051730 on chromosome 15q25.1[113, 114, 193, 203, 233]. The former SNP was not present in this study, however the SNP rs2736100 on chromosome 5p15.33, which lies in the same chromosomal region as rs402710 was present and produced an OR of 0.98 (95% CI: 0.86-1.11; p = 0.73), whereas rs1051730 produced an OR of 1.14 (95% CI: 1.00-1.31; p = 0.054) in the allelic model (Table 3.11).

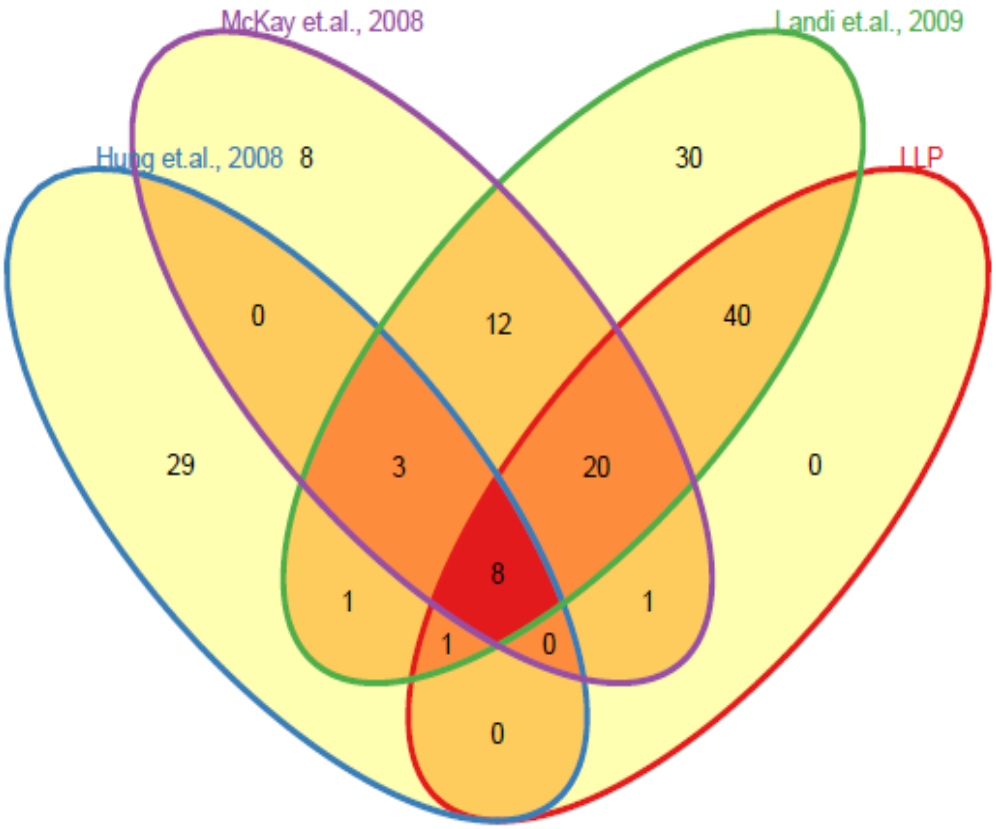
Furthermore, significant SNPs within chromosomal regions 5, 6 and 15, from significant publications[113, 193, 203, 233] were compared to the present LLP study using Venn diagrams (Figure 3.7 and Figure3.8). P values and odds ratios, where available, were extracted from the supplementary sections of these publications and merged with comparable SNPs from the LLP study (Table 3.11), leaving out those that were not significant (p >0.05) or not genotyped in the LLP population for the Venn diagram.

Figure 3.7: Venn diagram comparing LLP SNPs ($p < 0.05$) in chromosome 5, 6 and 15 to Wang *et al*., 2008 ($p < 10^{-4}$).



Wang *et al*. (2008) [233] reported SNPs significant at $p < 10^{-4}$ in an univariate model using only the UK population and in a meta-analysis using a fixed and random model for the UK, Texas and IARC populations. Thirty two SNPs were common to the LLP study and the unadjusted, fixed and random effect model from Wang *et al*. (2008)[233] while 38 SNPs which are specific to the LLP study, were extracted because they were present in the following significant studies[113, 193, 203, 233] (Figure 3.7). The unadjusted model in Wang *et al*. (2008)[233] was similar to the present study.

Figure 3.8: Venn diagram comparing LLP SNPs ($p < 0.05$) in chromosome 5, 6 and 15 to Hung *et al*., 2008 ($p < 10^{-5}$), Landi *et al*., 2009 ($p < 10^{-4}$) and McKay *et al*., 2008 ($p < 10^{-5}$).



Hung *et al*. (2008)[113] reported SNPs significant at $p_{(trend)} < 5\times10^{-5}$ in a multivariate model adjusted for age, sex and country with the variant coded in the log additive mode. Also, significant SNPs ($p < 10^{-5}$) in the 15q25.1 region, following fine genotyping, were also included. McKay *et al*. (2008)[193] reported SNPs significant at $p < 10^{-4}$ using two models, one, adjusted for age, sex and country and the other adjusted for age, sex, country and eigen values. The model without eigen values was used when comparing it to the present study. Landi *et al*. (2009)[203] reported significant SNPs ($p < 10^{-4}$) in a meta-analysis of 11 studies using an unadjusted model. Eight SNPs were common to all studies (Figure 3.8). None of the SNPs were exclusive to the LLP study.

For important chromosome regions in lung cancer, when compared to the previously published significant SNPs in GWAS, the most significant SNP in chromosome 5 in this study, rs401681 (OR=0.77; 95% CI: 0.67-0.88; p = 0.00014) was also significant in the unadjusted analysis carried out by Landi *et al*. (2009)[203] (OR=0.89; 95% CI: 0.86-0.92; p =6.65 x 10^{-11}) and Wang *et al*. (2008)[233] on the UK population (p = 0.00558) and in the multivariate analysis by McKay *et al*. (2008)[193] adjusting for age, sex, country without (OR=1.19; 95% CI: 1.11-1.28; p =2.00 x 10^{-06}) and with (OR=1.19; 95% CI: 1.11-1.28; p =3.00 x 10^{-06}) eigen values (Table 3.11).

rs3130564 (OR= 1.43; 95% CI: 1.24-1.65; p =7.44 x 10^{-07}) was the most significant SNP reported in this study for chromosome 6, and was also significant in the unadjusted analysis by Landi *et al*. (2009)[234] (OR=1.10; 95% CI: 1.06-1.15; p =1.21 x 10^{-05}) while rs4887077 (OR=1.16; 95% CI: 1.02-1.33; p =0.025) was also significant in unadjusted analysis of the UK population by Wang *et al*. (2008)[233] (p = 6.10 x 10^{-05}) and the multivariate models adjusted for age, sex and country without (OR=1.20; 95% CI: 1.12-1.29; p =3.00 x 10^{-07}) and with (OR=1.20; 95% CI: 1.12-1.29; p =6.00 x 10^{-07}) eigen values by McKay *et al*. (2008)[193] for chromosome 15. While these studies only look at the additive mode of inheritance, our study also covers dominant, recessive and genotypic models (Table 3.11).

The most significant SNP identified in this study lies within the *PRDM11* gene. The *PRDM* family of genes have recently gained interest as they have been associated with several human disease and cancers[235, 236]. Not much published information is available on *PRDM11,* although members of the gene family are found to be deregulated in several solid tumours, where they function as both tumour suppressors and drivers of oncogenic events[235].

The PRDM gene family have a PR domain and differing number of Zn finger repeats, except *PRDM11,* which does not have any Zn finger[236]. They play a key role in regulating expression, cell proliferation and differentiation through signal transduction[236]. Members

of the PRDM family are homologous to catalytic SET (Suppressor of variegation 3-9, Enhancer of zeste and Trithorax) domains that are histone methyltrasferases[236]. An interesting property of this group is that different molecular forms can result due to alternative splicing or by different promoters[236].

Potential tumour suppressors include *PRDM1*, *PRDM2*, *PRDM5* and *PRDM12*, inactivated in many cancers[237]. Inactivation of *PRDM1* and *PRDM2*, not necessarily together, is common in diffuse large B cell lymphoma (DLBCL)[237]. *PRDM1* is also silenced in natural killer cell lymphoma and *PRDM12* in chronic myeloid leukaemia[237].

Oncogenic properties are observed in *PRDM3*, *PRDM13*, *PRDM14* and *PRDM16.* Acute myeloid leukemia results if *PRDM3* or *PRDM16* devoid of *PRDI-BF1-RIZ1* homologous domain are expressed[237]. *PRDM14* is associated with breast cancer and lymphoid leukaemia while, *PRDM3* and *PRDM13* is linked to nasopharyngeal carcinoma and medulloblastoma, respectively[237].

It is hypothesized that tumourigenesis for the PR genes occurs in a *yin-yang* mechanism where an imbalance of PR-plus product (product with the PR domain) that is tumour suppressing and PR-minus product (product without the PR domain) that is oncogenic, leads to cell transformation[238]. The imbalance is caused by activation and inactivation of PR-minus and PR-plus product, respectively, or both[238]. The instability and altered state of these genes in cancers may be due to its location at the end of the chromosome. Inactivation of this gene is mostly through loss of expression[238].

PRDM11 is associated with thyroid function regulation[239] and in near-haploid lymphoblastoid leukaemia[240]. A SNP in *PRDM11* was identified by NGS, in near haploid lymphoblastoid leukaemia and was reported to be associated with epigenetic gene[240].

The second most significant gene *ZNF382* is a functional tumour suppressor in multiple carcinomas that may contribute to the initiation of apoptosis and inhibition of cell proliferation[241]. The antitumourigenic activity of this gene is caused by suppressing both NF-κB and the AP-1 signalling pathways while the epigenetic silencing of *ZNF382* may activate cancer signalling pathways during tumourigenesis[241].

PRDM genes and *ZNF382* are tumour suppressors that are inactivated in cancer. Mutations in such genes, unlike the present study, should show an increased risk of cancer. The protective nature of these mutations can be explained by two reasons.

The first reason may be that this variation is age related[242]. Low methylation of a tumour suppressor gene *THBS4*, was observed in tumour tissue, in a comparison between colorectal adenomas and normal colonoscopies of colorectal cancer patients, concluding that it was not associated with promoting colorectal neoplasia[242]. The study considers the behaviour to be linked to age, referred to as "type A"[242]. "Type A" genes are methylated in both normal and tumour tissue but the extent of methylation is proportional to the normal tissue age as oppose to "type C" which is specific to tumour tissue[242]. This study shows that though the gene is a tumour suppressor its methylation is higher in normal compared to tumour tissue[242].

The second reason may be that there was no adjustment for important confounding variables like age, sex and smoking pack years. For instance SNP rs401681 in chromosome 5 showed an significant increased risk when adjusted for age, sex, country and eigen values in the multivariate analysis[193] but a significant decreased risk in the univariate analysis[203, 233]. Controlling for covariates might have produced an increased risk for these significant SNPs.

The third significant SNP is located in gene *HMGA2*. *HMGA2* is an oncogene that functions by binding and inactivating pRB in lung carcinogenesis[243]. Its protein is reported to be overexpressed in non-small cell lung cancer[244]. The detection of overexpressed levels of *HMGA2* can be used as a prognostic factor for early detection of lung cancer as the overexpression has also been seen in non-cancerous cells[244]. Furthermore, *HMGA2* overexpression can be used as a molecular factor for non-small cell lung cancer[244]. The significance of this gene in NSCLC can be tested using knocked out mice[245].

Similar results were seen by Quaye *et al*. (2009)[246] that carried a logistic regression to assess the association of tag SNPs in oncogenes with ovarian cancer[246]. SNP rs11683487 in gene *NMI*, produced a decreased risk of ovarian cancer with heterozygous OR = 0.80 (95% CI: 0.69-0.93) and homozygous OR = 0.87 (95% CI: 0.71-1.02) (p_{trend} = 0.038)[246]. Dominant model was the best model for this SNP with OR = 0.81 (95% CI: 0.71-0.94; p=0.004)[246]. Furthermore, as mentioned above the odds could have increased if the model was adjusted for age, smoking pack years and sex.

Other genes that appeared to be significant in the various models include *ITIH2, PDGFB, DIAPH2, UHRF1, RPL21P66, RNU6-67 and RP11-563D10.1.1*[247]. *ITIH2* is reported to be downregulated in solid tumours of the breast, colon and lung, which may be associated with carcinogenesis and or progression of these malignancies[247].

UHRF1 is reported to be overexpressed in lung cancer and causes epigenetic changes of tumour suppressor genes by maintaining their promoters in a hypermethylated state[248]. The *PDGFB* gene is a growth factor and mutations in this gene is associated with meningioma and dermatofibrosarcoma protuberans[249]. It is involved in functions such as transcriptional misregulation in cancer, regulation of actin cytoskeleton, focal adhesion, gap junction, cytokine-cytokine receptor interaction and MAPK signalling pathway while *DIAPH2* is involved in cytokinesis[249]. There is no published information available for gene

RPL21P66, RNU6-67 and *RP11-563D10.1.1*. Genes significant at the $p < 1 \times 10^{-05}$ level in this study are described in Table 3.10.

The next steps, to extend this genome wide association study, would be to replicate the results in another population of cases and controls using all three models and to evaluate the significance of these SNPs after controlling for important covariates like age, sex and tobacco smoking, since these measurements were unavailable for the 1958 Birth Cohort controls[194]. If significant it would indicate that these polymorphisms are directly associated with lung cancer and unaffected by the adjusted confounders[112, 113].

The SNPs that have been identified could be evaluated to see whether they influence the gene expression of neighbouring genes, and linkage disequilibrium analysis could be performed to more accurately define the SNP, or SNPs, that influence cancer susceptibility[245]. If neighbouring genes are influenced by the identified SNPs, the role of these genes in cancer-associated pathways, including cell cycle progression, cellular growth, apoptosis or DNA repair, could be tested in cell lines and in knock out studies[245]. Furthermore, mapping these genes to various known pathways, carrying out SNPs interaction studies and connecting expression profiles of these genes may not only help identify crucial pathways but also help discover new ones (Chapter 5).

Table 3.11: Significant SNPs from chromosome 5, 6 and 15, identified in published genome wide association analysis.

				Hung *et al.*, 2008*	Landi *et al.*, 2009[£]	Wang *et al.*, 2008[£]			McKay *et al.*, 2008	
SNP	**Chr**	**p-value**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI) fixed**	**OR(95% CI) random**	**UKGWA p-value**	**OR(95% CI) Model 1***	**OR(95% CI) Model 2[δ]**
rs11950678	5	0.3321	0.90(0.72-1.12)	0.78(0.69-0.88)						
rs10072467	5			1.25(1.13-1.38)						
rs1366625	5			1.21(1.11-1.33)						
rs9291949	5				1.09(1.04-1.13)					
rs402710	5								1.22(1.13-1.32)	1.22(1.13-1.32)
rs2736100	5	0.7276	0.98(0.86-1.11)		1.12(1.08-1.16)				1.18(1.10-1.26)	1.19(1.11-1.27)
rs31489	**5**	**0.0002317**	**0.77(0.67-0.89)**		**0.89(0.86-0.92)**	**0.88(0.84-0.94)**	**0.88(0.84-0.94)**	**0.02351822**	**1.20(1.12-1.29)**	**1.20(1.11-1.29)**
rs329122	5	0.4869	0.95(0.83-1.09)		0.93(0.90-0.96)					
rs401681	**5**	**0.0001402**	**0.77(0.67-0.88)**		**0.89(0.86-0.92)**	**0.88(0.83-0.93)**	**0.88(0.83-0.93)**	**0.00558291**	**1.19(1.11-1.28)**	**1.19(1.11-1.28)**
rs4635969	**5**	**0.0006673**	**0.73(0.61-0.88)**		**0.88(0.84-0.92)**					
rs4975616	**5**	**0.0002789**	**0.78(0.68-0.89)**		**0.90(0.87-0.93)**	**0.88(0.83-0.93)**	**0.88(0.83-0.93)**	**0.00272402**	**1.17(1.09-1.26)**	**1.17(1.09-1.26)**
rs3130564	**6**	**7.44E-07**	**1.43(1.24-1.65)**		**1.10(1.06-1.15)**					
rs3132610	**6**	**2.17E-05**	**1.43(1.21-1.70)**		**1.18(1.12-1.25)**	**1.2(1.1-1.31)**	**1.2(1.06-1.35)**	**0.01528174**	**1.26(1.13-1.41)**	**1.25(1.12-1.40)**
rs3094694	**6**	**2.26E-05**	**1.39(1.19-1.62)**		**1.12(1.07-1.17)**					
rs3130544	**6**	**3.69E-05**	**1.42(1.20-1.68)**		**1.17(1.11-1.23)**					
rs3132580	**6**	**4.79E-05**	**1.40(1.19-1.65)**		**1.15(1.10-1.21)**					
rs2187668	**6**	**6.44E-05**	**1.42(1.19-1.68)**		**1.18(1.12-1.24)**	**1.22(1.12-1.33)**	**1.21(1.07-1.37)**	**0.00364272**		
rs3094054	**6**	**6.76E-05**	**1.41(1.19-1.68)**		**1.20(1.13-1.27)**					
rs4122189	**6**	**7.62E-05**	**0.71(0.60-0.84)**		**0.92(0.89-0.96)**					
rs3132622	**6**	**8.38E-05**	**1.30(1.14-1.49)**		**1.07(1.04-1.11)**					

				Hung *et al.*, 2008*	Landi *et al.*, 2009[£]	Wang *et al.*, 2008[£]			McKay *et al.*, 2008	
SNP	Chr	p-value	OR(95% CI)	OR(95% CI)	OR(95% CI)	OR(95% CI) fixed	OR(95% CI) random	UKGWA p-value	OR(95% CI) Model 1*	OR(95% CI) Model 2[δ]
rs630379	6	9.58E-05	1.31(1.15-1.51)		1.08(1.04-1.13)					
rs2844773	6	0.0001384	1.37(1.16-1.61)		1.11(1.06-1.17)					
rs7750641	6	0.0001689	1.38(1.17-1.64)		1.17(1.11-1.24)	1.23(1.13-1.34)	1.22(1.08-1.38)	0.00160221	1.25(1.12-1.40)	1.24(1.11-1.39)
rs389884	6	0.0001956	1.41(1.18-1.68)		1.19(1.13-1.26)	1.24(1.13-1.35)	1.21(1.03-1.43)	0.00128195	1.29(1.15-1.44)	1.28(1.14-1.43)
rs2734986	6	0.0002206	1.34(1.15-1.57)		1.15(1.09-1.20)	1.17(1.09-1.26)	1.17(1.09-1.26)	0.04663749	1.22(1.11-1.34)	1.21(1.10-1.33)
rs3130380	6	0.0002255	1.39(1.17-1.66)		1.20(1.13-1.27)	1.22(1.12-1.34)	1.22(1.12-1.34)	0.00242865		
rs659445	6	0.0002665	1.28(1.12-1.47)		1.08(1.04-1.12)					
rs3130350	6	0.0002805	1.39(1.16-1.66)		1.21(1.14-1.28)				1.28(1.14-1.43)	1.27(1.12-1.42)
rs3094073	6	0.0002997	1.34(1.14-1.58)		1.11(1.06-1.17)					
rs886424	6	0.0003218	1.36(1.15-1.60)		1.16(1.10-1.22)					
rs1794282	6	0.0004289	1.38(1.15-1.66)		1.20(1.13-1.26)	1.2(1.1-1.31)	1.18(0.99-1.4)	0.0085809	1.26(1.13-1.41)	1.25(1.12-1.40)
rs9261290	6	0.0005052	1.37(1.15-1.63)		1.20(1.13-1.27)	1.22(1.11-1.33)	1.21(1.11-1.33)	0.00677158	1.28(1.14-1.44)	1.28(1.13-1.44)
rs2517861	6	0.0005246	1.29(1.12-1.49)		1.10(1.05-1.14)					
rs535586	6	0.0005315	1.27(1.11-1.45)		1.08(1.04-1.12)					
rs8321	6	0.0006476	1.36(1.14-1.62)		1.20(1.14-1.28)	1.23(1.12-1.35)	1.23(1.12-1.35)	0.00421701	1.29(1.15-1.45)	1.29(1.14-1.45)
rs3099844	6	0.0007861	1.35(1.13-1.61)		1.15(1.09-1.21)	1.2(1.1-1.31)	1.19(1.04-1.36)	0.00326065		
rs2233956	6	0.001188	1.28(1.10-1.49)		1.12(1.07-1.17)					
rs259919	6	0.001372	1.25(1.09-1.43)		1.10(1.06-1.14)					
rs3132685	6	0.00163	1.33(1.11-1.58)		1.19(1.12-1.25)				1.28(1.14-1.43)	1.27(1.14-1.42)
rs3131379	6	0.001684	1.34(1.12-1.61)		1.20(1.14-1.27)	1.26(1.16-1.38)	1.24(1.05-1.47)	0.00020471	1.30(1.16-1.45)	1.29(1.15-1.45)
rs1233487	6	0.001945	1.25(1.09-1.44)		1.09(1.05-1.13)					
rs3094127	6	0.002014	1.27(1.09-1.47)		1.11(1.06-1.16)					

				Hung *et al.*, 2008*	Landi *et al.*, 2009[£]	Wang *et al.*, 2008[£]			McKay *et al.*, 2008	
SNP	Chr	p-value	OR(95% CI)	OR(95% CI)	OR(95% CI)	OR(95% CI) fixed	OR(95% CI) random	UKGWA p-value	OR(95% CI) Model 1*	OR(95% CI) Model 2[δ]
rs2535319	6	0.00202	1.23(1.08-1.41)		1.08(1.04-1.12)					
rs806977	6	0.002427	0.81(0.71-0.93)		1.08(1.04-1.12)					
rs2523987	6	0.002638	1.29(1.09-1.52)		1.14(1.08-1.20)					
rs2746150	6	0.002665	1.32(1.10-1.59)	1.41(1.21-1.65)	1.19(1.11-1.26)	1.23(1.12-1.35)	1.2(0.98-1.48)	0.01795292	1.31(1.16-1.47)	1.30(1.15-1.46)
rs2517598	6	0.00296	1.28(1.09-1.50)		1.12(1.06-1.17)					
rs1235162	6	0.003002	1.31(1.10-1.58)		1.20(1.13-1.28)	1.24(1.13-1.36)	1.23(1.09-1.39)	0.00868927	1.30(1.15-1.46)	1.29(1.15-1.46)
rs7775397	6	0.004471	1.30(1.09-1.57)		1.20(1.14-1.27)	1.23(1.12-1.34)	1.2(0.97-1.47)	0.00113253	1.30(1.16-1.46)	1.29(1.15-1.45)
rs3134942	6	0.004632	1.29(1.08-1.53)		1.15(1.09-1.21)					
rs2524005	6	0.004852	1.24(1.07-1.44)		1.11(1.07-1.16)	1.15(1.07-1.23)	1.15(1.07-1.23)	0.02974593	1.20(1.10-1.31)	1.20(1.10-1.31)
rs3129073	6	0.005119	1.24(1.07-1.45)		1.10(1.05-1.14)					
rs7762279	6	0.006519	1.30(1.08-1.57)		1.15(1.08-1.22)					
rs4324798	6	0.01095	1.28(1.06-1.54)	1.45(1.23-1.69)	1.16(1.09-1.24)	1.24(1.13-1.36)	1.21(0.97-1.52)	0.02099102	1.30(1.15-1.47)	1.29(1.14-1.46)
rs3131093	6	0.01101	1.28(1.06-1.54)	1.41(1.21-1.65)	1.16(1.09-1.24)	1.23(1.12-1.35)	1.2(0.97-1.49)	0.02094826		
rs13194504	6	0.01174	1.28(1.06-1.54)	1.43(1.22-1.68)	1.15(1.08-1.23)	1.21(1.1-1.33)	1.19(0.94-1.5)	0.05454524	1.29(1.14-1.46)	1.28(1.13-1.45)
rs3096697	6	0.01183	1.21(1.04-1.41)		1.10(1.05-1.15)					
rs1150735	6	0.01196	1.19(1.04-1.36)		1.08(1.05-1.12)					
rs10484399	6	0.01217	1.27(1.05-1.54)		1.14(1.07-1.22)					
rs2535238	6	0.01313	1.21(1.04-1.40)		1.10(1.05-1.15)					
rs3749971	6	0.01624	1.26(1.04-1.51)	1.39(1.2-1.62)	1.16(1.09-1.23)	1.21(1.11-1.33)	1.19(0.96-1.46)	0.02462257	1.28(1.14-1.44)	1.27(1.13-1.43)
rs3129791	6	0.01815	1.26(1.04-1.52)	1.42(1.21-1.66)	1.17(1.10-1.24)	1.23(1.12-1.35)	1.2(0.96-1.5)	0.02167858	1.29(1.14-1.45)	1.28(1.13-1.45)
rs3130893	6	0.01894	1.25(1.04-1.51)	1.41(1.21-1.65)	1.17(1.10-1.24)	1.23(1.12-1.36)	1.21(0.98-1.49)	0.01480863	1.29(1.14-1.45)	1.28(1.13-1.45)
rs2747457	6	0.02401	1.19(1.02-1.37)		1.10(1.06-1.15)					

				Hung *et al.*, 2008*	Landi *et al.*, 2009[£]	Wang *et al.*, 2008[£]			McKay *et al.*, 2008	
SNP	**Chr**	**p-value**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI) fixed**	**OR(95% CI) random**	**UKGWA p-value**	**OR(95% CI) Model 1***	**OR(95% CI) Model 2[δ]**
rs175597	**6**	**0.02824**	**1.22(1.02-1.46)**		**1.12(1.06-1.19)**	**1.19(1.09-1.3)**	**1.19(1.07-1.33)**	**0.08584685**		
rs13211507	**6**	**0.02937**	**1.24(1.02-1.49)**		**1.15(1.08-1.23)**	**1.21(1.1-1.33)**	**1.2(0.99-1.44)**	**0.11024779**	**1.29(1.14-1.46)**	**1.28(1.13-1.45)**
rs13194781	**6**	**0.03445**	**1.23(1.02-1.49)**		**1.14(1.07-1.21)**					
rs3095089	**6**	**0.03484**	**1.19(1.01-1.39)**		**1.11(1.06-1.16)**					
rs3135353	**6**	**0.03605**	**1.20(1.01-1.43)**		**1.12(1.06-1.17)**					
rs3129939	**6**	**0.04396**	**1.18(1.00-1.38)**		**1.10(1.05-1.15)**					
rs9267522	6	0.05083	1.17(1.00-1.38)		1.11(1.06-1.16)					
rs2844659	6	0.06016	1.16(0.99-1.35)		1.10(1.05-1.15)					
rs259940	6	0.05119	1.15(1.00-1.33)		1.10(1.06-1.14)	1.13(1.07-1.21)	1.13(1.07-1.21)	0.00700774	1.17(1.08-1.27)	1.17(1.08-1.26)
rs3115663	6	0.05347	1.17(1.00-1.37)		1.11(1.07-1.16)					
rs3129763	6	0.06113	1.16(0.99-1.35)		1.11(1.06-1.15)					
rs3130618	6	0.06881	1.16(0.99-1.36)		1.11(1.07-1.16)					
rs1245371	6	0.07475	1.14(0.99-1.31)		1.10(1.06-1.14)				1.16(1.08-1.26)	1.16(1.08-1.26)
rs3893464	6	0.08414	1.12(0.98-1.28)		1.07(1.04-1.11)					
rs3806033	6	0.08796	1.12(0.98-1.28)		1.08(1.04-1.11)					
rs3129055	6	0.1033	1.13(0.98-1.30)		1.09(1.05-1.13)					
rs3117292	6	0.1231	0.90(0.78-1.03)		0.93(0.90-0.96)					
rs9393692	6	0.135	1.11(0.97-1.26)		1.08(1.04-1.12)					
rs2523554	6	0.1403	1.11(0.97-1.26)		1.08(1.05-1.12)				1.19(1.11-1.28)	1.19(1.10-1.28)
rs2256543	6	0.1409	1.10(0.97-1.26)		1.09(1.06-1.13)	1.14(1.08-1.21)	1.14(1.08-1.21)	0.00850992	1.17(1.09-1.26)	1.17(1.09-1.26)
rs2523946	6	0.146	0.91(0.79-1.04)		0.93(0.90-0.96)					
rs9295663	6	0.1898	1.10(0.95-1.28)		1.10(1.05-1.14)					

				Hung *et al.*, 2008*	Landi *et al.*, 2009[£]	Wang *et al.*, 2008[£]			McKay *et al.*, 2008	
SNP	**Chr**	**p-value**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI) fixed**	**OR(95% CI) random**	**UKGWA p-value**	**OR(95% CI) Model 1***	**OR(95% CI) Model 2[δ]**
rs9505900	6	0.2747	0.93(0.81-1.06)						1.17(1.09-1.25)	1.17(1.09-1.25)
rs6917441	6	0.5986	1.04(0.89-1.22)						1.17(1.08-1.27)	1.17(1.08-1.27)
rs1007475	6	0.6192	0.96(0.84-1.11)	1.22(1.11-1.34)						
rs3129054	6	0.6404	0.97(0.84-1.12)		0.92(0.89-0.96)					
rs429083	6	0.6504	0.97(0.85-1.11)		1.08(1.04-1.11)					
rs1545092	6	0.6781	0.97(0.85-1.11)			1.12(1.06-1.19)	1.12(1.06-1.19)	0.02920815	1.16(1.08-1.25)	1.15(1.07-1.24)
rs1925439	6	0.9332	1.01(0.81-1.25)		1.12(1.06-1.19)					
rs4286803	6	0.9425	1.00(0.87-1.14)						1.16(1.08-1.25)	1.17(1.09-1.26)
rs3117143	6			1.43(1.21-1.68)	1.17(1.10-1.25)				1.31(1.15-1.48)	1.30(1.15-1.47)
rs3117582	6				1.22(1.15-1.29)	1.3(1.19-1.42)	1.28(1.07-1.52)	6.24E-06	1.30(1.16-1.46)	1.29(1.15-1.45)
rs1270942	6				1.19(1.13-1.26)	1.24(1.13-1.35)	1.22(1.04-1.43)	0.00067162	1.28(1.14-1.43)	1.27(1.13-1.42)
rs9262143	6				1.19(1.13-1.26)	1.24(1.14-1.35)	1.23(1.1-1.38)	0.00158545	1.26(1.13-1.41)	1.25(1.12-1.40)
rs3130805	6								1.25(1.13-1.39)	1.23(1.11-1.37)
rs7452888	6								1.18(1.10-1.27)	1.18(1.10-1.27)
rs1150752	6				1.19(1.12-1.27)					
rs3132631	6				1.18(1.10-1.26)					
rs3132630	6				1.17(1.10-1.25)					
rs1233579	6			1.4(1.19-1.64)	1.16(1.09-1.23)	1.22(1.11-1.34)	1.19(0.96-1.48)	0.02981285		
rs3129962	6				1.15(1.08-1.23)					
rs9379494	6				1.15(1.08-1.23)					
rs2734583	6				1.15(1.08-1.22)					
rs3094061	6				1.14(1.07-1.22)					

				Hung *et al.*, 2008*	Landi *et al.*, 2009[£]	Wang *et al.*, 2008[£]			McKay *et al.*, 2008	
SNP	**Chr**	**p-value**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI) fixed**	**OR(95% CI) random**	**UKGWA p-value**	**OR(95% CI) Model 1***	**OR(95% CI) Model 2[δ]**
rs209181	6				1.11(1.06-1.17)					
rs4678	6				1.11(1.06-1.16)					
rs2734985	6				1.11(1.06-1.15)					
rs2844657	6				1.10(1.05-1.14)					
rs6457374	6				1.10(1.05-1.14)					
rs422331	6			1.21(1.11-1.33)		1.14(1.07-1.2)	1.12(1.02-1.24)	0.04563638		
rs4887077	**15**	**0.02482**	**1.16(1.02-1.33)**			**1.18(1.12-1.25)**	**1.18(1.12-1.25)**	**6.10E-05**	**1.20(1.12-1.29)**	**1.20(1.12-1.29)**
rs11638372	**15**	**0.02622**	**1.16(1.02-1.33)**		**1.19(1.15-1.23)**	**1.18(1.12-1.25)**	**1.18(1.12-1.25)**	**5.29E-05**	**1.20(1.12-1.29)**	**1.19(1.11-1.28)**
rs6495314	**15**	**0.03045**	**1.16(1.01-1.32)**		**1.19(1.15-1.24)**	**1.18(1.12-1.25)**	**1.18(1.12-1.25)**	**0.00014268**	**1.21(1.13-1.30)**	**1.21(1.13-1.30)**
rs8034191	**15**	**0.03494**	**1.16(1.01-1.33)**	**1.32(1.21-1.45)**	**1.29(1.25-1.34)**	**1.3(1.23-1.38)**	**1.3(1.23-1.38)**	**1.35E-06**	**1.34(1.25-1.44)**	**1.34(1.25-1.44)**
rs6495309	15	0.05308	0.85(0.72-1.00)		0.78(0.75-0.82)	0.77(0.72-0.82)	0.77(0.67-0.87)	1.80E-10	1.23(1.12-1.34)	1.23(1.13-1.34)
rs1051730	15	0.05376	1.14(1.00-1.31)	1.3(1.19-1.43)	1.31(1.27-1.36)	1.3(1.23-1.38)	1.3(1.23-1.38)	4.00E-07	1.35(1.25-1.45)	1.35(1.26-1.45)
rs13180	15	0.05922	0.88(0.76-1.01)		0.86(0.83-0.89)				1.20(1.12-1.29)	1.20(1.11-1.29)
rs4887053	15	0.0665	0.85(0.72-1.01)		0.84(0.80-0.88)	0.8(0.75-0.86)	0.8(0.72-0.9)	8.38E-08	1.19(1.09-1.29)	1.18(1.08-1.29)
rs1394371	15	0.0693	1.14(0.99-1.31)		1.20(1.16-1.25)	1.21(1.14-1.29)	1.21(1.14-1.29)	0.00013006	1.24(1.15-1.33)	
rs1002941	15	0.4946	0.95(0.82-1.10)	1.23(1.11-1.36)						
rs2036534	15				0.79(0.76-0.82)	0.77(0.71-0.82)	0.77(0.68-0.87)	1.92E-10	1.23(1.12-1.34)	1.23(1.12-1.34)
rs1996371	15					1.18(1.12-1.25)	1.18(1.12-1.25)	0.00014124	1.21(1.13-1.30)	1.21(1.13-1.30)
rs10519203				0.76(0.7-0.84)						
rs1317286				1.34(1.23-1.47)						
rs1504550				1.29(1.18-1.42)						
rs16969968				1.32(1.2-1.44)						

				Hung *et al.*, 2008*	Landi *et al.*, 2009[£]	Wang *et al.*, 2008[£]			McKay *et al.*, 2008	
SNP	**Chr**	**p-value**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI)**	**OR(95% CI) fixed**	**OR(95% CI) random**	**UKGWA p-value**	**OR(95% CI) Model 1***	**OR(95% CI) Model 2[δ]**
rs17405217				1.29(1.18-1.42)						
rs17483548				1.3(1.18-1.42)						
rs17483721				1.27(1.16-1.40)						
rs17483929				1.29(1.18-1.42)						
rs17484235				1.32(1.21-1.45)						
rs17484524				1.3(1.19-1.43)						
rs17486278				1.3(1.19-1.42)						
rs17487223				1.28(1.17-1.40)						
rs2009746				1.3(1.18-1.42)						
rs2036527				1.36(1.24-1.49)						
rs2568494				1.33(1.21-1.46)						
rs2656052				1.35(1.23-1.49)						
rs2656065				1.29(1.18-1.41)						
rs7180002				1.3(1.18-1.42)						
rs7181486				1.3(1.18-1.42)						
rs8031948				1.36(1.24-1.49)						
rs931794				0.77(0.7-0.84)						
rs951266				1.31(1.19-1.43)						
rs9788721				0.76(0.7-0.84)						

SNP significant in all studies, including LLP are highlighted in bold; *adjusted for age, sex and country; [£]unadjusted; [δ]adjusted for age, sex, country and eigen value

CHAPTER 4

GENOME WIDE SURVIVAL ANALYSIS

4.1 Aim

Lung cancer is the leading cause of death due to cancer in males and the second most prominent cause of death in females, worldwide, in 2008 [38]. The five-year overall survival for lung cancer is 16% in the USA and 7.8% and 9.1% in men and women, respectively in the UK[99]. In England and Wales, 20% of lung cancers are small cell (SCLC) while 80% are non-small cell (NSCLC) where squamous cell carcinoma represent 43.75%, 27% are adenocarcinomas and 10% are large cell carcinomas[250].

The aim of this study was to identify Single Nucleotide Polymorphisms (SNPs) associated with the survival of non-small cell lung cancer (NSCLC) patients and, separately, with the survival of patients with either early- or advanced-stage NSCLC. To this end, survival was analysed using Cox Proportional Hazard regression after controlling for the effect of age at diagnosis, cell type, stage and smoking pack years. A further aim was to investigate the cumulative effect of significant SNPs on patient survival. The study considered the overall survival and cause-specific survival (i.e. lung cancer-associated) of NSCLC patients separately.

4.2 Introduction

Survival studies have been carried out in various cancers to identify factors associated with shorter survival[251-255]. Multivariate Cox proportional hazard regression analysis identified that liver metastasis and total number of all chemotherapy cycles were significant in extensive small cell lung cancer patients[256]; non-curative resection and tumour location on the gallbladder neck were significant in gallbladder cancer patients with resection with

curative intent[257]; age, tumour stage and nodal status, number of lymph nodes retrieved, operative method, lymphovascular invasion, perineural invasion, postoperative chemotherapy, and preoperative serum CEA level >/= 2.4 ng/mL were independent predictors for 5-year overall survival in colorectal cancer[258]; pathological lymph node pN2 status was associated with overall survival in breast cancer (stage I to III) after treatment with surgery and adjuvant therapy[259]; and age, pathological stage and tumour size were significant in the overall survival analysis of gastric cancer patients[260]. Similar survival studies were also conducted in lung cancer using clinical factors (Table 4.1a).

These factors could be integrated into statistical models, for example to predict responsiveness to specific treatments, thereby improving patient outcome and clinical management[251, 253]. In addition to the clinical and epidemiological factors described above, it may be possible to identify inherited SNPs that are associated with post-diagnosis cancer survival, similarly to the identification of SNPs associated with cancer incidence[113, 114, 193, 203]. Such SNPs may enable physicians to devise treatment based on an individuals' requirement[251] and survival-GWAS studies may also shed some light on the limited knowledge available on various biological processes involved in survival of lung cancer patients.

4.2.1 Genome Wide Survival Analysis in Lung Cancer

Advances in genotyping technologies have permitted the detection of disease-associated loci on a genome wide scale[107, 202]. Most of the genome wide association studies (GWAS) published in lung cancer research have concentrated on disease incidence, identifying SNPs that increase lung cancer susceptibility rather than analysing SNPs associated with the

survival of diagnosed patients[113, 114, 193, 203]. The SNPs previously identified in lung cancer GWAS include rs402710 within gene *CLPTM1L* (5p15.33), rs2736100 within gene *TERT* (5p15.33) and rs1051730 within gene *CHRNA3* (15q25.1)[113, 114, 193, 203].

Of those studies which have addressed the association between SNP inheritance and lung cancer survival, many have been performed in the context of a therapy with the aim of identifying SNPs for subsequent patient stratification[251-255] (Table 1a). For instance, Hu *et al*. (2012)[252] identified survival-associated SNPs in Chinese patients with advanced stage NSCLC receiving first line platinum based chemotherapy; Sato *et al*. (2011)[251] identified SNPs associated with the survival of advanced NSCLC Japanese patients treated with Carboplatin and Paclitaxel; Lee *et al*. (2012)[255] identified polymorphisms in Korean advanced NSCLC patients receiving systemic chemotherapy; and Tan *et al*. (2011)[254] evaluated survival in both SCLC and NSCLC Caucasian cases receiving platinum-based chemotherapy. Other studies that have addressed survival in Caucasians include Wu *et al*. (2011)[253] and Huang *et al*. (2009)[261]. Wu *et al*. 2011[253] considered the survival of advanced stage NSCLC patients treated with first line platinum-based chemotherapy using germline variants while Huang *et al*. (2009)[261] used SNPs from tumour-derived DNA to evaluate survival in early stage NSCLC cases.

Limitations of the Huang *et al*. (2009) study[261] include the following: i) the datasets used in the discovery and validation phases were incomparable because the discovery dataset was made up of samples from different ethnicities while the ethnicity of samples from the validation cohort is not reported; ii) although study populations of the size used are not unfounded (for example, similar sized studies are common in comparative mRNA expression-GWAS studies), these sample sizes are still relatively modest (discovery set: n= 100; validation set: n = 89); iii) minimal confidence can be ascribed to the accuracy of the called genotypes, since the SNPs were assessed in tumour-derived DNA and lung tumours

are subject to frequent mutations and genomic instability[261]. In contrast, in the Wu *et al*. (2011) study[253], blood was extracted for genotyping after treatment of the patients with a chemotherapeutic agent (which will reduce the stability of the genomic markers) and the authors failed to correct for multiple comparisons, and therefore a high number of false-positive associations are expected within this study[253].

Although there are large differences in the design and study populations used in the above papers, it is notable that SNPs identified in one study have not yet been identified by any other (Table 4. 1a). SNPs identified so far lie within genes involved in tumour suppression, the initiation and regulation of translation, development, apoptosis, inflammation, adipogenesis, osteoblastosis, cell adhesion and regulation (Table 4.1b).

Table 4.1a: Publications on lung cancer survival

Author	Cases	Inclusion criteria	statistical model	Significant SNPs
Hu *et al*., 2012[252]	Discovery - (Chinese cohort 1 - 303; Chinese cohort 2 - 225; total - 528); replication 1 - 340 Chinese patients; replication 2 - 409 Caucasian patients	Aim - Identify SNPs influencing the overall survival of patients receiving platinum based chemotherapy. Inclusion criteria - Stage III/IV NSCLC treated with first line platinum based chemotherapy without surgery.	Multivariate Cox proportional hazard regression model adjusted for age, gender, histology, stage and smoking status. SNPs treated in the additive mode.	12 SNPs were identified in the meta analysis of the 2 Chinese discovery cohorts. The HRs presented are for SNPs that were significant in the pooled analysis of Chinese populations (3 cohorts) in the study. rs7629386 (HR=1.65; 95% CI: 1.30-2.09 ;p-value: 3.63×10^{-5}); rs969088 (HR=1.43; 95% CI: 1.24-1.66 ;p-value:1.75×10^{-6}); rs3850370 (HR=1.38; 95% CI: 1.19-1.60 ;p-value: 2.92×10^{-5}); rs41997 (HR=0.66; 95% CI: 0.56-0.78 ;p-value: 4.19×10^{-7}); rs12000445 (HR=0.67; 95% CI: 0.57-0.80 ;p-value: 7.12×10^{-6}).
Huang *et al*., 2009[261]	Discovery -100 patients from Massachusetts general hospital. Validation - 89 patients from National Institute of Occupational Health, Norway.	Aim - To identify SNPs in tumour tissue associated overall survival of early stage NSCLC cases. Inclusion criteria - Early stage (IA, IB, IIA and IIB) NSCLC patients.	SNPs were identified in an univariate model and validated in a multivariate model adjusted for age, sex, clinical stage (IA,IB,IIA and IIB),cell type(squamous cell carcinoma vs adenocarcinoma), smoking pack years and FDR	Univariate analysis identified 50 SNPs significant at the 2.5 × 10-4 significance levels. Pooled analysis of the two cohorts produced the following hazard ratio for the SNPs that were significant in the validation cohort. Significant SNPs include rs10176669(HR=2.40; 95% CI: 1.65-3.49 ;p-value: 1.74×10^{-6}); rs4438452(HR=2.25; 95% CI: 1.54 - 3.28; p-value: 9.97×10^{-6}); rs12446308 (HR=2.90; 95% CI: 1.93- 4.36 ;p-value:2.88×10^{-8}); rs13041757 (HR=2.04; 95% CI: 1.48 - 2.81; p-value:

				6.08 × 10^{-6}); rs10517215 (HR=2.44; 95% CI: 1.61 - 3.71; p-value: 2.45 × 10^{-5}).
Sato *et al.*, 2011[251]	Discovery only - 105 Japanese patients	Aim - Identify SNPs associated with OS in Japanese patients treated with Carboplatin and Paclitaxel. Inclusion criteria - Stage III/IV, no prior chemotherapy, surgery and/or radiotherapy, patient older than 20 years and Eastern Cooperative Oncology Group performance status between 0-2.	Univariate and multivariate Cox proportional hazard regression after adjusting for performance status and gender. Adjusted for Holm's correction for multiple testing.	The significant SNPs in the multivariate analysis were rs1656402 (p-value: 4.5 × 10^{-7}); rs1209950 (p-value: 6.5 × 10^{-5}); rs9981861 (p-value: 9.2 × 10^{-7})
Tan *et al.*, 2011[254]	Discovery - 1183 Caucasian patients (222 SCLC and 961 NSCLC)	Aim - To identify germline variants selected using lymphoblastic cell lines (LCL) that influence the overall survival of patients receiving platinum based chemotherapy. Inclusion criteria - Pathologically confirmed primary lung cancer cases treated with platinum based chemotherapy.	Multivariate Cox proportional regression model adjusted for disease stage was used to identify SNPs in 1183 lung cancer patients, which were initially identified by conducting a SNP versus cisplatin IC_{50} and SNP versus expression of 283 (91 African-Americans; 96 Caucasian-Americans and 96 Han Chinese-American unrelated subjects) lymphoblastoid cell lines (LCL).	168 SNPs were selected for genotyping in the lung cancer patients. The most significant SNPs were rs11169748 (HR= 1.75; 95% CI: 1.03-2.97; p value: 0.039) and rs2440915 (HR= 1.41; 95% CI: 1.08-1.83; p value: 0.012) in NSCLC. Of the 19 genes tested in a knockdown experiment, significant genes include DAPK3 and METTL6, whose expression level was correlated with rs11169748 and rs2440915, respectively.

Lee *et al*., 2012[255]	Discovery - 384 Korean NSCLC patients	Aim - Identify SNPs of prognostic significance in advanced NSCLC Korean patients treated with systemic chemotherapy. Inclusion criteria - Stage IV patients receiving systemic chemotherapy, without any prior therapy or surgery on whom follow up data is available were selected.	SNPs in the dominant or additive inheritance mode were tested using a Cox proportional regression model in a 1000 datasets generated by bootstrap resampling to evaluate the overall survival after adjusting for age, ECOG performance status, smoking history, histology type, number of metastatic sites at diagnosis, use of platinum-based chemotherapy and use of EGFR-TKIs. Significant SNPs were further evaluated in a subgroup analyses of patients treated with platinum based chemotherapy (n = 254), never smokers with adenocarcinoma histology with (n=178) or without EGFR-TKIs therapy (n=215).	17 SNPs were significant in the overall and sub group analysis. The most significant SNP in the overall survival analysis was rs1571228 in the dominant model [AG+GG to AA], (HR= 0.53; 95% CI: 0.42 - 0.67; p-value = 2.025×10^{-7}).
Niu *et al*., 2012[262]	Discovery - A total of 874 (76 SCLC and 798 NSCLC) lung cancer cases were analysed.	Aim - To identify SNPs associated with overall survival in lung cancer patients treated with taxanes, wherein the SNPs were discovered using a genome wide association analysis of 276 IC_{50} cytotoxicity values for taxanes using lymphoblastoid cell lines	Disease stage among age at diagnosis, gender, smoking status and treatment was the only variable selected via a backward regression to be included in the final multivariate Cox regression model.	153 candidate SNPs were genotyped in 874 lung cancer cases. 4 SNPs were significant in the NSCLC cases and 2 SNP were significant in the SCLC cases in the survival analysis. The most significant SNP in NSCLC was rs1106697 (HR: 1.237; p value: 0.007). None of the significant genes from NSCLC cases

		and 1.3 million SNPs.		were significant in the knockdown experiment.
Wu *et al*., 2011[253]	Discovery -213 MD Anderson Discovery Population; Validation 1 - 945 Mayo clinic validation population; Validation 2 - 420 PLATAX validation population.	Aim - To evaluate SNPs associated with decreased overall survival in advanced NSCLC patients receiving Platinum based chemotherapy. Inclusion criteria -White ever smoker, stage III/IV NSCLC without surgery, treated with first line platinum based chemotherapy with or without radiotherapy	SNPs were identified using dominant, recessive and additive model in a multivariate Cox proportional hazard regression after adjusting for age, sex, pack years, clinical stage(IIIA,IIIB[dry],IIIB[wet],IV) and pre-treatment performance status(0,1, or 2-4).Genetic model with the least P value was considered.	60 SNPs were selected for validation. Significant SNPs include rs1878022 $_{pooled}$ (HR=1.33; 95% CI: 1.19-1.48;p-value: 5.13×10^{-7}); rs10937823$_{MD\ Anderson\ +\ Mayo\ clinic(}$HR=1.82; 95% CI: 1.42-2.33 ;p-value: 1.73×10^{-6}).

Table 4.1b: Gene/ closest gene for SNPs identified by various publications of survival in NSCLC cases.

Marker	BP position	Gene	Cytogenetic position of gene	gene summary
rs7629386[a]	40966907	*RPS27P4*	3p22.1	-
rs969088[a]	26389262	*LOC100131678*	5p14.1	-
rs3850370[a]	78534906	*FRDAP*	14q24.3	-
rs41997[a]	117991895	*ANKRD7*	7q31	Reported in alcohol drinking behaviour[263].
rs12000445[a]	23426271	*SUMO2P2*	9p21.3	-
rs10176669[b]	169084859			Functions in cellular responses and p38 mitogen activated protein kinase signalling pathway. Inactivation increases cell apoptosis[264].
rs4438452[b]	169071348	*STK39*	2q24.3	
rs13041757[b]	45600280	*EYA2*	20q13.1	Involved in apoptosis and its upregulation promotes tumour growth[261].
rs10517215[b]	30774083	*PCDH7*	4p15	Encodes a protein involved in intercellular recognition and adhesion[261].
rs12446308[b]	6417933	*RBFOX1*	16p13.3	Involved in neurodegenerative diseases[265].
rs1571228[c]	18930222	*FAM154A*	9p22.1	Identified as a variant associated with height in a Korean population that may suggest a possible function in cell growth[255].
rs1106697[d]	155672944	*LOC100996445*	7	-
rs1656402[e]	233426526	*EIF4E2*	2q37.1	Plays a crucial role in the initiation and regulation of translation, and is upregulated in NSCLC[251].
rs1209950[e]	40173528	*ETS2*	21q22.2	Regulates development and apoptosis. The encoded protein is a proto-oncogene and regulates telomerase. [provided by RefSeq, Jan 2012][249]
rs9981861[e]	41415044	*DSCAM*	21q22.2	Encodes a Down syndrome cell adhesion molecule of the immunoglobulin family and its expression increased in small cell than NSCLC cases[251].

rs11169748[f]	51579171	*POU6F1*	12q13.13	Involved in the proliferation of ovarian adenocarcinoma (clear cell)[266].
rs2440915[f]	61673772	*CCDC6*	10q21	Encodes a protein that function as a tumour suppressor.[provided by RefSeq, Sep 2010][249]
rs1878022[g]	108699032	*CMKLR1*	12q24.1	Involved in inflammation, adipogenesis and osteoblastogenesis[253].
rs10937823[g]	7480422	*SORCS2*	4p16.1	Contains a domain named VPS10 that functions in intracellular trafficking and lysosomal processing[253].

a-Hu *et al*., 2012; b-Huang *et al*.,2009; c-Lee *et al*.,2012; d-Niu *et al*., 2012; e-Sato *et al*., 2011; f-Tan *et al*., 2011; g-Wu *et al*.,2011

4.3 Statistical Concepts Underlying Survival Analysis

Analysis of lifetime, survival time or failure time, defined as the time to the occurrence of an event of interest for individuals in a population, is an important concept in various fields, for example, to investigate the survival probability of diseased individuals or the warranty of a product[267]. In the case of the lung cancer study developed here, lifetime refers to the lifespan from a particular time point (i.e. from the time of diagnosis till death or censoring) [267]. The relevant statistics for an individual / study are the time scale, time origin and time to an event of interest (death or censoring)[267].

Consider T to be a non-negative variable representing the time to an event[267]. In terms of lung cancer this would be the survival time of a patient[267]. Depending on the study design, it would be the time from diagnosis till death or time till the end of the study or if the patient is lost to follow up, the date last seen[267].

For a continuous T, the probability density function is given by

$$\mathrm{F(t)} \;=\; \mathrm{Pr(T} \leq \mathrm{t)} \;=\; \int_0^t f(a)da \qquad [267]$$

and the probability of an individual surviving until time t is given by the survivor function

$$\mathrm{S(t)} \;=\; \mathrm{Pr(T} \geq \mathrm{t)} \;=\; \int_t^\infty f(a)da \qquad [267]$$

At any given time t, the rate of death or failure is given by the hazard function

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t+\Delta t \mid T \geq t)}{\Delta t}$$

$$= \frac{f(t)}{S(t)} \qquad [267]$$

In terms of lung cancer survival time, the hazard function (or the hazard rate) up to time t, is the probability of death or failure during the time interval $[t, t+\Delta t)$ and the cumulative hazard function is given by

$$\lambda(t) = \int_0^t h(a)da \qquad [267]$$

The above set of equations cover the various functions used in studying the survival analyses when the time variable is continuous[267].

4.4 Method for Analysing Survival Data

4.4.1 Censoring

Unlike other statistical datasets, survival data is rarely complete[267]. A major constraint is censoring, where it is not possible to obtain the exact survival times for every individual in

the study, although the fact that an individual survived beyond a certain time point may be documented[268].

An individual's survival data may be censored due to i) loss of the patient to follow up; ii) the patient dropping out of the study; or iii) it is end of the study and the patient has not observed the event[268, 269]. Three types of censoring include left censoring, right censoring and interval censoring[269]. Left censoring occurs when the event of interest is already observed before recruiting the individual into the study[269]. Right censoring occurs when the exact event time is not known but is only known to exceed or be equal a certain time point[268]. Interval censoring is when the exact time of the event is not known but it is known to occur within a certain interval[269].

4.4.2 Semi Parametric Models

4.4.2.1 Kaplan Meier Method

The Kaplan Meier method is an empirical method for survival analysis that may be applied when there are varying survival times and not all individuals in the study experience the event of interest[270, 271]. For a given individuals, the method requires the survival time, the status of the individual at the end of the study period (or at censoring) and the observational groups to which they belong[270]. Assumptions of the method include that the likely survival time for any individual in the study is the same whether they are censored or not, whether recruited early or late during the study and whether the event of interest occurred on the date detected[271].

The Kaplan Meier curve is plotted with the survival time on the X axis and the cumulative probability of survival on the Y axis[270]. It displays the proportion of participants experiencing an event over the course of study[272]. The estimated cumulative survival and the Greenwood standard error is given by

$$\hat{S}(t) = \prod_{t_i<t}(1-\frac{d_i}{n_i})$$

$$SE[\hat{S}(t)] = \sqrt{[\hat{S}(t)^2]\sum\frac{d_i}{n_i(n_i-d_i)}}$$

$\hat{S}$ is the survival at time t and, d_i and n_i are the number of events (failures) and number of individuals at risk, at time t_i[273].

4.4.2.2 Cox Proportional Hazard Model

The Cox Proportional Hazard model (also in Chapter 2) is given by

$$\lambda(t) = \lambda_o(t)\exp(\sum\beta x) \quad [175]$$

$\lambda(t)$ is the event rate at time t expressed as the function of risk variables, $\lambda_0(t)$ is the event rate at baseline; i.e. measurements at the beginning of the study and $\exp(\sum\beta x)$ is the proportionality constant indicator for the specified risk factors[175]. This model is semi parametric as $\lambda_o(t)$ is unspecified and is used widely as the effect can be estimated without the knowledge of $\lambda_o(t)$ [175].

The Cox model survival function is given by

$$S(t) = S_o(t)^{\exp \sum_{n=1}^{p} \beta_n x_n} \quad [175]$$

The hazard ratio (HR) is given as the ratio of hazard for one individual to the hazard for another individual[175].

$$\text{HR} = \frac{\lambda(t)^*}{\lambda(t)}$$

$$= \frac{\hat{\lambda}_o(t) \exp(\sum \beta x^*)}{\hat{\lambda}_o(t) \exp(\sum \beta x)}$$

$$= e^{\sum_{k=1}^{p} \beta_k (x_k^* - x_k)} \quad [175]$$

4.4.2.2.1 Proportionality Hazard Assumption

An important assumption of the Cox model is that of 'proportional hazards', i.e. that the hazard for one individual is proportional to the hazard for another individual[175]. For time independent covariates, the relative hazard for two individuals, i and j, is given by

$$\frac{\lambda_o(t) \exp(\beta x_i)}{\lambda_o(t) \exp(\beta x_j)} = \frac{\exp(\beta x_i)}{\exp(\beta x_j)} \quad [274]$$

while for time dependent variable, the relationship is given by

$$\frac{\exp(\beta x_i(t))}{\exp(\beta x_j(t))} \quad [274]$$

Proportionality hazard assumptions can be tested by visual inspection of the survival curve for time independent variable while for time dependent variable Schoenfeld residuals can be used[274].

The proportionality hazard assumption test using Schoenfeld residuals provides a test statistics (measurement) making it the most feasible test, with larger global statistics depicting non-proportionality[274].

4.5 Materials and Methods

One hundred and eighty five NSCLC cases from Liverpool were identified. Blood DNA was extracted using Qiagen kits and genotyped using the 300K HumanHap Illumina bead chip array. The genotype data were quality controlled to include single nucleotide polymorphisms (SNPs) with i) a minor allele frequency >1%; ii) a genotypic call rate of > 95%; iii) to exclude SNPs with a Hardy Weinberg equilibrium of $p < 0.001$ and iv) to identify population outliers. The data was also checked to remove any duplicates, related individuals and individuals with a gender discrepancy between SNP calls and epidemiological data. Every individual had a genotype call rate of > 95%.

The survival status for each case was determined using the ONS (Office for National Statistics) registry data, the last ONS update being in February 2012. Cause specific death was identified if the cause of death was reported with ICD-10 codes "C34" ('Malignant neoplasm of lung and bronchus') or "C780 "('Secondary malignant neoplasm of lung') while the survival status for overall analysis was death due to any cause. The survival time was calculated using the date of diagnosis and either the date of death or the date last reported alive.

SNPs were coded in an additive mode (0, 1, 2) referring to the number of minor alleles carried by an individual[261]. Significant SNPs were identified using Cox proportional hazard

analysis to evaluate cause-specific and overall survival for all, early and advanced stage NSCLC cases. The regression analysis was adjusted for age at diagnosis, histological type (adenocarcinoma or squamous cell carcinoma), smoking pack years, sex and stage (as ordered variable): I, II, III and IV for all, IA, IB, IIA and IIB for early and IIIA, IIIB and IV for advanced stage cases.

Kaplan Meier curves were plotted for significant SNPs ($p \leq 10^{-6}$) and the difference between allele groups were tested using the log rank test. The proportional hazard assumption was assessed using the Schoenfeld residual (log transformed) for each of the models. The cumulative effect of the alleles ($p \leq 10^{-6}$) associated with shorter survival was tested for all, early and advanced stage NSCLC cases for both cause specific and overall survival. All quality controls were conducted in PLINK[227] and analyses were carried out using the "survival"[274, 275] package in R[229]. Manhattan and Kaplan Meier plots, and Schoenfeld residuals were obtained using packages such as "calibrate"[230] and "survival"[274,275], respectively, in R[229].

4.6 Results

Survival information and genotype data were available for 185 individuals. The association between genotype and survival after lung cancer diagnosis was addressed within this group. As such, overall and cause-specific (ie, lung cancer-associated) survival analysis was performed for the 185 NSCLC cases, and then the same analyses were performed for early stage (stages I and II, N = 107) and advanced stage (stages III and IV, N = 78) cases, separately (tumour staging was defined at diagnosis). Characteristics of the study population are shown in Table 2.2. There are fewer female patients with advanced stage

cancer but otherwise the population characteristics are comparable between early and advanced stage cases.

Table 4.2: Population characteristics of NSCLC cases

Population Characteristics	All NSCLC cases	Early stage cases	Advanced stage cases
Total patients	185	107	78
Sex, No. (%)			
Male	109 (58.92)	55 (51.40)	54 (69.23)
Female	76 (41.08)	52 (48.60)	24 (30.77)
Age at diagnosis, mean (standard deviation)	67.08 (8.19)	67.06 (8.10)	67.12 (8.37)
Smoking pack years, mean (standard deviation)	41.59 (21.55)	38.97 (20.55)	45.19 (22.48)
Clinical stage, No (%)			
Stage IA		26 (24.30)	
Stage IB	77 (41.62)	51 (47.66)	
Stage IIA		3 (2.80)	
Stage IIB	30 (16.22)	27 (25.23)	
Stage IIIA			28 (35.90)
Stage IIIB	51 (27.57)		23 (29.49)
Stage IV	27 (14.59)		27 (34.62)
Histological type			
Adenocarcinoma	88	54	34
Squamous cell carcinoma	97	53	44

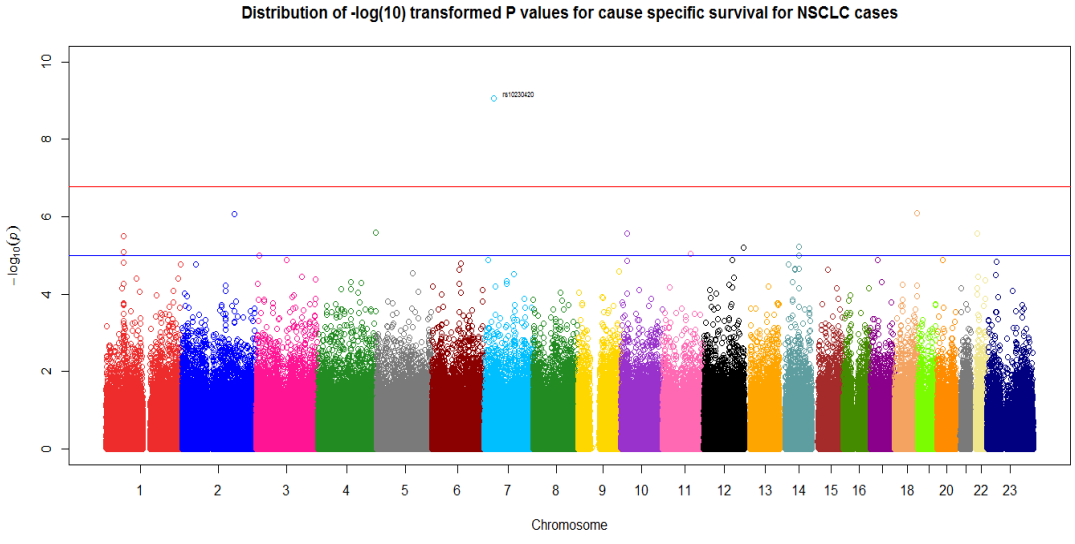
The median survival time for the lung cancer cases is shown in Table 4.3 (including stratification for early and advanced stage cases). A higher proportion of the advanced-stage cases died within the study than did early-stage cases (93.59% versus 65.42%) and the median survival time for advanced stage cases is correspondingly lower (11.7 versus 41.2).

Table 4.3: Median survival time distribution in the NSCLC cases

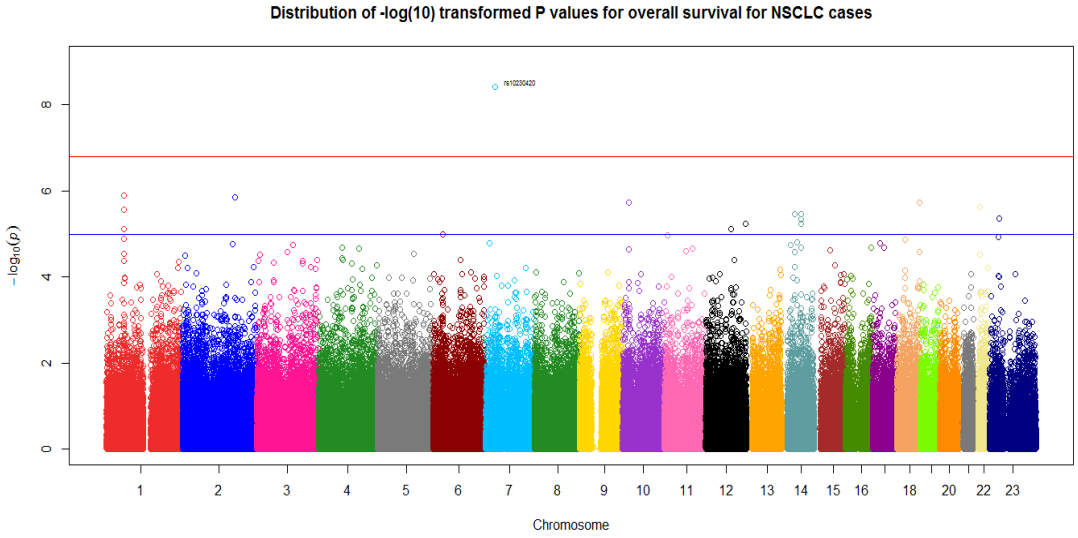
Population	Total cases	Events	Median survival time (95% CI)
Cause specific survival			
All	185	129	27.1 (20.2-37.0)
Early	107	58	57 (38.8-89.8)
Advanced	78	71	11.7 (10.3-18.9)
Overall survival			
All	185	143	25.2 (19.5-34.8)
Early	107	70	41.2 (37.0-65.2)
Advanced	78	73	11.7 (10.3-18.9)

Figure 4.1: Manhattan plots for allelic association with cause-specific (panels a, c and e) and overall (panels b, d and f) survival in the NSCLC cases. The plotted portion of each SNP corresponds to the genomic location and negative log of the observed p-value. The red and blue lines correspond to the Bonferroni correction and p = 10^{-5} levels, respectively. All NSCLC cases (panels a-b), early-stage cases (panels c-d) and advanced-stage cases (panels e-f) are shown.

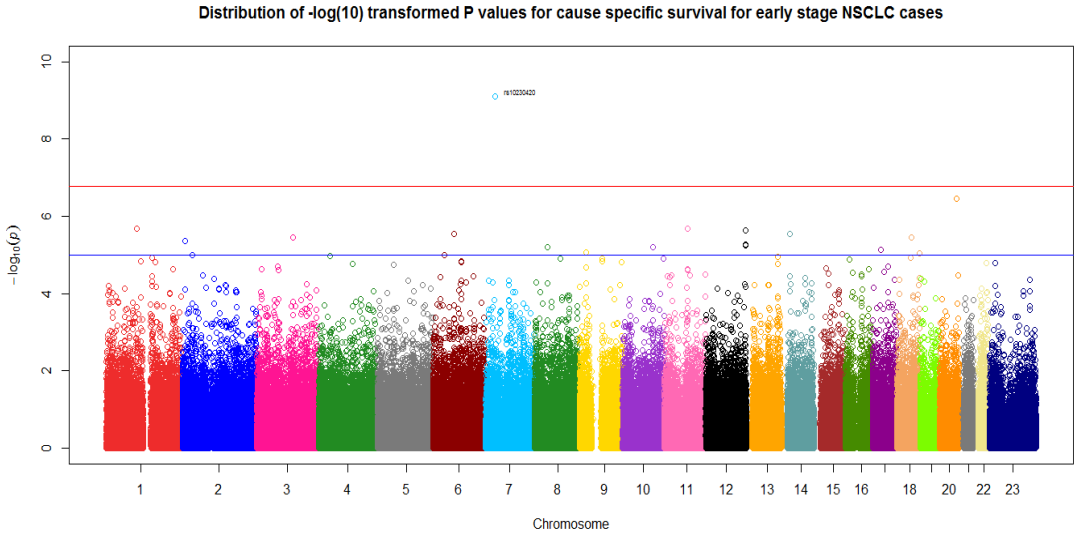
a)



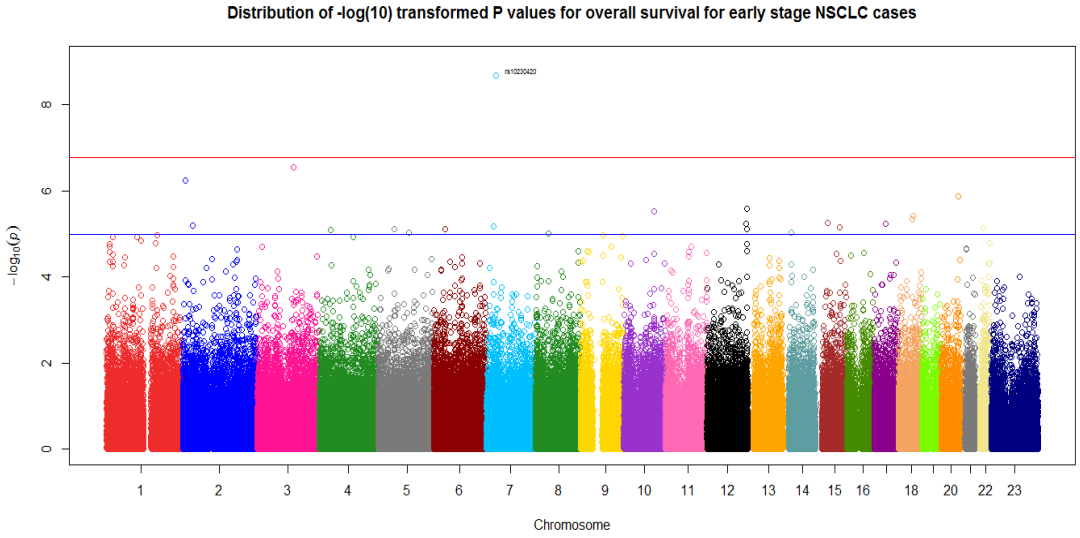
b)


Distribution of -log(10) transformed P values for overall survival for NSCLC cases
rs10230420
$-\log_{10}(p)$
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
23
Chromosome

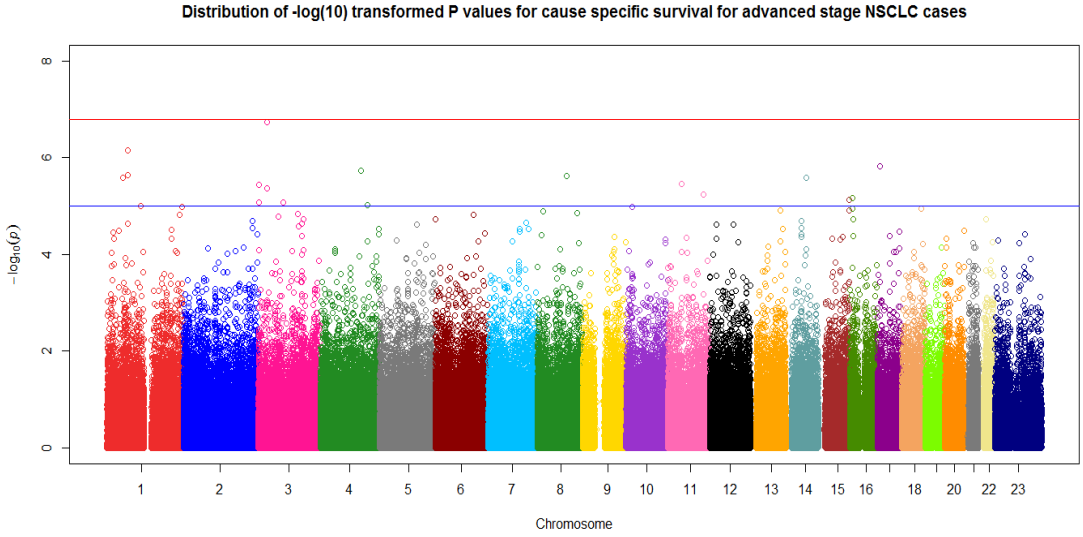
c)


Distribution of -log(10) transformed P values for cause specific survival for early stage NSCLC cases
rs10230420
$-\log_{10}(p)$
0
2
4
6
8
10
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
23
Chromosome

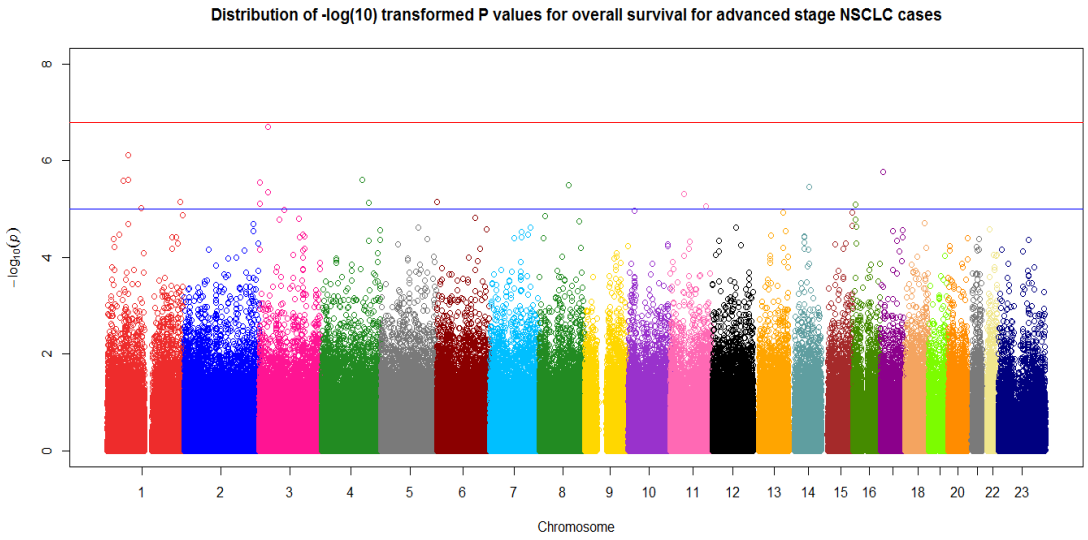
d)


Distribution of -log(10) transformed P values for overall survival for early stage NSCLC cases
rs10230420
$-\log_{10}(p)$
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
23
Chromosome

e)


Distribution of -log(10) transformed P values for cause specific survival for advanced stage NSCLC cases
$-\log_{10}(p)$
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
23
Chromosome

f)



The 185 cases were tested for 307002 SNPs that passed quality control in a Cox proportional hazard regression model after adjusting for age at diagnosis, sex, smoking pack years, stage (I, II, III, IV) and histological type (adenocarcinoma or squamous cell carcinoma). For the early stage cases, 306097 SNPs that passed the quality control criteria were tested using the Cox proportional hazard regression model after adjusting for age at diagnosis, sex, smoking pack years, stage (IA, IIA, IB, IIB) and histological types while for the advanced stage cases 302703 SNPs that passed the quality control criteria using the Cox proportional hazard regression model after adjusting for age at diagnosis, sex, smoking pack years, stage (IIIA, IIIB, IV) and histological type. P-values for the association with survival in each analysis are depicted in Figure 4.1 and the results for cause-specific and overall survival are described in the following two subsections.

4.6.1 Cause Specific Survival

Figure 4.2, Figure 4.3 and Figure 4.4 depict survival curves for patients with the three possible genotypes for SNPs found to be significant ($p < 10^{-6}$) in all, early and advanced stage cause-specific survival analysis. The KM estimator for right censoring computes an estimated survival function, the jump in value corresponds to the cumulative survival probability observed at that particular time (y-axis) and the markings on the survival curves represent censoring times [276]. The survival curves are supplemented with the Log-rank test p values that indicate whether the survival curves for the allele groups are different [276]. In the cause specific survival (Table 4.4) analysis, rs10230420 was significant in all NSCLC cases and early stage NSCLC cases while for the advanced stage cases no SNPs were significant at the Bonferroni correction level. The Bonferroni correction level for all, early and advanced NSCLC cases were 1.629E-07, 1.633E-07 and 1.652E-07, respectively. Since this correction level is very conservative and could lead to the loss of important survival associated SNPs, a cut off level of $p \leq 10^{-6}$ was chosen. Three SNPs for all NSCLC and early stage NSCLC cases while 2 SNPs for advanced stage NSCLC cases were significant at the $p \leq 10^{-6}$ level.

The multivariate Cox proportional hazard regression model that discovered significant SNPs ($p \leq 10^{-6}$) rs10230420, rs9949512 and rs2139133 in all NSCLC cases, rs3746619 and rs3827103 in early NSCLC cases and rs1868110 and rs2206779 in advanced NSCLC cases were significant when tested for proportionality hazard assumption seen by the global p value (Table 4.4). Cox proportional hazard being a semi parametric test, requires the testing of the proportionality hazard assumption which is obtained using the Schoenfeld residuals in R. The proportionality hazard assumption is met if the p-value produced for the model was not significant ($p > 0.05$).

SNPs rs3746619 and rs3827103 have identical survival curves and are located on the same chromosomal arm. They have the same minor allele frequencies and produced the same hazard ratios in the cause specific analysis of the early stage NSCLC cases. Since they are located in close proximity and have produced same survival hazard ratios they could be in linkage disequilibrium (Table 4.4). When allelic association with cause-specific survival was assessed in all of the NSCLC cases, 84 SNPs were found to be significant at the $p \leq 10^{-4}$ level. The corresponding analysis identified 153 and 128 SNPs when restricted to early- and late-stage cases, respectively.

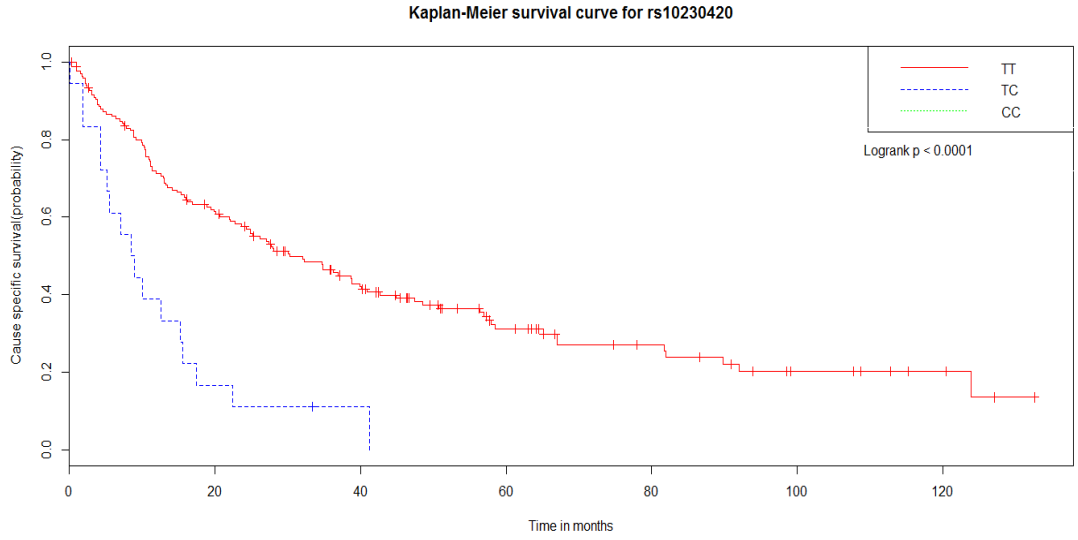
Table 4.4: SNPs significant in the cause specific survival analysis at $p \leq 10^{-6}$. Non-significant p-value for Schoenfeld residual indicate fulfilment proportionality hazard assumption

SNP	Genotype (alive/dead due to lung cancer)	MAF	HR(95% CI)	p-value[$]	p-value[*]
All NSCLC					
rs10230420	**TT (55/112); TC (1/17); CC (-/-)**	**0.05**	**6.2 (3.46-11.11)**	**0.0745**	**8.80E-10**
rs9949512	**CC (19/68); CT (28/51); TT (9/10)**	**0.28**	**1.85 (1.45-2.37)**	**0.1929**	**7.91E-07**
rs2139133	**TT (34/50); TC (17/57); CC (5/22)**	**0.35**	**1.87 (1.46-2.41)**	**0.3235**	**8.54E-07**
Early stage NSCLC cases					
rs10230420	TT (48/45); TC (1/13); CC (-/-)	0.07	10.06 (4.82-21)	0.0310	7.74E-10
rs3746619	**AA (47/48); AC (2/10); CC (-/-)**	**0.06**	**8.62 (3.77-19.75)**	**0.0521**	**3.46E-07**
rs3827103	**AA (47/48); AG (2/10); GG (-/-)**	**0.06**	**8.62 (3.77-19.75)**	**0.0521**	**3.46E-07**
Advanced stage NSCLC cases					
rs1868110	**GG (6/46); GT (1/21); TT (0/3)**	**0.18**	**3.36 (2.13-5.31)**	**0.3103**	**1.84E-07**
rs2206779	**TT (6/49); TC (1/15); CC (0/7)**	**0.19**	**3.09 (1.98-4.82)**	**0.4256**	**7.06E-07**

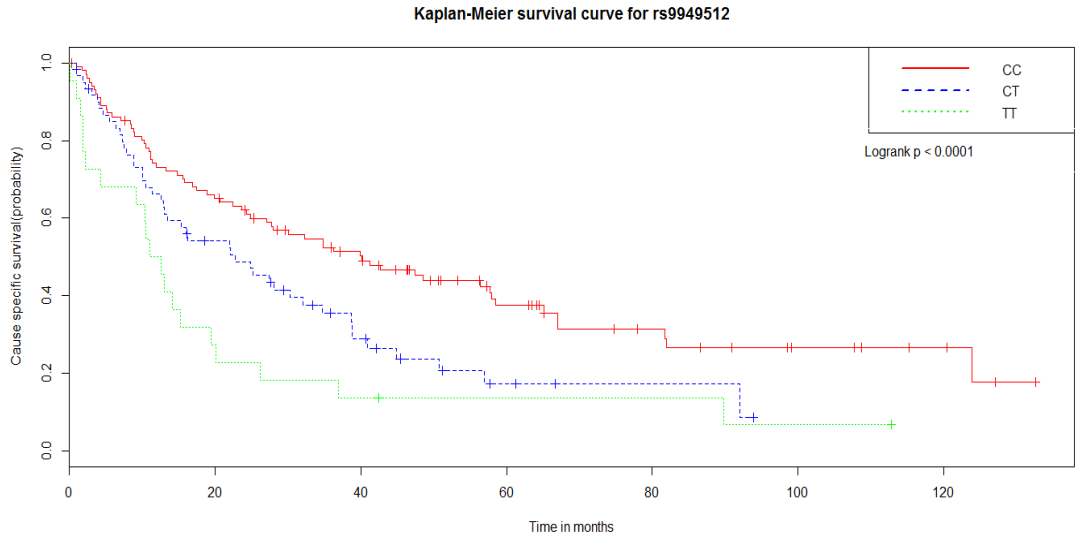
*Cox proportional hazard model after adjusting for age, sex, smoking pack years, stage and histological type. [$] Schoenfeld residual p-value for the Cox proportional hazard model. Bold entries depict proportional hazard assumption satisfaction.

Figure 4.2: Kaplan-Meier plots for SNPs that were significant at the $p < 10^{-6}$ level in cause-specific survival analysis for all NSCLC cases. Major-allele homozygotes and heterozygous individuals are shown in red and blue, respectively; minor-allele homozygotes are shown in green, where possible. Vertical ticks on survival curves denote censoring while differences between survival curves is tested using log rank test.

a)



b)



c)

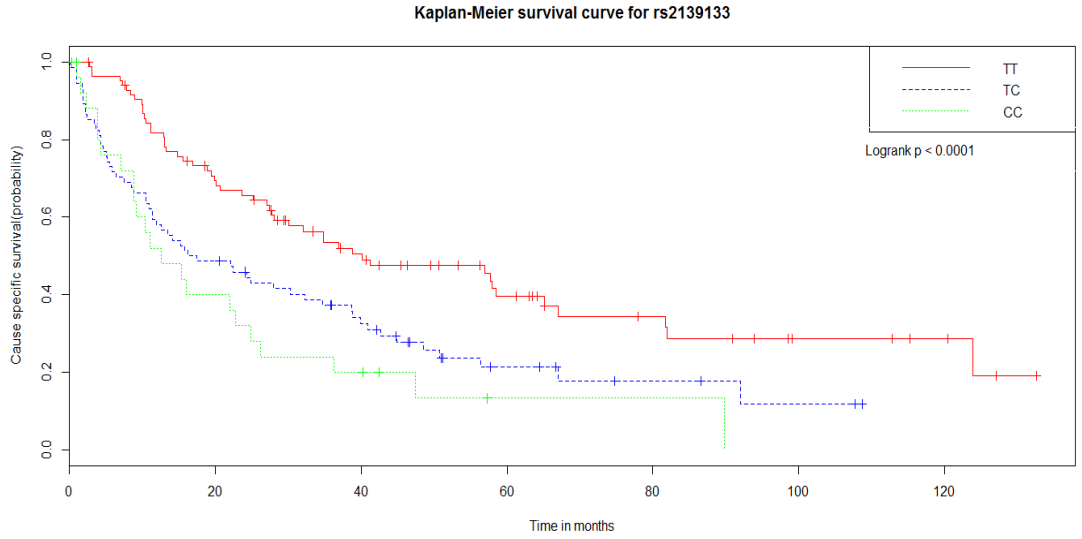
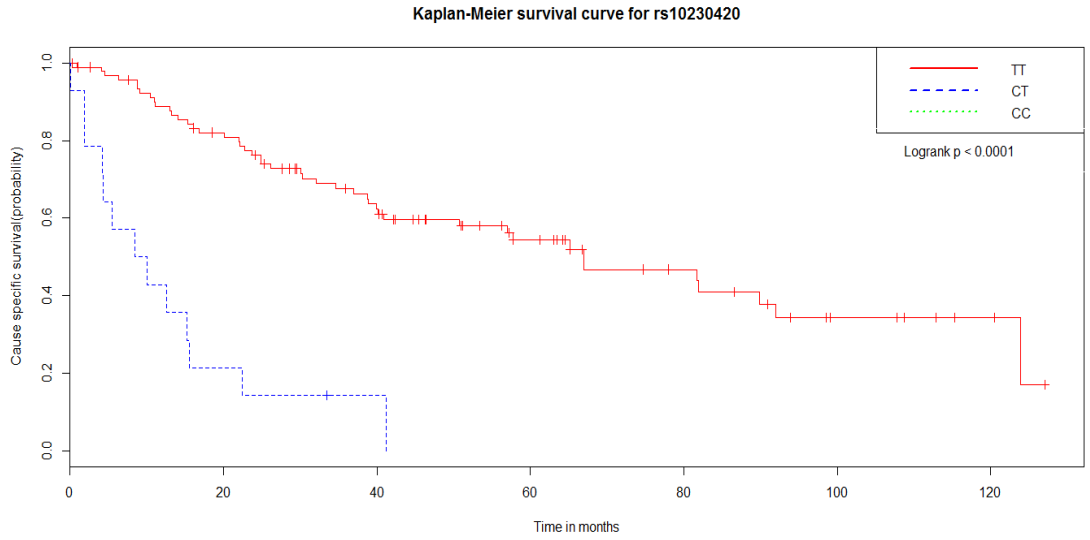
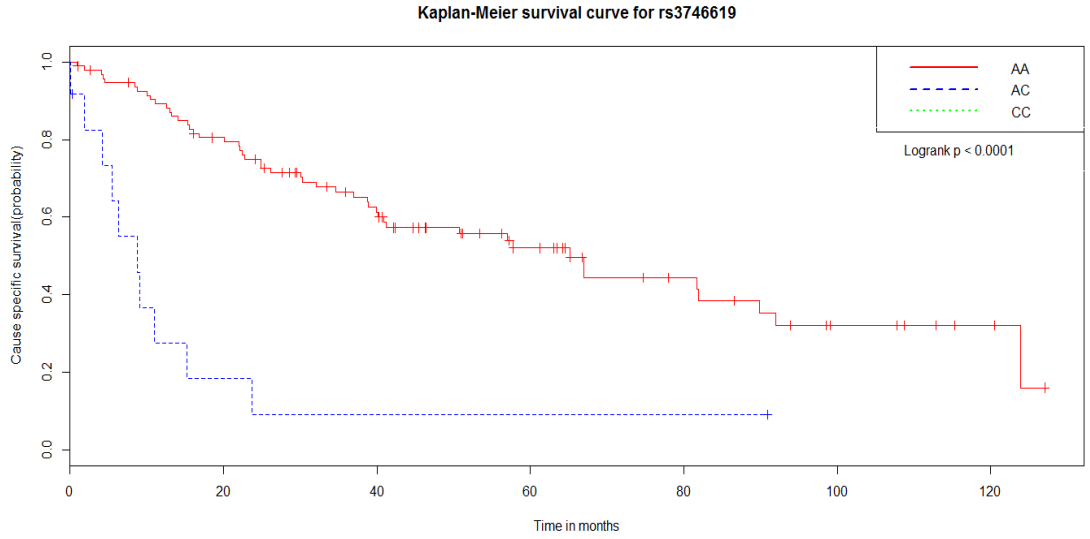


Figure 4.3: Kaplan-Meier plots for SNPs that were significant at the $p < 10^{-6}$ level in cause-specific survival analysis for early NSCLC cases (refer to Fig 4.2's legend).

a)



b)


Kaplan-Meier survival curve for rs3746619
AA
AC
CC
Logrank p < 0.0001
Cause specific survival(probability)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Time in months

c)

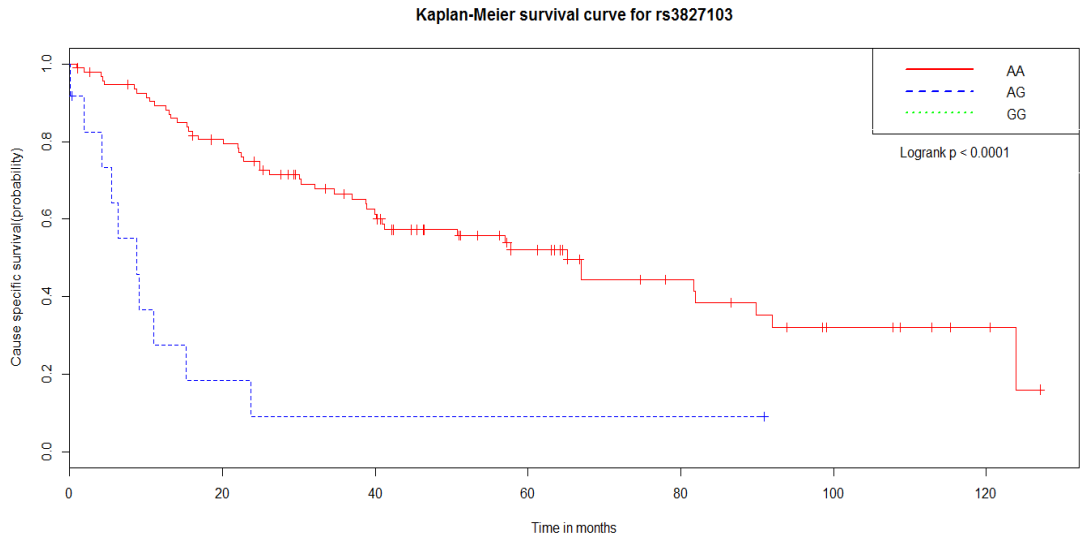
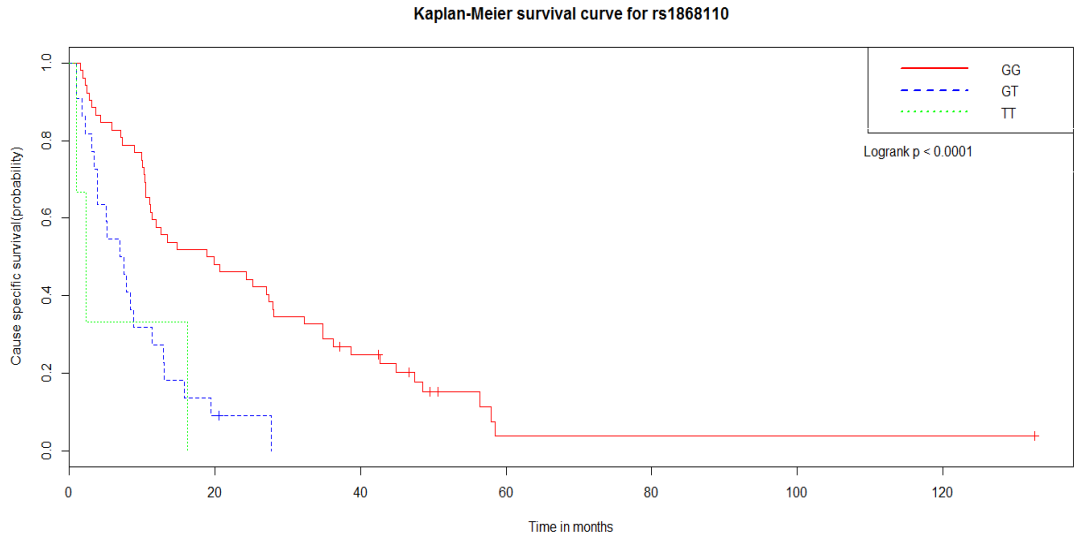
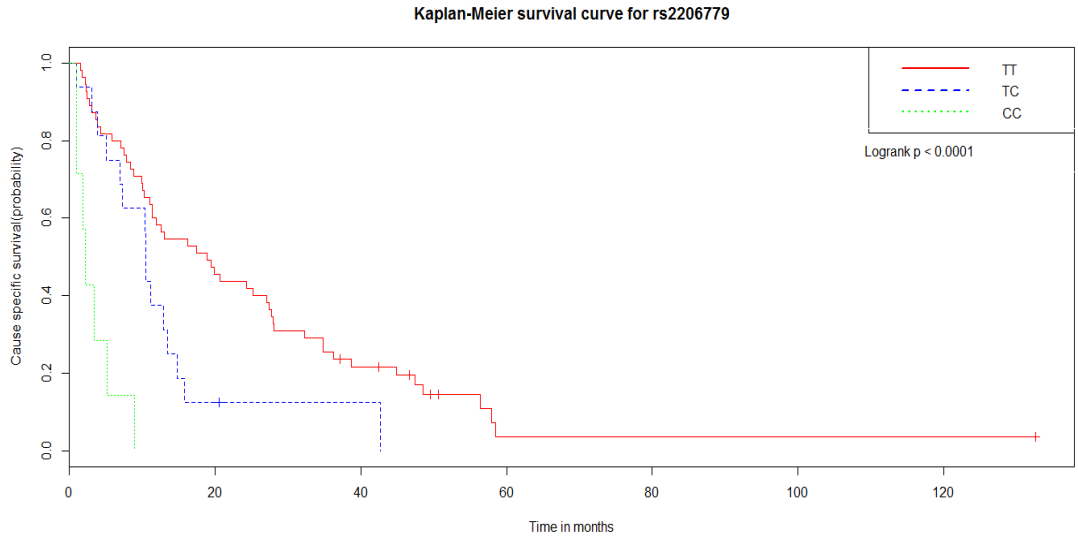

Kaplan-Meier survival curve for rs3827103
AA
AG
GG
Logrank p < 0.0001
Cause specific survival(probability)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Time in months

Figure 4.4: Kaplan-Meier plots for SNPs that were significant at the $p < 10^{-6}$ level in cause-specific survival analysis for advanced NSCLC cases (refer to Fig 4.2's legend).

a)



b)



4.6.2 Overall Survival

Figure 4.5, Figure 4.6 and Figure 4.7 depict the SNPs significant in all, early and advanced stages overall survival analysis. In the overall survival analyses (Table4. 5), rs10230420 was identified significant at the Bonferroni correction level in all NSCLC cases and early stage NSCLC cases, while for the advanced stage NSCLC cases no SNPs were significant at the Bonferroni correction level. The figures of KM survival curves for allele categories of SNPs also have the log rank p value printed [276]. The vertical ticks indicate censoring on the survival curves plotted with the survival probability on the y axis and time on the x axis [276]. The Bonferroni correction level for all, early and advanced NSCLC cases were 1.629E-07, 1.633E-07 and 1.652E-07, respectively. Since this correction level is very conservative and could lead to the loss of important survival associated SNPs, a cut off level of $p \leq 10^{-6}$ was chosen. One SNP for all NSCLC, 3 SNPs for early stage NSCLC cases while 2 SNPs for advanced stage NSCLC cases were significant at the $p \leq 10^{-6}$ level.

Multivariate Cox proportional hazard model that identified SNPs ($p \leq 10^{-6}$) rs1868110 and rs2206779 in advanced stage NSCLC cases were significant when tested for proportionality hazard assumption depicted by the non-significant " global" p-value while all SNPs ($p \leq 10^{-6}$) depicted non-proportionality in the total and early stage NSCLC cases (Table 4.5).

For the total NSCLC cases, 90 SNPs for the overall survival analysis were significant at $p \leq 10^{-4}$ level while for the early stage and advanced stage NSCLC cases, 118 SNPs and 125 SNPs were significant at $p \leq 10^{-4}$.

Table 4.5: SNPs significant in the overall survival analysis at $p \leq 10^{-6}$. Non-significant p-value for Schoenfeld residual indicate fulfilment proportionality hazard assumption.

SNP	Genotype (alive/dead)	MAF	HR(95% CI)	p-value[$]	p-value[*]
All NSCLC					
rs10230420	TT (41/126); CT (1/17); CC (-/-)	0.05	5.65 (3.18-10.05)	0.0104	3.76E-09
Early stage NSCLC cases					
rs10230420	TT (36/57); CT (1/13); CC (-/-)	0.07	8.93 (4.36-18.28)	0.0038	2.07E-09
rs2056533	CC (37/61); TC (0/9); TT (-/-)	0.04	9.03 (3.9-20.92)	0.0249	2.84E-07
rs6708630	TT (33/53); CT (4/15); CC (0/2)	0.11	4.02 (2.33-6.95)	0.0030	5.73E-07
Advanced stage NSCLC cases					
rs1868110	**GG (4/48); TG (1/21); TT (0/3)**	**0.18**	**3.35 (2.12-5.29)**	**0.3233**	**2.01E-07**
rs2206779	**TT (4/51); CT (1/15); CC (0/7)**	**0.19**	**3.08 (1.97-4.81)**	**0.4768**	**7.71E-07**

[*] Cox proportional hazard model after adjusting for age, sex, smoking pack years, stage and histological type. [$] Schoenfeld residual p-value for the Cox proportional hazard model. Bold entries depict proportional hazard assumption satisfaction.

Figure 4.5: Kaplan-Meier plots for SNPs that were significant at the $p < 10^{-6}$ level in overall survival analysis for all NSCLC cases (refer to Fig 4.2's legend).

a)

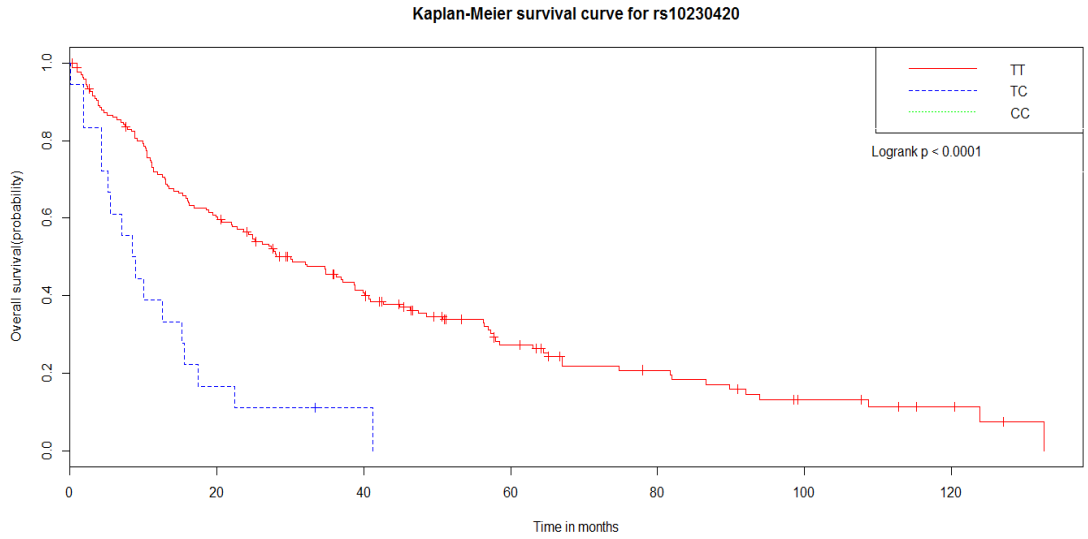
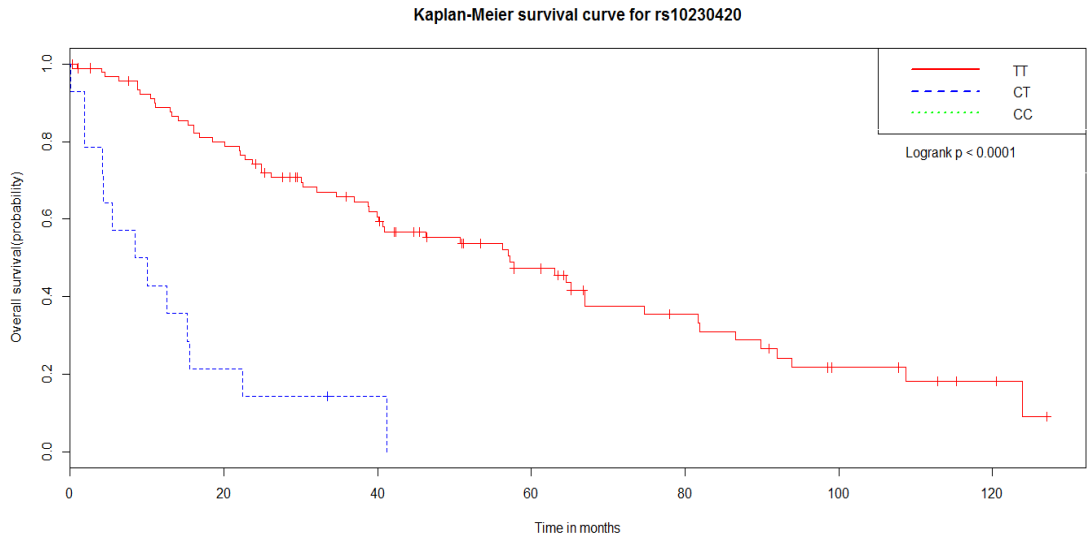
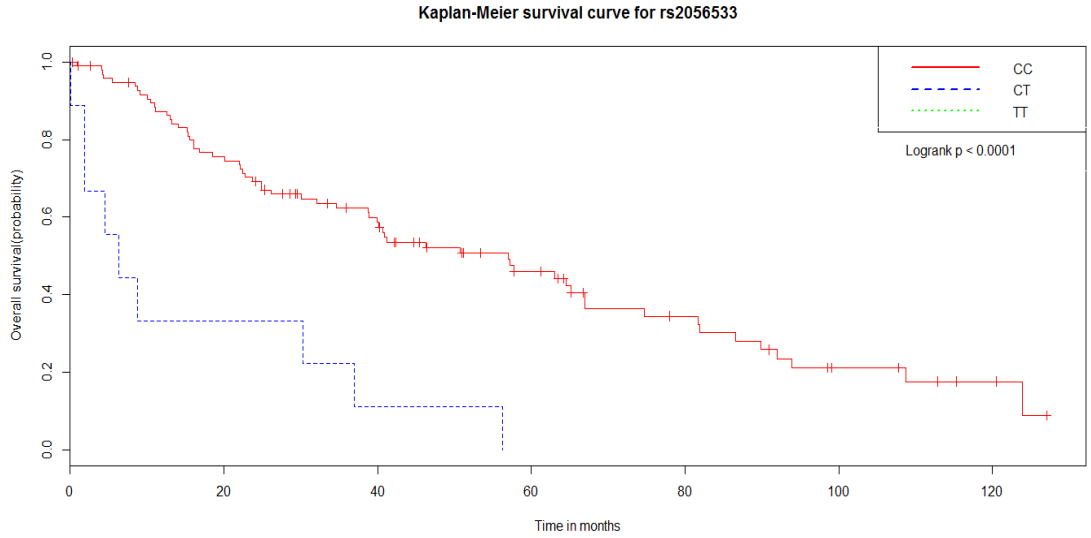


Figure 4.6 Kaplan-Meier plots for SNPs that were significant at the $p < 10^{-6}$ level in overall survival analysis for early NSCLC cases (refer to Fig 4.2's legend).

a)



b)


Kaplan-Meier survival curve for rs2056533
CC
CT
TT
Logrank p < 0.0001
Overall survival(probability)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Time in months

c)

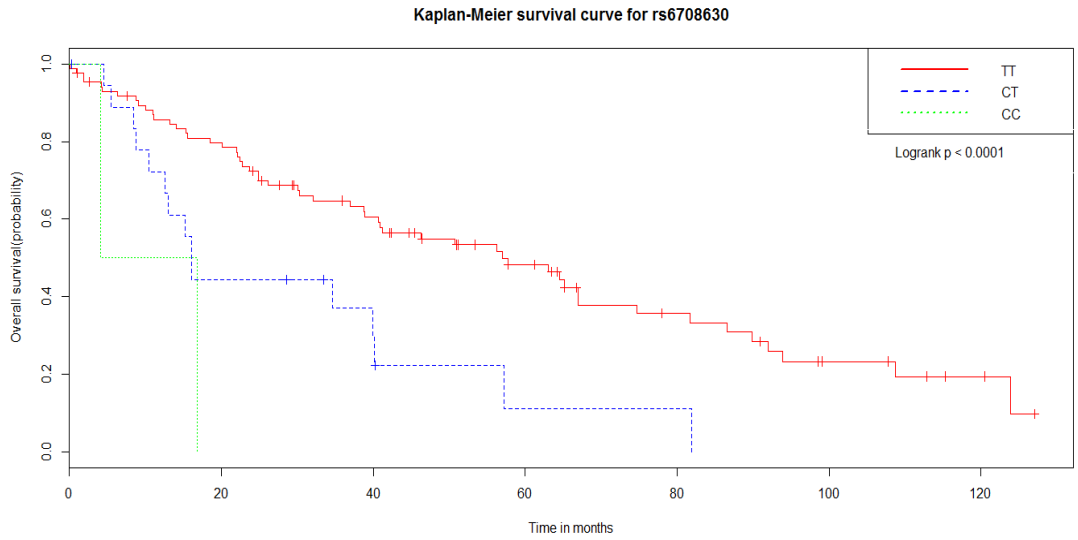
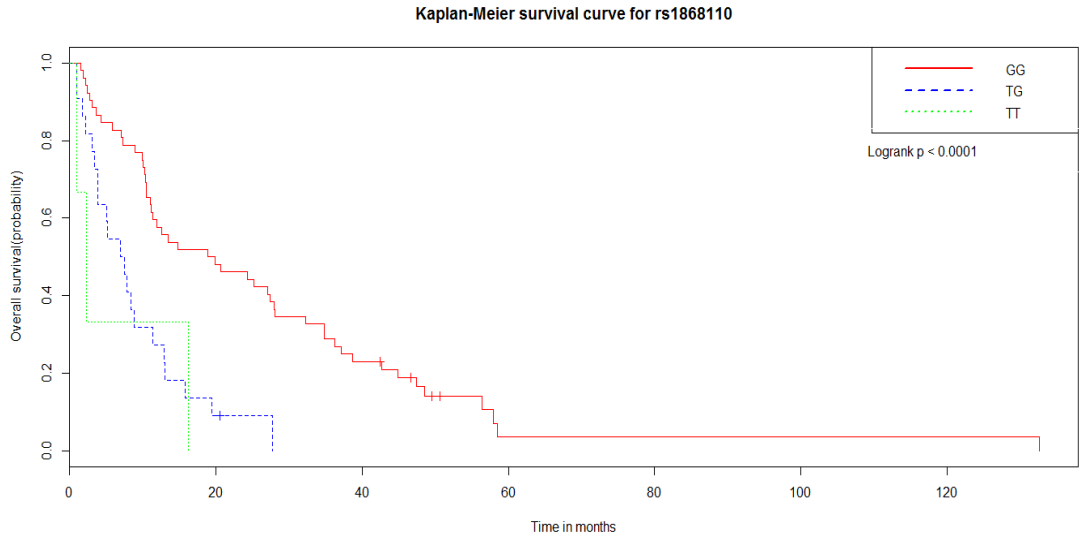
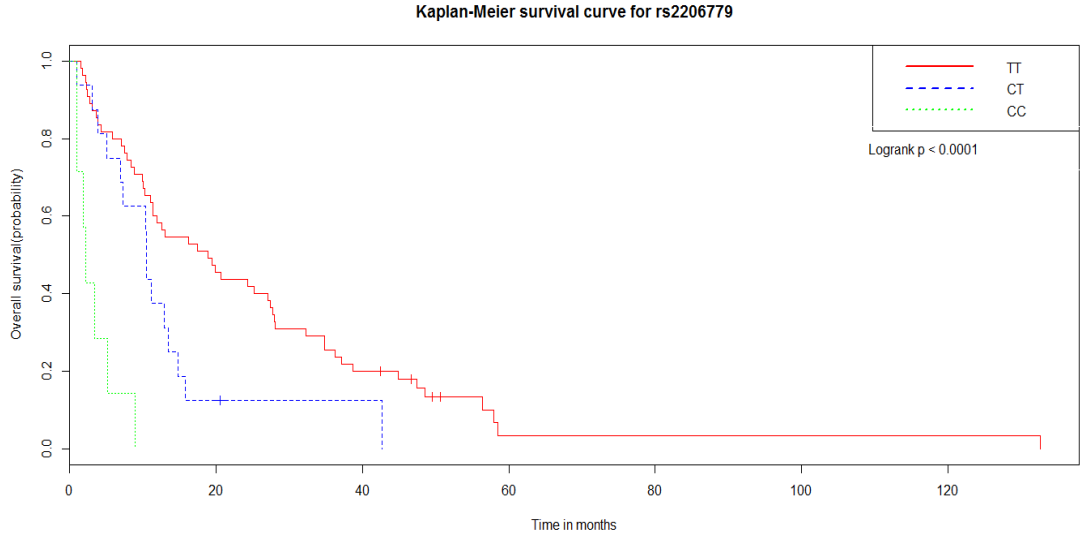

Kaplan-Meier survival curve for rs6708630
TT
CT
CC
Logrank p < 0.0001
Overall survival(probability)
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Time in months

Figure 4.7: Kaplan-Meier plots for SNPs that were significant at the $p < 10^{-6}$ level in overall survival analysis for advanced NSCLC cases (refer to Fig 4.2's legend).

a)



b)



The SNPs identified in the overall-survival analysis (Table 4.5 and Figures 4.5-4.7) are broadly consistent with those identified in cause-specific analysis (Table 4.4 and Figures 4.2-4.4): rs10230420, rs1868110, rs2206779 were each significant at the $p < 10^{-6}$ level for the corresponding stage-group (rs10230420 for all cases and early stage analysis; rs1868110 and rs2206779 for advanced stage cases). However, two SNPs (rs2056533 and rs6708630) that were significant in the early-stage overall survival analysis were not significant in cause-specific analysis (wherein their p-values were 3.58E-06 and 4.43E-06). Similarly, the linked SNPs rs3746619 and rs3827103 were significant in the cause-specific analysis but not in the overall survival analysis (where their p-values were both 1.32E-06).

4.6.3 Genes Identified in the Survival Analysis

Table 4.6: Description of genes harbouring significant SNPs.

SNP	Chromosome position	Gene/ Closest gene	Gene summary	Cytogenic position
rs6708630	10225807	*CYS1*	Involved in foetal kidney development[277].	2p25.1
rs1868110	27010197	*NEK10*	Involved in processes like mitosis, cell cycle, DNA repair, check point control, genotoxic stress. Also associated with breast and lung cancer[278].	3p24.1
rs10230420	29949780	*WIPF3*	Regulates cytoskeletal organisation playing a role in cell differentiation and spermatogenesis[279]	7p14.3
rs3746619	54823805	*MC3R*	Polymorphism in this gene is associated with obesity in humans[249].	20q13.2-q13.3
rs3827103	54824029			
rs2206779	69118705	*AF357533*	-	1q31.3
rs9949512	76641845	*SALL3*	Mutations in this gene may be associated with congenital disorder. Gene silencing is associated with oncogenesis through accelerated methylation[249].	18q23
rs2056533	114485145	*ZBTB20*	Involved in oncogenesis, haematopoiesis and immune responses[280]	3q13.2
rs2139133	171193917	*MYO3B*	Involved in protein kinase activity, motor activity and ATP binding[279]. Somatic mutation (nonsense) in this gene was identified in lung adenocarcinomas[281].	2q31.1-q31.2

The functional annotations of genes in which significant SNPs were located are tabulated in Table 4.6 (annotations obtained from NCBI for genes lying closest to a significant SNP). Genes involved in cytoskeletal organisation, gene expression, protein kinase, motor and ATP binding activity and oncogenesis were identified in the comparison for all NSCLC cases (*MYO3B, SALL3, WIPF3*). For the early stage NSCLC cases gene involved in obesity and

oncogenesis were identified (*ZBTB20, MC3R, CYS1, WIPF3*), while for advanced stage NSCLC cases, genes involved in various cellular functions were identified (*NEK10, AF357533*).

4.6.4 Joint Survival Analysis

The joint survival analysis was carried out by summing the total number of minor alleles identified at the $p < 10^{-6}$ level to assess the level of risk for carrying significant risk alleles. Separate cumulative analysis was carried out for all, early and advanced stage NSCLC cases using a Cox proportional hazard regression model adjusted for age at diagnosis, smoking pack years, stage and histological cell type, using the same dataset. A similar analysis was carried out by Huang *et al*. (2009)[261], to test the cumulative effect of the 5 SNPs that were discovered and validated by this study.

The number of risk alleles in the model was treated as categorical variable to study the risk associated with the effect of carrying a particular number of risk alleles. This model is similar to the genotypic model where risk estimates are obtained for the homozygous and heterozygous risk alleles using the major allele homozygote as the baseline variable.

For the early stage cause-specific survival analysis, one of the two SNPs: rs3746619 and rs3827103, were dropped from the joint analysis because they produced the same hazard ratios and minor allele frequencies, suggesting linkage disequilibrium.

For joint analyses (Table 4.7) for every model tested for the cause specific survival satisfied the proportionality hazard assumption detected by Schoenfeld residuals while for the overall survival analysis, the all and early stage model produced test statistics detecting

non-proportionality. For each of the model a trend of increased hazard ratio with increasing number of risk alleles was observed.

Table 4.7: Joint effect of significant SNPs decreasing survival at $p \leq 10^{-6}$ for both cause specific and overall survival analysis.

Cause Specific survival			**Overall survival**		
All NSCLC cases					
No. of risk allele	frequency	HR(95% CI)	No. of risk allele	frequency	HR(95% CI)
0	50	Referent	0	167	Referent
1	58	2.1 (1.26-3.50)	1	18	5.65 (3.18-10.05)
2	51	3.22 (1.92-5.39)			
3	14	8.27 (3.99-17.13)			
≥4	12	10.49 (4.97-22.14)			
Schoenfeld p-value - 0.155			Schoenfeld p-value - 0.01036		
Early stage NSCLC cases					
0	86	Referent	0	72	Referent
1	16	8.83 (4.10-19.01)	1	24	9.73 (4.89-19.37)
2	5	36.03 (10.83-119.94)	2	11	20.82 (8.7-49.83)
Schoenfeld p-value - 0.0954			Schoenfeld p-value - 0.0037		
Advanced stage NSCLC cases					
0	41	Referent	0	41	Referent
1	19	3.31 (1.75-6.26)	1	19	3.27 (1.73-6.19)
2	12	8.59 (3.6-20.52)	2	12	8.45 (3.55-20.15)
3	5	50.07 (14.37-174.54)	3	5	51.23 (14.69-178.66)
Schoenfeld p-value - 0.337219			**Schoenfeld p-value - 0.41013**		

Cox proportional hazard model after adjusting for age, sex, smoking pack years, stage and histological type.

4.7 Discussion

This study investigated the association of SNPs with the survival of NSCLC patients. Additionally, subgroup-specific SNPs were identified for both early and advanced-stage NSCLC cases. Survival analysis for all NSCLC cases identified three distinct SNPs (one associated with both cause-specific and overall survival; the other two being specific for overall survival) at the $p \leq 10^{-6}$ significance level. There was evidence that the proportional hazards assumption failed for the SNP rs10230420 in the overall survival study ($p = 0.010$), although there was little evidence to doubt this assumption for this SNP in the cause-specific survival study ($p = 0.075$) or for either of the other SNPs identified when considering all NSCLC cases. Proportional hazards are a fundamental assumption of the multivariate Cox regression models used here (as detailed in the introduction to this chapter) and elsewhere[175, 274]. Nonetheless, for rs10230420 in both cause specific and overall survival analysis of all NSCLC cases, the KM curves are well separated with a significant p-value for the log rank test (Figure 4.1a and Figure 4.2a).

For the survival of early-stage lung cancer cases, five SNPs were identified (rs10230420, rs3746619, rs3827103, rs2056533 and rs6708630, discussed above, in both cause-specific and overall survival analyses; the neighbouring SNPs rs3746619 and rs3827103 in cause-specific; rs2056533 and rs6708630 in overall survival analyses).Two of these SNPs, produced a non-significant tests statistics, satisfying the proportionality hazard assumption. Finally, for the advanced-stage lung cancer cases, two SNPs were identified in both the cause-specific and overall survival analysis. Neither of these SNPs showed evidence to dispute the proportional hazards assumption. The overlap in results for the advanced stage cases could be due to the number of events being close (71/78 for cause specific and 73/78 for overall) (Table 4.3, Table 4.4 and Table 4.5).

In the previously carried lung cancer research (Table 4.1a), only Wu *et al*. (2011)[253] tested for the proportionality hazard assumption. Since all the research involves identifying SNPs associated with lung cancer testing for proportionality hazard assumption is crucial, as the Cox proportional hazard model depends on it.

Functions of SNP-associated genes: SNP rs10230420 is located in the intron of *WIPF3* and was found to be associated with decreased overall and cause specific survival in all NSCLC cases. *WIPF3* is a member of the Wiskott–Aldrich syndrome protein (*WASP*)- interacting protein (*WIP*) family made up of the *WIP, WIPF3*, and the WIP- and CR16-homologous protein (*WICH/WIRE*) gene and is regulated by corticosteroids [282]. It is expressed in both the brain and testis and plays a role in spermatogenesis[282]. No role of this gene is reported in lung cancer and therefore needs evaluation.

SNP rs1868110 is located near genes *NEK10* (27,080,151 bp), *LOC101929642* (27,151,574 bp*), LRRC3B* (26,664,300 bp) and *MINOS1P3* (245,856 bp)[249]. No information is available for gene *MINOS1P3* while *LOC101929642* is a protein coding gene with no information. *NEK10* belongs to the *NEK* (NIMA-related kinase) serine threonine protein kinase family whose members function during mitosis, cell cycle, check point control, DNA damage repair and genotoxic (ultra violet, ionising radiation ,etoposide) stress[278, 283]. NEK10 was found to be associated with breast cancer identified by a large GWAS association study[278, 283]. A kinome analysis depicted that this gene was the only kinase tested in a group of 20 primary lung neoplasms and 7 lung cancer cell lines with multiple non-synonymous somatic mutations whose frequency matches that to the BRAF and *STK11/ LKB1* genes associated with lung cancer[278, 284]. Copy number deletion for 1 sample was observed for this gene, out of 17 samples, in a genomic study of smokers and non-smokers with NSCLC[285]. Mutational frequency analysis was the initial step but further research on its mechanism, including

expression and sequencing studies need to be carried out to improve the understanding of its role in lung cancer.

LRRC3B is a tumour suppressor gene expressed in normal brain, kidney and lung tissue[286, 287]. Its protein is located in the nucleus and functions in interaction, recombination, transcription, development, immune response, DNA repair, signal transduction, cell adhesion, expression and apoptosis[286, 287]. The gene's downregulation is associated with acute leukaemia, its promoter hypermethylation is significantly higher in colorectal tissue[286] and is epigenetically silenced in gastric cancer[287]. Other cancers that have reported the downregulation of *LRRC3B* include brain, breast, colon, prostate and testis[287]. *LRRC3B* is a tumour suppressor implicated in many cancers[286, 287] but has not yet been identified in lung cancer. Since it's published in many cancers it may have a common mechanism leading to oncogenesis and could serve as a potential biomarker.

SNP rs2056533 is located in intron of gene *ZBTB20* identified to be linked to decreased overall survival in early stage NSCLC patients[288]. *ZBTB20* belongs to a class of transcription factors associated with biological functions including transcription, proliferation, cell morphogenesis and death[288]. Expressed by hippocampal progenitor, it is known to play a crucial role in hippocampal development and its expression is associated with poor prognosis in hepatocellular carcinoma[288]. Downregulation of *ZBTB20* expression is involved in oncogenesis and also plays a role in haematopoiesis and immune responses[280]. Further work is necessary to identify the mechanisms involving this gene that may have implication in the pathogenesis of lung cancer.

SNPs rs3746619 and rs3827103, located in the *5'-UTR* and exon of the *MCR3* gene, respectively were associated with the decreased cause specific survival of early stage NSCLC cases. They produced the same hazard ratios in the cause specific survival analysis and may be in linkage disequilibrium. *MC3R* is a 7-transmembrane G-protein coupled

receptor expressed in hypothalamic nuclei that control human weight by modulating the body mass index (BMI), subcutaneous fat mass and insulin levels suggesting its association with human obesity[289]. Copy number deletion was observed for 1 sample out of 17, for this gene in a genomic study of smokers and non-smokers with NSCLC[285]. Though it's associated with obesity its role in lung cancer is still to be discovered.

SNP rs2139133 located in the intron of *MYO3B* was linked to decreased overall and cause specific survival in all NSCLC cases. *MYO3B* is expressed in the retina, kidney and testis and is associated with Bardet-Biedl syndrome characterised by dysmorphic extremities, retinal dystrophy, weight gain, renal deformities and malfunction, male hypogenitalism, mental retardation and diabetes mellitus[290]. Screening of lung cancer samples have identified missense germline and nonsense somatic mutations in a study conducted on coding exons of 518 protein kinases[281]. Expression and sequencing analysis of this gene will further increase the information available on its role in lung cancer.

SNP rs9949512 is located close to *SALL3* (76,829,394bp) and *LOC645321* (76,740,275bp)[249], and was found to be associated with decreased cause specific survival in all NSCLC cases. *SALL3* is a member of the sal-like (sall) gene family associated with embryonic development and consists of *sall1, sall2, sall3* and *sall4.* Loss of the *sall3* gene leads to palate deficiency, abnormalities in cranial nerves, and perinatal lethality and is also one of the genes whose deletion leads to the 18q deletion syndrome resulting in hearing loss, mental retardation, midfacial hypoplasia, late growth and limb deformities[291]. *SALL3* inhibits methylation and its decreased mRNA transcription is reported to be due to promoter hypermethylation, in hepatocellular carcinoma while its methylation levels are increased in bladder cancer[291]. Copy number deletion for 1 sample out of 17, was observed for this gene in a genomic study of smokers and non-smokers with NSCLC[285]. Though it's implicated in bladder and

hepatocellular carcinoma, no studies have been reported on its role in lung cancer and therefore research to understand its mechanism in lung cancer is necessary.

Design issues for lung cancer GWA survival studies: Previously published GWA survival studies (Table 4.1a) and this study, used Cox proportional hazard regression analysis to identify SNPs associated with survival in lung cancer cases[251-255, 261, 262]. While all the published studies (Table 4.1a) evaluated only overall survival, this study looked at both overall and cause specific survival in NSCLC patients[251-255, 261, 262]. Furthermore, while most studies looked at either early stage or advanced stage lung cancer populations (Table 4.1a), this study evaluated all NSCLC cases collectively, together with subgroup analysis only of early and advanced NSCLCs[251-255, 261, 262].

Most studies evaluated survival using the additive model but Wu *et al*. (2011)[253] was the only publication that evaluated all three; additive, dominant and recessive model while Lee *et al*. (2012)[255] used dominant and additive model. While Sato *et al*. (2011)[251], Huang *et al*. (2009), [261] and this study adjusted for multiple testing the rest of the publications[252-255, 262] did not. Niu *et al*. (2012)[262], Lee *et al*. (2012)[255], Sato *et al*. (2011)[251] and Tan *et al*. (2011)[254] were the only studies that did not validate their study using another population while Huang *et al*. (2009)[261], Wu *et al*. (2011)[253] and Hu *et al*. (2012)[252] validated their study using another population.

This study has identified genes not previously reported in lung cancer survival studies adding more potential genes to the list that needs further functional evaluation. This study also needs to be replicated to validate the findings. This can be done by carrying out the same statistical procedure in a different but ethnically same population and checking to see whether the same SNPs from the discovery population are significant[253, 261]. Another

drawback may be the smaller sample size of the study, though similar sample sizes have been used by other lung cancer survival studies[251, 264]. Genes identified by this study do not act singly but their action regulate the function of other genes, hence pathway analysis to evaluate molecular networks[292] in disease causation is necessary. This study has also looked at the joint analysis of SNPs significant at $p \leq 10^{-6}$ associated with decreased survival with all NSCLC cases together with subgroup analysis for early and advanced NSCLC cases.

This study identified various SNPs associated with NSCLC together with early and advanced stage NSCLC subgroup analysis. SNPs identified through such analysis can serve as crucial biomarkers for clinical purposes and personal treatment though they still need to be validated using a different dataset.

CHAPTER 5

IDENTIFICATION OF IMPORTANT PATHWAYS ASSOCIATED WITH SURVIVAL IN LUNG CANCER

5.1 Aim

The aim of this project was to identify biologically-related gene sets that form survival-associated molecular pathways in lung cancer patients. Both overall and cause-specific lung cancer survival analysis were considered. Genotypes for all assayed SNPs that are linked to at least one member of a given gene pathway were analysed using the random forest technique for different split rules, concurrently using data from the genome-wide dataset for non-small cell lung cancer (NSCLC) cases, as used in chapter 4. Pathway annotations for human genes were obtained from the *Homo sapiens* subset of the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database and associated SNPs were identified using National Centre for Biotechnology Information (NCBI) database.

5.2 Introduction

Although highly penetrant disease-risk alleles can be identified by pedigree analysis, a large number of common diseases, including lung cancer, are influenced by interactions between multiple risk loci (and non-genetic contributions from environmental, occupational and lifestyle factors)(as discussed in Chapter 1). Genome-wide association studies provide the standard method for identifying these low penetrance risk loci[111, 233, 234, 293, 294]. The large number of suggested loci, the uncertain functional effects of inheriting a risk allele and the poor reproducibility of disease-association for individual loci across different studies means that insight into the disease mechanism rarely follows directly from GWA studies (Chapter 3). An alternative to the single locus approach is to analyse SNPs associated with sets of

biologically-connected genes (for example, genes encoding proteins that function in the same signalling pathway), leading to an increased power to identify whether there is a pathway-level association of SNP inheritance with a given trait[295]. Pathway analysis of survival data would help understand the various biological mechanisms from lung carcinogenesis till death, potentially leading to reduced disease mortality and improving patient care[253, 296, 297].

5.2.1 Annotation Databases

A variety of gene annotation databases have been developed that are used to provide functional insights into the results of genome-scale analyses[298-300]. The databases provide varying levels of data access and differ in the methods of gene annotation and their relevance to disease biology.

Gene Ontology: Gene Ontology (GO) comprises three hierarchies of biological information relating, respectively, to the cellular components, molecular functions and biological processes occurring in organisms [298, 299]. The database is hierarchical in that more specific terms are nested within more general terms: for example, the nucleolus is subordinate to the nucleus in the cellular component ontology[298]. In addition, GO provides mappings between genes and terms within the GO hierarchy if a given gene (or its product) is known or predicted to play a role in the cellular component, pathway or activity to which a GO term refers[298]. Despite many annotations being based on electronic prediction, the annotations are reasonably accurate [301] and GO is frequently used in pathway analysis

because of the breadth of processes and functions covered by its approximately 30000 terms [302].

Kyoto Encyclopaedia of Genes and Genomes: Whereas GO documents the relations between different biological pathways and cellular components, the Kyoto Encyclopaedia of Genes and Genomes (KEGG) was initiated to document the relations between different biomolecules, both within and between species[300]. Three main components of KEGG are databases containing, respectively, information about genes (the KEGG 'Gene Universe'), biochemicals (the 'Chemical Universe') and functional annotations of proteins (the 'Protein Network')[300]. Although the 'gene' and 'chemical' databases can be used to identify sets of genes or proteins that share sequence similarity or perform similar enzymatic reactions[300], it is the KEGG protein network that is most relevant to the pathway-level analyses described here.

KEGG 'PATHWAY' is the main database within the KEGG protein network and comprises a series of pathway maps[303]. A pathway map is a molecular network that is relevant to a specific biological process or function and consists of nodes (which mainly represent proteins, but may also include genes or small molecules) and edges that connect relevant nodes (edges may document pairs of proteins/genes that function in the same reaction or which exhibit a physical or regulatory interaction)[303]. Hence, KEGG provides some information regarding the possible role of a protein in a given biological pathway[303]. Unlike GO, the pathway maps within KEGG include annotations relevant to non-physiological processes, such as disease pathways and networks relevant to drug development[298, 299, 303-306]. The systems level information in the KEGG PATHWAY database has largely been manually curated providing another benefit over GO[307-309].

5.2.2 Annotational and Methodological Challenges

Genomic and proteomic studies are generating high volumes of data that are relevant to specific biological processes[310]. For any given study, large number of variables (SNPs, mRNA / protein expression levels) may show an association with a disease process or respond to a particular stimulus[310]. Analysing this data in a pathway context provides a means to obtain biological insight but requires appropriate analytical methods that have a secure statistical basis and annotation databases that are unbiased, up to date and relevant to the disease / biological phenomenon under study[310]. Notably, the initial pathway analysis tools were developed for gene expression microarray studies and may not be directly relevant to GWAS studies and also, the quality of databases depends on the accuracy of gene prediction, the degree of automated curation, the pathways selected for curation and on publication bias within the literature[310].

To directly ascertain a biological role for a human gene requires experimental evidence[310]. As a consequence, functional annotation lags behind genomic identification for most genes and, in a given database, the level of support for the stored annotations may vary quite widely[310]. Indeed, GO, and some other databases, employ automated gene annotation systems to predict functional roles using a variety of data sources (for example, functionally-characterised orthologues, coexpression studies, protein-protein interactions, literature mining)[298, 310]. In October 2007, >95% of the annotations in GO were not manually curated, and the removal of these electronically-sourced associations reduced the number of annotated genes from 18,587 to 11,890[310].

The pathway knowledge databases typically lack information regarding the biological context under which the supporting experimental evidence was obtained[310]. As a result,

context-specific pathway associations that may be irrelevant to a study (for example, genes that function in a pathway in a tissue-specific manner) cannot be filtered out prior to performing a pathway analysis[310]. Moreover, databases such as GO only store associations between genes and pathways, rather than including the biological connections between the genes within that pathway (for example, kinase-substrate relations and protein-protein interactions, as included in KEGG pathway maps)[310]. Inclusion of dynamic and contextual information in the functional annotations of genes may enhance the possibilities of pathway analysis[310]. However, to capitalise on such fine-grained information may require considerably more advanced statistical tools than are currently available.

In addition to the choice of statistical methodology, the choice of annotation database to use and the pathways within that database to consider have a large impact on the outcome of a pathway analysis study[311]. Seemingly identical pathways may have different levels of definition within distinct databases[311]. For example, apoptosis is a single pathway in KEGG, but in GO, apoptosis can be further divided into inductive, regulatory and tissue-specific pathway components[311]. Allied to this, there is an ill-defined boundary between the pathway components and those molecules that regulate or are regulated by it[311]. Due to differences in the associated gene lists for related pathways between databases, results of a pathway analysis can be inconsistent across different databases[311].

The number of comparisons performed must be controlled for when testing multiple null hypotheses, such as for example, when testing a large collection of pathways for association with a disease state[312]. Bonferroni correction is based on the assumption that the tested hypotheses are independent, which is false in many pathway studies since the included gene lists may overlap between any pair of pathways[312]. Such stringent multiple testing methods reduce the power to detect true associations with a pathway[312]. Bonferroni or Sidak methods are therefore conservative[312], while procedures that control

the proportion of false positive significant pathways, such as that developed by Benjamini and Hochberg[313] may be more appropriate. A limitation of all of these approaches is that, depending on the pathway selection procedure, they assume the independence of all pathways that are nominally significant prior to controlling for multiple comparisons[312]. Bootstrapping methods, whereby datasets are generated by sampling with replacement, can be used to bypass the latter issue, but are computationally demanding[312].

Assigning SNPs to genes is an important step in determining the SNP complement of a pathway[312]. The most common method is to consider any SNP that lies within a gene-proximal region as a SNP associated with that gene[312]. The region typically includes the gene and a window of between 5kb[314] and 500kb[308] of surrounding sequence. Other issues that arise include the set of genes to include from the genome build (i.e. whether non-coding, pseudogenes or predicted genes should be included), the size of the sequence window and how to deal with SNPs that are attributed to more than one gene[312].

Depending on the choice of pathway analysis method, the number of SNPs and/or genes annotated to a pathway can influence the likelihood of a type-I error for that pathway[312]. This occurs, for example, when the method is based on taking the mean or the least of the SNP-level p-values as a representative gene-level p-value (discussed in Holmans *et al*. (2010)[312]). Since genes are of non-uniform size, larger genes will tend to have more associated SNPs than smaller genes[312]. Additionally, a larger choice for the sequence window will lead to more SNPs being assigned to a gene[312]. For a given gene, this both increases the possibility of finding a functional association, if one exists, and also leads to the consideration of many SNPs that may have no functional relationship with the gene[312]. Large windows also increase the overlap between neighbouring genes, thus increasing the number of shared SNPs[312]. Some studies have chosen to disregard SNPs that lie in these

regions of overlap[315], although an alternative is to perform permutation methods for these overlapping SNPs[312].

The high correlation seen between SNPs that lie in linkage disequilibrium may affect any pathway level analysis that considers SNP-level data independently[312]. Similar effects occur in gene expression data (due to cross reactivity of microarray probes and co-regulation of gene expression) where this interdependence is corrected for by permutation of phenotype labels amongst the samples[312]. An identical method has been proposed to deal with the interdependence between SNPs in pathway analysis however this method is computationally intensive[312].

5.2.3 Pathway Analysis

Pathway analysis tests for an association between a trait and a set of genes: the pathway[295]. The pathway tests employed are either self-contained or competitive[295]. In the former case, the association of a trait with a pathway is determined independently of its association with all other pathways; whereas in the latter case, summary statistics computed for each of a set of pathways are compared (perhaps controlled for pathway size, etc.) with those pathways having extreme values of the summary statistic considered to be trait-associated (under the assumption that most pathways are not associated with the trait)[295].

Design choices must be made prior to pathway-level analysis of genotypic data to ensure that both the statistical approach and the biological question are appropriate and, importantly, to ensure that the computing requirements of the method (processing,

memory, storage) can be met by the available resources on a reasonable timescale[295]. If the volume of genotypic data is large, it may be necessary to analyse only a restricted number of pathways[295]. Self-contained tests must be applied when using this candidate pathway approach, since pathway selection may bias the results of competitive pathway analysis[295].

There are good reasons to consider genome-wide pathway analysis for all pathways if it is computationally feasible to do so[295]. The hypothesis-free nature of such an approach means that novel aspects of disease aetiology may be revealed, thus potentially improving our knowledge of disease biology[295]. However, such approaches require adjustment for the analysis of multiple testing, and so have lower power to detect genuine associations of a pathway with a trait than do candidate-pathway approaches[295].

5.2.4 Pathway Analysis in Lung Cancer

Although pathway analysis to compare lung cancer cases against control subjects has been performed using genotype data[113, 193, 203, 233], no studies have addressed the association of pathways with lung cancer survival. This may be due to the unavailability of data for survival type analysis or the methodological challenge in modifying the existing case-control analysis methods, to analyse lifetime data.

Pathway analysis has been performed using gene set enrichment analysis (GSEA) for expression data of lung adenocarcinomas with “good” and “poor” as outcome variables[316]. The correlation of gene expression with group label (good or poor) was used to rank genes representing a defined set (based on either pathway or location)[316]. The pathway-

associated genes were checked to see whether they occur at random with respect to an ordered list of all analysed genes[316]. If the position of the genes was non-random, the pathway may be phenotypically related and its over-representation was determined using an enrichment score[316]. The significance of the score was determined by a phenotype-based permutation method[316]. This study identified 17 gene sets associated with amino acid and nucleotide acid metabolism, immune modulation and mTOR signalling[316].

Another study carried out a comparison of pathway analysis methods using Caucasian case-control SNP datasets[317]. Two datasets were used. Each dataset was generated by combining two existing studies, a 'Central European and Toronto' dataset (CETO; cases=2258/controls=3027) and a 'Germany and Texas' dataset (GRMD; cases=1639/controls =1618) to achieve similar sample sizes and power to detect an association[317]. Individuals were genotyped using the Illumina platform on either the HumanHap 300 or HumanHap 550 chip arrays[317]. Four methods (EASE, GenGen, SLAT and mSUMSTAT) were used to identify important lung cancer pathways[317]. The methods range from simple (EASE) to the more complex (GenGen, SLAT and SUMSTAT), developed to handle various issues associated with linkage disequilibrium and gene size[317].

SNPs were included if the minor allele frequency (MAF) was >0.01, Hardy Weinberg equilibrium (HWE) p-value was ≥0.001 in controls and a >95% genotyping rate was observed[317]. Additionally, individuals with misreported gender and >10% missing genotype were excluded[317]. SNPs within 20kb around the gene were included, including only pathways with a minimum of 15 genes and maximum of 200 genes[317]. X^2 test statistics from unconditional logistic regression adjusted for sex, age and country of origin were used in EASE, GenGen and mSUMSTAT, while unadjusted regression X^2 test statistics were used for SLAT[317]. For GenGen and mSUMSTAT, one thousand permuted logistic regressions were conducted by shuffling the case/control status for every run[317]. The SNP with the most

significant X^2 test statistic obtained from the above methods, was used to represent each gene[317]. Test statistics ($p \leq 0.05$) obtained for each SNP were used in EASE[317]. EASE score for enrichment representation was calculated using a modified Fisher Exact probability and FDR was calculated to control for multiple testing[317].

GenGen was conducted by calculating a Kolmogorov-Smirnov-like running sum statistic on the ranked X^2 test statistics (obtained through logistic regression), in descending order[317]. Pathway p-values were obtained through permuted normalised enrichment score (NER) derived using the SNP test statistics[317]. The genes from pathways were represented by their most significant SNPs[317].

The modified-SUMSTAT is similar to GenGen except that pathway-level significance is determined by averaging the X^2 test statistics[317]. The normalised test statistics and permutation of phenotype is what makes it different from the original SUMSTAT methods[317]. Unlike the other methods used by Fehringer *et al*. (2012) all SNP test statistics defining a gene are used for pathway analysis. The observed and permuted data for phenotype is used to determine pathways with test statistics reaching a specific threshold[317]. For all methods, multiple testing was controlled by the FDR method[317]. The SLAT program computes its own association statistics with the response variable, without adjusting for any covariates; and the pathway level test statistics are obtained by using the p-values that meet a specific threshold[317].

Nerve impulse, Ras-GEF and LDL binding pathways were significant in both the populations when using the EASE method; acetylcholine receptor, heme metabolic, porphyrin metabolic, pigment biosynthetic and the 4 iron,4 sulphur cluster-binding pathways were significant in both populations for the mSUMSTAT method while regulation of cell migration was significant in both populations for the SLAT method[317]. Pathways that were significant when using multiple tools on a single dataset included chloride ion binding,

interleukin-2 biosynthetic, regulation of cell migration, acetylcholine receptor, heme metabolic, complement activation, chromatin assembly and regulation of cell migration[317].

A two staged, random forest technique devised by Chung *et al*. (2012), was applied to a lung cancer GWAS data set of 663 cases and 642 controls genotyped at 496,761 loci using the Illumina 550K platform[318].

Training and testing sets were created in a random forest (RF) method whereby, in a given iteration, samples were assigned to a training set (with replacement) and the rest to a testing set, to develop SNP-based classification trees[318]. The above process was repeated, to create a forest composed of classification trees[318]. The random forest was then used to classify each sample (in the testing set) based on how frequently the sample was classified to each category across those classification trees, where the sample was present in the testing dataset[318]. Classification error rate and the variable importance are also calculated[318].

For the two-stage RF-based (TRF) pathway method, the above RF steps are carried out with SNPs in a pathway and the process is repeated for SNPs whose variable importance exceeds a pre-specified threshold[318]. The pathway is scored based on the prediction error rate (of the former step)[318]. The process is repeated by permuting the case-control status, a predefined number of times to calculate the p value for the pathway[318]. Significant pathways identified include cyanoamine acid metabolism, Fc gamma R-mediated phagocytosis, p53 signalling and pentose phosphate pathway[318].

Lee *et al*. (2013), conducted a pathway-based analysis on Korean lung cancer patients consisting of 869 cases and 1533 controls[319]. Eight hundred and eighty pathways were analysed using GSEA and validated by ARTP[319]. Pathways were filtered from the discovery dataset using GSEA if the p-value ≤ 0.025 and FDRs ≤ 25% and considered if the ARTP p-

value ≤ 0.01[319]. The majority of the subjects were genotyped using Affymetrix Genome-Wide Human SNP Array 5.0 (Affymetrix, Santa Clara, CA, USA) [319]. SNPs were filtered (MAF, HWE, genotyping call rate) for various quality control methods[319]. Pathways containing between 20 and 200 genes were included from pathway databases such as BioCarta, KEGG, Reactome and other curated pathway sets[319]. SNPs that were within 20 kb of a gene were included in the analysis[319]. Multivariate logistic regression was conducted using age, sex and smoking status from which p-values were derived for SNPs, for use in GSEA[319]. Each gene was represented by its most significant SNP and one thousand permutations of phenotype were conducted to calculate the test statistics for each gene[319]. An enrichment score was calculated for every pathway using the ranked gene list, and normalised to make comparisons between different sized pathways[319]. Multiple testing was controlled for by FDR (false discovery rate)[319]. The validation was carried out using ARTP, as described below, with the same number of permutations[319].

For the ARTP method, a truncation point (numeric value) is predefined, where the total numbers of p-values, arranged in ascending order; equal to the truncation point (predefined), are multiplied to obtain the p-value for the pathway[319]. The former arranged p-values are obtained from the significant SNPs in genes that are associated with a given pathway[319]. This method (ARTP) employs a combined statistics method that does not depend on pathway size[319].

Eleven of the 880 pathways were significant in the GSEA and subsequent ARTP pathway analysis for the additive and dominant model, namely, G1/S transition, cell cycle, G1/S check point, ABC transporters and signalling pathways (VEGF, phosphatidylinositol , Inositol phosphate metabolism, NRAGE (JNK), cell death (NRAGE, NRIF and NADE) and p75 NTR receptor mediated)[319]. Activation of the pre-replicative complex was significant in the dominant but not additive model when tested using ARTP[319].

A study on Han Chinese individuals was carried out to identify significant lung cancer pathways using 1473 cases and 1962 controls in the discovery stage and 858 cases and 1115 controls in the validation stage[320]. Candidates were genotyped using the Affymetrix genome wide human SNP Array 6.0. Misreported gender, familial relationship, extreme heterozygosity and genotype call (<95%) rate and outliers were a part of the individual based quality control exclusion criteria, together with the regular SNP based quality control (MAF, HWE, genotype call rate (<95%)) including only autosomal SNPs in the analysis[320].

One hundred and ninety one pathways from Biocarta and 177 pathways from KEGG were used, only considering those with a minimum of 10 and a maximum of 200 genes[320]. The gene boundary for SNP allocation was taken to be 50kb downstream or upstream[320]. Statistics associated for each SNP were obtained through logistic regression, after adjusting for age, sex, pack years of smoking and four principal components obtained from EIGENSTRAT 3.0[320]. Each gene was represented by its most significant SNP and the pathway was evaluated using GenGen[320].

Kolmogorov-Smirnov-like running sum statistics were used to obtain pathway-level enrichment scores (GenGen software[308])[320]. The case-control status was shuffled 1000 times and used to obtain permuted pathway associations, giving the normalised enrichment score adjusted for gene sizes[320]. The significance was set at $p \leq 0.05$ and FDR ≤ 0.5[320]. Twenty two pathways were identified in the discovery phase but only four ("achPathway" – the role of nicotine acetylcholine receptors in the regulation of apoptosis; "At1rPathway" – angiotensin II mediated activation of JNK pathway via Pyk2 dependent signalling; "metPathway" – signalling of hepatocyte growth factor receptor; and "rac1Pathway" – Rac1 cell motility signalling pathway) of them were significant in the replication and combined study[320].

Though little has been published using SNP data in survival analysis[296] the case-control studies mentioned above gives us various statistical approaches that may be altered to form the basis of survival data analysis. The case-control pathway studies have used various methodologies (GSEA, GenGen, SLAT, mSUMSTAT, TRF) and identified various pathways associated in ethnically different populations (Caucasian, Chinese, Korean)[317-320]. They have also compared different methods across two different study populations and results from different methods within a single population[317]. One study validated their results using a different method rather than a different population for both the dominant and additive model[319] while another conducted the analysis, separately, by population and also a combined analysis[320]. The results of these published case-control pathway studies are poorly reproduced between studies which may be due to the use of different methodologies and populations of different ethnicities.

5.2.5 Imputation

Genotype imputation increases the power of pathway analysis by detecting additional associations[321]. Missing genotypes may be inferred using haplotype Hidden Markov Models (HMMs). The HapMap CEU panel release 27 (NCBI build 36) can be used as the reference panel for Caucasians[201]. Haplotype HMMs are developed using phased genotype data, alternating between the sampling and model building processes[322]. The genotypes are sampled for every individual depending on the genotypic data and the current haplotype HMM built using phased haplotypes[322]. The phased haplotype for the first iteration is obtained by using HapMap allele frequencies to impute genotypes and then phasing heterozygous genotypes randomly[322]. For biallelic markers, the ungenotyped markers are

imputed by averaging the posterior allelic probability for that site[322]. This is obtained by summing the probability of the HMM states obtained using a reference panel with repeated iterations tending to improve the accuracy of imputation[322]. HMM-based imputation can be carried out using the programs BEAGLE, IMPUTE MACH and fast-phase from BIMBAM[321].

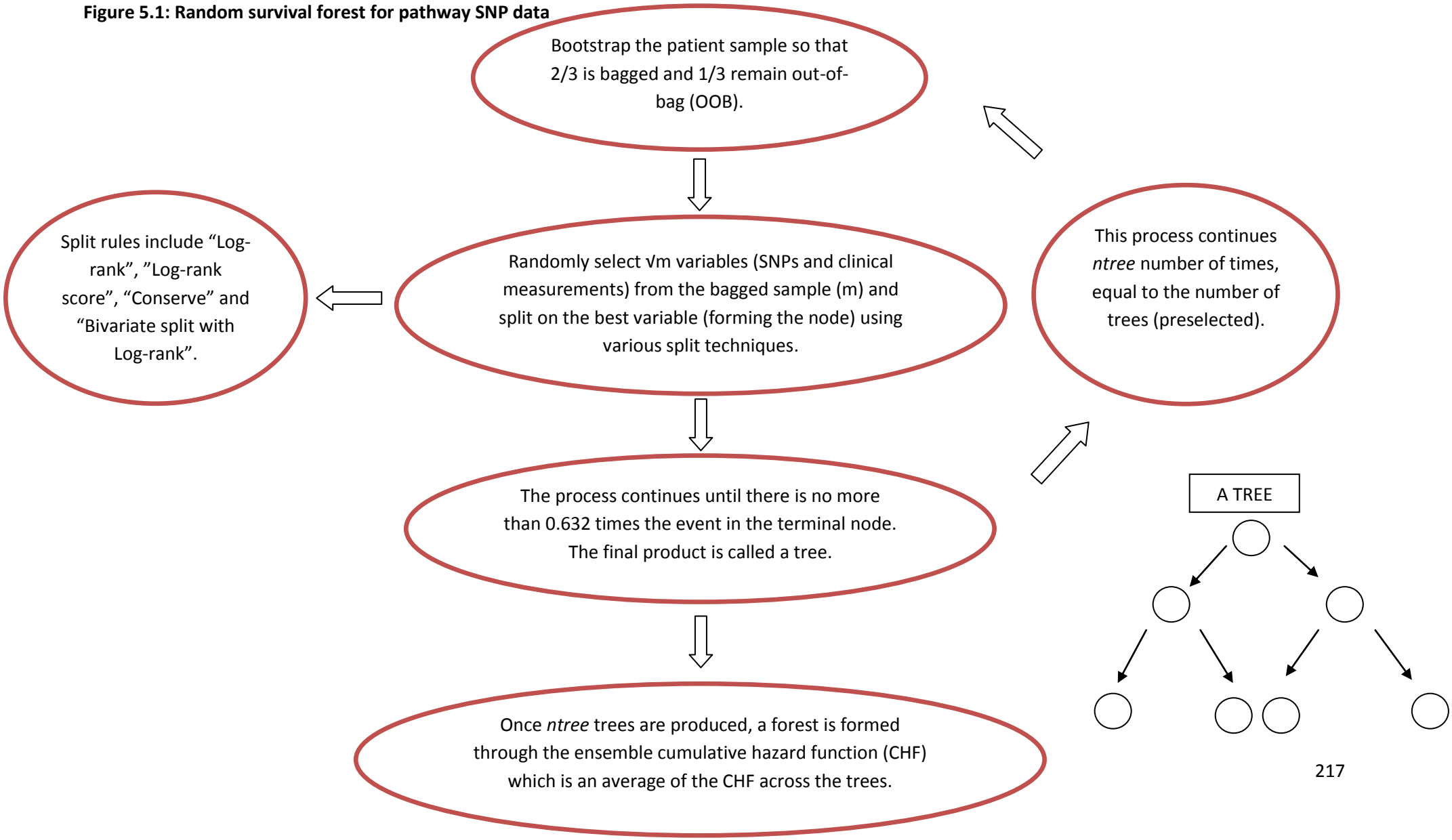
5.2.6 Random Forest Method

Random forest is a classification algorithm, where a set of random vectors are sampled independently to form a tree and many trees, collectively form a forest[323]. All trees forming the forest have the same distribution[323]. This approach is closely allied to random survival forest wherein randomisation is introduced while initiating a tree and splitting nodes to grow the trees[323]. It is a simple procedure requiring only the number of randomly selected predictors, the number of trees grown in the forest and the splitting rule[323].

Figure 5.1 explains the generation of a forest with reference to the SNP based random survival forest for pathways[296, 323]. The first step involves bootstrapping the sample (lung cancer patients) in such a way that two thirds of the sample is bagged leaving one third of it out-of-bag (OOB)[296, 323]. A survival tree is grown using the SNPs that are bagged[296, 323]. The tree is grown by splitting at each node, using √m randomly selected predictors where m is the number of bagged predictors[296, 323]. Of these √m predictors, the one that maximises the difference is used to split the node into daughter nodes using the survival split criteria[296, 323]. This process ends when terminal nodes contain no more than 0.632 times the number of events[296, 323]. The OOB samples are used to determine the variable importance and the predictive error rate[296, 323].

The expected number of events and times obtained through a 10 fold cross validation where 90% is used to train the data while the rest (10%) is used for predicting the event and survival time[323]. Individuals are grouped into high and low risk, and a log rank test is used to determine the significance of the pathway[296, 323]. The discriminatory ability of the prediction technique was also evaluated using an area under the receiver operating characteristic curve[296, 323].

Figure 5.1: Random survival forest for pathway SNP data



Splitting rules for random forests

Splitting rules are applied to split nodes into daughter nodes in such a way that it maximises the difference between the two sets of data[323]. Consider n individuals at node h which will be divided into two nodes on a predictor x that maximise the difference between the daughter nodes[323]. The following are the splitting rules used in the random forest technique.

Log-rank:

For a predictor x (here, the genotype at a specific locus or phenotype measurements) at value c, the LR statistic is defined as

$$LR(x,c) = \sum_{i=1}^{N}\left(d_{i,1} - X_{i,1}\frac{d_i}{X_i}\right)\left[\sum_{i=1}^{N}\frac{X_{i,1}}{X_i}\left(1-\frac{X_{i,1}}{X}\right)\left(\frac{X_i - d_i}{X_i - 1}\right)d_i\right]^{1/2}$$

Where X_i is the number of individuals at risk ($X_i = X_{i,1} + X_{i,2}$, where $X_{i,1}$ and $X_{i,2}$ are the respective values for the two daughter nodes) and d_i is the number of deaths at time t_i [323]. The output obtained using the above equation is the measure of node separation, the value being directly proportional to the extent of the difference between the two groups[323].

Conserve:

The Nelson-Aalen cumulative hazard estimator for daughter j is

$$\widehat{H}_j(t) = \sum_{t_{i,j} \le t} {}^{r_{i,j}}/_{V_{i,j}}$$

Where $r_{i,j}$ and $V_{i,j}$ are the number of deaths and the number of individuals at time t_i in daughter node j[323]. The Nelson Aalen estimator can be written to fit the survival data in the daughter node j in the form

$$\sum_{l=1}^{n_j} \widehat{H}_j(T_{l,j}) = \sum_{l=1}^{n_j} \varepsilon_{l,j}$$

where $T_{l,j}$ and $\varepsilon_{l,j}$ are the time and censoring indicator pairs for l=1,…n_j. Let the $T_{(1),j} \le T_{(2),j} \le \cdots \le T_{(nj),j}$ be the ordered time intervals for daughter j and $\varepsilon_{(l),j}$ be the censoring indicator for $T_{(l),j}$. For k=1,...,n_j [323]

$$Conserve\ (x, c) = \left(\frac{1}{X_{1,1} + X_{1,2}} \sum_{j=1}^{2} X_{1,j} \sum_{k=1}^{n_j - 1} \left| O_{k,j} \right| \right)$$

where $O_{k,j} = \sum_{l=1}^{k} \widehat{H}_j \left(T_{(l),j} \right) - \sum_{l=1}^{k} \varepsilon_{(l),j}$ [323]

Conserve test for x at value c (*Converse(x,c)*) produces a small value if the nodes are well separated and therefore the magnitude of node separation is represented as 1/(1+*Conserve(x,c)*)[323].

Log-rank score:

Let $x_1 \le x_2 \le \cdots \le x_n$ be a set of ordered predictor variable x, and ϑ_k the indicator variable, 1 if the event is observed and 0 otherwise[323].

$$LRS(x,c) = \frac{\sum_{x_l \le c}\left(a_l - \sum_{k=1}^{\sigma_l}\frac{\vartheta_k}{n-\sigma_k+1} - n_1\left(\overline{a_l - \sum_{k=1}^{\sigma_l}\frac{\vartheta_k}{n-\sigma_k+1}}\right)\right)}{\left[n_1\left(1-\frac{n_1}{n}\right)s_a^2\right]^{0.5}}$$

s_a^2 is the variance of $a_1 - \sum_{k=1}^{\sigma_l}\frac{\vartheta_k}{n-\sigma_k+1}$ which is the formula for computing the ranks for each survival time T_l and σ_k = #[t : $T_t \le T_k$] represents the total number of events (death/censor) observed at or before T_t . The absolute value of LRS for x at value c (LRS (x,c)) is the measure of node separation[323].

As depicted in Figure 5.1, the variables (SNPs and clinical measurements) in the bagged sample are used to select the best variable that can maximise the difference in the daughter nodes[296, 323]. This variable splits the bagged sample of patients using the different split rules to form the tree. Many trees thus form the random forest[296, 323].

Bivariate random survival forest with log-rank split rule:

When the random survival forest method is used with the above split criteria, $(m)^{0.5}$ predictors are randomly sampled for splitting and the node is generated using a single predictor[296]. However for the bivariate random survival forest with log-rank split (bRSF LR) method, LR split is used for the node split using pair of predictors that determine the best split that maximises the survival difference between daughter nodes[296].

These are some of the used split-criteria for random forest techniques, each employing a different method to calculate the extent of node separation[296]. Of all the above split criteria available, the bivariate random survival forest with LR splitting criteria, which takes into account the correlation between SNPs and unlike other methods that use a single variable

during the splitting procedure, uses two variables, was the best according to Pang *et al*. (2010)[296]. This was because it produced a small type I error relative to the other methods when tested using simulated datasets[296]. It was also the best in terms of power for a sample size of as small as 50 using simulated datasets[296].

Though pathway analysis using the random forest technique this method is recently being tested for survival analysis of SNP data[296], it has made the analysis of pathway-based survival data possible.

5.3 Material and Methods

One hundred and eighty five NSCLC cases from Liverpool were identified. Blood DNA was extracted using Qiagen kits and genotyped using the 300K HumanHap Illumina bead chip array. The CEU HapMap3 dataset was utilised to carry out imputation of missing genotypes in every chromosome[201]. Outliers from the case dataset, identified using the nearest neighbour technique for outlier detection in PLINK[227] were removed and the remaining data was merged with the HapMap 3 CEU population. The genotype dataset was quality controlled to include single nucleotide polymorphisms (SNPs) with a minor allele frequency >1% and genotypic call rate of >95% and to exclude SNPs with a Hardy Weinberg equilibrium p-value < 0.001. The data were checked to remove any duplicates, related individuals and discordant sex. Every individual had a genotype call rate of >95%. Imputation was carried out using BEAGLE 3.3.2[322] by first phasing the reference (CEU HapMap3) dataset.

The survival status for each case was determined using the ONS (Office for National Statistics)[125] registry data, the most recent ONS update being in February 2012. Cause specific death was identified if the cause of death was reported as "C34" ('Malignant neoplasm of lung and bronchus') or "C780" ('Secondary malignant neoplasm of lung') (ICD-10) while the survival status for overall analysis was death due to any cause. The survival time was calculated using the date of diagnosis and date of death or date last reported alive.

SNPs were coded in an additive mode (0, 1, 2) with reference to the number of minor alleles carried by an individual[261]. A selection of 18 pathways from the KEGG website for species *Homo sapiens* were utilised to run this analysis (http://rest.kegg.jp/list/pathway/hsa)[324, 325]. A perl script was written to extract the genes using the KEGG link http://rest.kegg.jp/link/genes/a where "a" was replaced by the relevant pathway reference number[324, 325]. Those genes that were extracted for each KEGG pathway were used to extract pathway-associated SNPs using two files downloaded from dbSNP in July 2012 (Build 132, https://cgsmd.isi.edu/dbsnpq/downloads.php)[326]. The first file, a table titled "GeneToName", contained the Entrez gene ID, gene symbol, gene name, gene type and taxonomy id for each gene, and the other, titled "_loc_snp_gene_ref", contained the SNP id, contig id, numeric NCBI Entrez gene ID, gene symbol, start and stop position of the genes and accession of RefSeq mRNA associated with the Entrez ID, the functional properties of the SNP, base position, allele and codon position and type (whether coding or non-coding)[326]. A perl script was written to extract genes from pathways in KEGG and SNPs from genes using the above two files. The SNPs were filtered to include only those that were not in high LD (section 3.2.7) (r^2>0.8)[297]. This was done by identifying SNP pairs in LD (r^2>0.8) within a distance 250Kb and removing one of them and was performed using scripts written in R[229] and PLINK[227]. Hence, the final dataset contained SNPs that were not in high LD (r^2>0.8).

The survival pathway analysis was run to identify pathways associated with cause-specific and overall survival analysis included age at diagnosis, histological types (adenocarcinoma versus squamous cell carcinoma), smoking pack years, sex and stage (I, II, III and IV) and the LD-filtered SNPs.

'Pwayrfsurvival', a script developed in R[229] by Pang *et al*. (2010), was modified for use with the dataset[296]. It was applied to identify pathways associated with overall and cause-specific survival using various split algorithms including log-rank, log-rank score, conserve and bivariate with log-rank. It depends on other R packages such as "brsf"[327], "survival"[328, 329] and "survivalROC"[330]. FDR-adjusted p-values were calculated using R[229]. This was done by first ranking the p-values in ascending order and choosing the minimum of either the corresponding p-value or the value obtained by multiplying the rank number of the p-value with the nominal p-value (0.05) and dividing it by the total number of tests[331]. All analysis were conducted using BEAGLE[322] , perl, R[229] and PLINK[227].

5.4 Results

The population characteristics of the 185 NSCLC cases from Liverpool are tabulated in the previous chapter (Table 4.2 and Table 4.3). A set of 18 pathways were selected to test for significance in the random forest survival pathway analysis in both cause-specific and overall survival analysis using the log-rank, log-rank score, conserve and bivariate random survival forest with log-rank split rule. The number of pathways was preselected due to time restrictions, though the methods could have been applied to all pathways from the KEGG database. The KEGG pathways that were included because of their association with lung cancer were alcoholism[332], apoptosis[333], base excision repair (BER)[334], cell cycle[335] ,

Chemical carcinogenesis[336], ECM receptor interaction[337], Erbβ signalling pathway[338], Insulin secretion[339], Mismatch repair (MMR)[340], NF-κβ signalling pathway[341] , nicotine addiction[342], NSCLC, Notch signalling pathway[343], Nucleotide excision repair (NER)[344], p53 signalling pathway[345], Small cell lung cancer, TGF-β signalling pathway[346] and VEGF signalling pathway[347].

The output of the Pwayrfsurvival script is a p-value and an area under the receiver operative characteristic curve (AUC) value for each pathway that was considered; these are computed by comparing the predicted events and times to those observed in the original dataset[296].

Important pathways were identified via a log rank test computed by grouping individuals into high and low risk groups of approximately equal sizes[296]. This grouping depends on the expected survival times and events computed using a ten-fold cross validation in which 90% of the sample was trained to be tested in the remaining 10% at each fold[296]. A small p-value of the log rank test was indicative of the importance of a given pathway in lung cancer survival[296].

The predictive accuracy of this pathway-based method for determining patient survival, was evaluated by using AUC employing the expected survival times and events to evaluate how these SNPs predict a lung cancer free survival[296]. A large AUC value indicates a good prediction. This analysis is a function of the time t and proceeds through estimating the sensitivity and specificity at different cut offs[296].

Table 5.1: Results for cause specific and overall random forest pathway survival analysis using log-rank split rule

Pathway	No. of SNPs	Log-rank					
		Cause specific			Overall		
		p value	FDR adjusted	AUC	p value	FDR adjusted	AUC
Alcoholism	870	**0.0313**	**0.0332**	0.61	0.3547	0.3547	0.57
Apoptosis	372	**4.26E-08**	**1.92E-07**	*0.75*	**3.28E-08**	**1.97E-07**	*0.76*
Base excision repair	81	**3.39E-09**	**2.08E-08**	*0.77*	**1.48E-08**	**1.33E-07**	*0.77*
Cell cycle	397	**1.99E-09**	**2.08E-08**	*0.72*	**1.96E-09**	**3.53E-08**	*0.72*
Chemical carcinogenesis	178	**8.67E-08**	**3.12E-07**	0.67	**2.81E-06**	**8.43E-06**	0.69
ECM receptor interaction	1097	**0.0007**	**0.0009**	0.60	**0.0041**	**0.0053**	0.65
Erbβ signalling pathway	1078	**0.0437**	**0.0437**	0.60	**0.0087**	**0.0097**	0.57
Insulin secretion	1439	**0.0001**	**0.0002**	0.63	**0.0020**	**0.0035**	0.58
Mismatch repair	103	**2.20E-07**	**6.60E-07**	*0.72*	**1.01E-06**	**3.64E-06**	*0.71*
NF kappa β signalling pathway	471	**7.47E-06**	**1.68E-05**	0.67	**0.0025**	**0.0038**	0.64
Nicotine addiction	574	**0.0005**	**0.0007**	*0.71*	**0.0022**	**0.0036**	*0.70*
Non-small cell lung cancer	730	**0.0047**	**0.0053**	0.57	**0.0036**	**0.0049**	0.54
Notch signalling pathway	347	**0.0012**	**0.0015**	0.62	**0.0061**	**0.0073**	0.65
Nucleotide excision repair	130	**3.46E-09**	**2.08E-08**	*0.70*	**6.38E-06**	**1.64E-05**	0.69
p53 signalling pathway	228	**4.86E-06**	**1.25E-05**	0.67	**3.99E-07**	**1.80E-06**	*0.70*
Small cell lung cancer	911	**0.0002**	**0.0003**	0.62	**0.0016**	**0.0031**	0.65
TGF beta signalling pathway	412	**1.02E-05**	**2.04E-05**	0.56	**9.31E-05**	**0.0002**	0.59
VEGF signalling pathway	455	**0.0002**	**0.0003**	0.56	**0.0143**	**0.0151**	0.48

Italics -AUC≥0.70; **bold - p≤0.05**

Table 5.2: Results for cause specific and overall random forest pathway survival analysis using bivariate random survival forest with log-rank split

Pathway	No. of SNPs	Bivariate random survival forest with log-rank					
		Cause specific			Overall		
		p value	FDR adjusted	AUC	p value	FDR adjusted	AUC
Alcoholism	870	0.7274	0.7274	0.50	0.6504	0.8362	0.51
Apoptosis	372	**4.19E-06**	**1.51E-05**	*0.77*	**0.0074**	**0.0166**	*0.73*
Base excision repair	81	**2.68E-07**	**4.02E-06**	*0.78*	**8.90E-08**	**1.60E-06**	*0.79*
Cell cycle	397	**4.18E-06**	**1.51E-05**	*0.71*	**0.0021**	**0.0083**	*0.70*
Chemical carcinogenesis	178	**4.23E-05**	**0.0001**	0.66	**0.0029**	**0.0086**	*0.71*
ECM receptor interaction	1097	0.1066	0.1599	0.61	0.1949	0.2923	0.64
Erbβ signalling pathway	1078	0.5642	0.5974	0.59	0.9987	0.9987	0.52
Insulin secretion	1439	0.4887	0.5498	0.46	0.8024	0.9576	0.43
Mismatch repair	103	**6.70E-07**	**4.02E-06**	*0.73*	**7.01E-05**	**0.0006**	*0.72*
NF kappa β signalling pathway	471	**0.0015**	**0.0034**	0.52	**0.0035**	**0.0089**	0.53
Nicotine addiction	574	0.1057	0.1599	0.64	0.4166	0.5768	0.66
Non-small cell lung cancer	730	0.2307	0.2768	0.49	0.8938	0.9576	0.49
Notch signalling pathway	347	**0.0364**	0.0655	0.59	0.0815	0.1466	0.62
Nucleotide excision repair	130	**2.19E-05**	**6.57E-05**	0.67	**0.0001**	**0.0008**	0.68
p53 signalling pathway	228	**5.18E-07**	**4.02E-06**	0.67	**0.0023**	**0.0083**	0.67
Small cell lung cancer	911	0.2046	0.2630	0.61	0.1897	0.2923	0.63
TGF beta signalling pathway	412	**0.0051**	**0.0103**	0.49	0.0790	0.1466	0.49
VEGF signalling pathway	455	0.1487	0.2059	0.38	0.9044	0.9576	0.38

Italics -AUC≥0.70; **bold - p≤0.05**

Table 5.3: Results for cause specific and overall random forest pathway survival analysis using conserve split rule

Pathway	No. of SNPs	Conserve					
		Cause specific			Overall		
		p value	FDR adjusted	AUC	p value	FDR adjusted	AUC
Alcoholism	870	0.4226	0.4226	0.56	**9.35E-01**	0.935	0.56
Apoptosis	372	**1.05E-05**	**3.78E-05**	*0.78*	**4.47E-05**	**0.0001**	*0.78*
Base excision repair	81	**3.51E-11**	**6.32E-10**	*0.81*	**1.74E-09**	**2.93E-08**	*0.81*
Cell cycle	397	**2.90E-06**	**1.31E-05**	*0.71*	**3.04E-07**	**1.82E-06**	*0.71*
Chemical carcinogenesis	178	**6.41E-07**	**3.85E-06**	0.66	**6.00E-05**	**0.0002**	0.66
ECM receptor interaction	1097	**0.0250**	**0.0322**	0.62	0.4761	0.5041	0.62
Erbβ signalling pathway	1078	**0.0018**	**0.0028**	0.60	**9.39E-02**	0.1207	0.60
Insulin secretion	1439	**2.14E-04**	**0.0004**	0.50	**0.0197**	**0.0273**	0.50
Mismatch repair	103	**8.75E-05**	**0.0002**	*0.78*	**3.26E-09**	**2.93E-08**	*0.78*
NF kappa β signalling pathway	471	**0.0001**	**0.0002**	0.57	**0.0009**	**0.0016**	0.57
Nicotine addiction	574	0.0851	0.0958	0.68	**4.65E-03**	**0.0070**	0.68
Non-small cell lung cancer	730	**3.62E-02**	**0.0434**	0.51	0.3173	0.3570	0.51
Notch signalling pathway	347	**1.76E-02**	**0.0244**	0.65	**2.03E-03**	**0.0033**	0.65
Nucleotide excision repair	130	**7.21E-09**	**6.49E-08**	*0.76*	**1.25E-05**	**4.50E-05**	*0.76*
p53 signalling pathway	228	**5.89E-05**	**0.0002**	0.63	**4.76E-07**	**2.14E-06**	0.63
Small cell lung cancer	911	0.3077	0.3258	*0.75*	0.2098	0.2518	*0.75*
TGF beta signalling pathway	412	**0.0004**	**0.0006**	0.51	**7.11E-05**	**0.0002**	0.51
VEGF signalling pathway	455	**1.31E-04**	**0.0003**	0.51	**0.0007**	**0.0015**	0.51

Italics -AUC≥0.70; **bold - p≤0.05**

Table 5.4: Results for cause specific and overall random forest pathway survival analysis using log-rank score split rule

Pathway	No. of SNPs	Log-rank score					
		Cause specific			Overall		
		p value	FDR adjusted	AUC	p value	FDR adjusted	AUC
Alcoholism	870	0.9492	0.9492	0.56	0.3178	0.3575	0.60
Apoptosis	372	**9.98E-05**	**0.0003**	*0.70*	**2.00E-06**	**1.20E-05**	*0.70*
Base excision repair	81	**3.71E-09**	**6.68E-08**	*0.76*	**5.42E-08**	**9.76E-07**	*0.76*
Cell cycle	397	**5.36E-06**	**1.93E-05**	0.66	**0.0003**	**0.0008**	0.63
Chemical carcinogenesis	178	**3.94E-05**	**0.0001**	0.66	**8.87E-05**	**0.0003**	0.69
ECM receptor interaction	1097	**0.0116**	**0.0161**	0.57	**0.02860**	**0.0468**	0.68
Erbβ signalling pathway	1078	0.3932	0.4163	0.62	0.10934	0.1514	0.64
Insulin secretion	1439	**2.49E-03**	**0.0045**	0.58	0.15246	0.1830	0.55
Mismatch repair	103	**7.86E-07**	**3.64E-06**	0.69	**2.13E-07**	**1.92E-06**	*0.71*
NF kappa β signalling pathway	471	**0.0050**	**0.0074**	0.55	**0.00053**	**0.00120**	0.54
Nicotine addiction	574	**6.09E-02**	0.0685	0.68	0.14427	0.18295	0.61
Non-small cell lung cancer	730	**4.78E-02**	0.0574	0.47	0.43797	0.46374	0.49
Notch signalling pathway	347	**3.44E-03**	**0.0056**	0.58	**0.03260**	**0.04890**	0.61
Nucleotide excision repair	130	**5.19E-07**	**3.64E-06**	*0.72*	**6.20E-06**	**2.79E-05**	0.68
p53 signalling pathway	228	**8.08E-07**	**3.64E-06**	0.60	**3.77E-05**	**0.00014**	0.60
Small cell lung cancer	911	**2.89E-02**	**0.0372**	0.55	0.63497	0.63497	0.59
TGF beta signalling pathway	412	**1.67E-03**	**0.0038**	0.55	**0.00144**	**0.00287**	0.56
VEGF signalling pathway	455	**2.08E-03**	**0.0042**	0.54	**0.00984**	**0.01772**	0.50

Italics -AUC≥0.70; **bold - p≤0.05**

Table 5.5: Outcome summary of the results for the four split rules

Pathway	Cause specific survival				Overall survival			
	Log-rank	bivariate RSF with log-rank split	Conserve	Log-rank score	Log-rank	bivariate RSF with log-rank split	Conserve	Log-rank score
Alcoholism	++	--	--	--	--	--	+-	--
Apoptosis	++	++	++	++	++	++	++	++
Base excision repair	++	++	++	++	++	++	++	++
Cell cycle	++	++	++	++	++	++	++	++
Chemical carcinogenesis	++	++	++	++	++	++	++	++
ECM receptor interaction	++	--	++	++	++	--	--	++
Erbβ signalling pathway	++	--	++	--	++	--	--	--
Insulin secretion	++	--	++	++	++	--	++	--
Mismatch repair	++	++	++	++	++	++	++	++
NF kappa β signalling pathway	++	++	++	++	++	++	--	++
Nicotine addiction	++	--	--	+-	++	--	--	--
Non-small cell lung cancer	++	--	++	+-	++	--	--	--
Notch signalling pathway	++	+-	++	++	++	--	--	++
Nucleotide excision repair	++	++	++	++	++	++	++	++
p53 signalling pathway	++	++	++	++	++	++	--	++
Small cell lung cancer	++	--	++	++	++	--	--	--
TGF beta signalling pathway	++	++	++	++	++	--	--	++
VEGF signalling pathway	++	--	++	++	++	--	--	++

++=significant (p<0.05) after FDR correction; +-= significant (p<0.05) before FDR correction; -- = not significant (p<0.05) before and after FDR correction

The most significant pathway for random survival forest (RSF) with log-rank split was 'cell cycle' in the cause specific and overall survival while 'BER' was the most significant for the cause specific and overall RSF analysis using bivariate RSF with log-rank split, log-rank score split rule and the conserve split rule.

The pathways that were both significant ($p \leq 0.05$) and produced an AUC of ≥ 0.7 included 'BER', 'MMR', 'NER', 'cell cycle', 'apoptosis' and 'nicotine addiction' for cause-specific RSF using log-rank split while 'BER', 'MMR', 'cell cycle', 'apoptosis', 'nicotine addiction' and 'p53 signalling pathway' were pathways significant for the overall RSF using log-rank split (Table 5.1).

For the cause specific bivariate RSF with log-rank split, the significant pathways ($p \leq 0.05$) with AUC ≥ 0.7 were 'apoptosis', 'BER', 'MMR' and 'apoptosis', while 'BER', 'MMR', 'apoptosis', 'cell cycle' and 'chemical carcinogenesis' were significant ($p \leq 0.05$; AUC ≥ 0.7) for the overall bivariate RSF with log-rank split (Table 5.2).

For the cause specific and overall RSF with conserve split rule, 'apoptosis', 'cell cycle', 'NER', 'BER' and 'MMR' were significant ($p \leq 0.05$; AUC ≥ 0.7) pathways (Table 5.3), while for the log-rank score split rule (Table 5.4) 'apoptosis', 'BER' and 'MMR' were significant in the overall RSF pathway analysis while 'BER', 'NER' and 'apoptosis' were significant in the cause specific RSF pathway analysis.

Table 5.5 displays the result for the pathway outcome using the four split rules. 'Apoptosis', 'BER', 'cell cycle', 'chemical carcinogenesis', 'MMR' and 'NER' were significant after controlling for FDR in all of the RSF methods, for both the cause-specific and overall-survival approaches. These pathways could be analysed further in lung cancer survival research. There were more significant pathways in the cause-specific analysis compared to the overall survival analysis, for the four split rules. 'Erbβ signalling', 'nicotine addiction',

'NSCLC' and 'SCLC' pathways were not significant in any of the overall survival analysis for the different split rules except log-rank while alcoholism is the only pathway that was not significant for any of the cause specific analysis for the different split rules except log-rank.

For the pathways that were significant in all of the analyses for the different split rules, the number of SNPs ranged from 81 to 397. There were other pathways with a larger complement of SNPs but which were insignificant in some of the analyses suggesting that the significance of the pathway in the random survival technique may not depend on the number of SNPs.

The results also show that the log rank split analysis was significant for almost all analysed pathways. A high level of similarity in the p-values for pathways was observed across the overall-survival and cause-specific survival analyses using the bivariate split with log rank. Additionally, a different distribution of results was observed for the conserve and the log rank score split rule between the cause specific and overall survival analyses. The nature of these results for the same pathways using different split rules suggests that some may be overly conservative while others may have a tendency to produce false positive results.

5.5 Discussion

Cancer is a complex process involving multiple pathways of genes that act in synergy[348]. These pathways may be studied to understand their role in tumour biology and knowledge gained from such studies may ultimately aid in early detection and improve patient outcome[348]. Pathways are a better option than a single gene or SNP analysis as identifying a single variable does not explain its role in the disease but a pathway is already a set of linked variables, thus a better option for future research[113, 193, 203, 233, 317-320].

This study conducted a pathway analysis using variations on the random forest algorithm with different split techniques and identified a series of pathways that significantly associated with survival (Table 5.1 to Table 5.4). A selection of 18 pathways were analysed for their association with lung cancer, within this set, pathways that were both significantly associated with lung cancer survival ($p \leq 0.05$) and had an AUC of ≥ 0.7 were identified. The identified pathways collectively included 'BER', 'MMR', 'NER', 'cell cycle', 'apoptosis' and 'chemical carcinogenesis'. An interesting point to note here is that only 'cell cycle' pathway from the above list was significant in the genome wide incidence (case-control) pathway analysis (discussed in section 5.2.4) published by Lee *et al*. (2013)[319].

Another interesting observation to note is that 'NSCLC' and 'SCLC' pathway were not significant in all analyses. This may be due to the incomplete dataset regarding the pathway elements comprising the NSCLC and SCLC, as a pathway for cancer is a complex interplay of many pathways.

Generally, gene families involved in lung cancer tumour growth and metastasis include growth factor signalling, second messengers, cell cycle regulation, apoptosis/senescence, adhesion, migration, DNA repair and differentiation[348]. Epidermal growth factor receptors (EGFRs) are overexpressed and frequently are subject to activating mutations in NSCLCs where they stimulate cell proliferation, while second messengers transmit cell proliferation signals within cells[348]. Cell cycle regulatory genes participate in various phases of the cell cycle, controlling the proliferation of the dividing cell[348]. The cell cycle pathway, for instance, could be deregulated when genes like *TP53* and *RB* mutate[348]. *TP53*, referred to as the "guardian of the genome", is activated during genotoxic stress[348]. Its many roles include halting the cell cycle at the G1-S check-point (through reduction of *RB* phosphorylation), stimulating DNA repair and initiating cellular apoptosis (through regulating *BAX/BCL2* gene expression)[348]. The main agents in the apoptotic pathway include *TP53* and *BCL2* proto-

oncogene that protects against apoptosis[349]. In cancer, however, inactivation of *TP53* causes uncontrollable growth through bypassing apoptosis[348]. *BCL2* expression is more frequently elevated in SCLCs (75-95%) than NSCLCs[349]. There are various carcinogenic agents including chemicals such as aromatic amines, aldehydes and benzene[350] that may bring about the above mentioned changes. These changes would continue to remain or progress if not treated and therefore would influence not only the incidence but survival of the cancer. Furthermore, incidence analysis may detect polymorphisms that increase the likelihood of developing cancer (see section 5.2.4), whereas survival analysis may identify polymorphisms that increase the aggressiveness of cancer and that would lead to shorter survival.

Apoptosis or programmed cell death is a characteristic of normal cells while the key features of cancer cells include immortality, potential to replicate incessantly and resistance to anti-growth signals[349]. They may also initiate angiogenesis and metastasise, that spread beyond the primary tumour and survive in distant tissues[349]. This occurs through DNA alterations[349]. The major risk factor for causing such alteration in lung cancer is the interindividual differences in metabolising tobacco smoke carcinogens and its active compounds, and repairing the DNA damage caused by it[349]. Cigarette smoking causes changes in DNA resulting in mutagenesis, therefore several DNA repair pathway play their crucial role in eliminating DNA adducts and restoring the genetic stability of the genome[351]. A 14-fold increase in lung cancer risk is associated with an average smoker[349]. DNA repair pathways include base excision repair (BER), nucleotide excision repair (NER) and mismatch repair (MMR) that function on small lesions, bulk lesions and replication errors, respectively[351]. The other DNA repair pathways include single and double strand DNA break repair[351].

The MMR system is activated at the post-replicative phase, rectifying errors that have evaded the DNA checks by DNA polymerase[352]. Therefore, an alteration in the system may lead to genetic instability and make the cell susceptible to mutagenic transformation[352]. The system involves interaction between proteins including hMSH2, hMSH3, hMSH6, hMLH1, hPMS2 and hMLH3[352]. DNA repair by MMR is carried out by the complex interplay of the above components[352]. Various combinations of these components bind to mismatched nucleotides; single and larger insertion/deletion variants and loops, to allow for DNA repair[352].

The NER repair mechanism takes place in the following way. *XPC* and *hHR23B* form a complex to initiate the repair[353]. The damaged site is excised by a *TFIIH* complex that includes the helicases *XPB* and *XPD*[353]. These helicases also regulate strand separation at the damaged site [353]. In the NER reaction, damage is confirmed by *XPA*, which senses an open DNA conformation that is crucial for the repair mechanism[353]. The opened DNA complex is stabilised by replication protein A (RPA) and permits the positioning of *XPG*[353]. Excision of DNA at the 5' end of the lesion is carried out by an endonuclease complex formed by *ERCC1* and *XPF*[353]. The damaged site is removed and the void filled by replication factor, thus completing the repair process[353].

The BER pathways occurs through the participation of DNA glycolases, apurinic/apyrimidinic endonuclease (*APE1*), polymerase β (polβ), DNA ligase-III – *XRCC1* complex which are involved in cleaving the bond to the damaged nucleotide, cleaving the sugar phosphate at the 5' side and adds nucleotides from the 3' side, interact with polβ and completes the repair process by repair patch, respectively[354]. But for longer repairs, the above pathway changes following the action of polβ[354]. Polδ/ε adds a few more bases at the 3' end producing a flap, which is removed by flap endonuclease 1 (*FEN1*) added by proliferating cell nuclear antigen (*PCNA*)[354]. Finally, DNA ligase I completes the repair[354].

In lung cancer cells however, polymorphisms in the above pathway components ('MMR', 'NER', 'BER') causes the repair mechanisms to falter thus leading to the proliferative growth of cancer cells[351]. Hence, the result that SNPs within these pathways may contribute to lung cancer survival is not unfounded.

The above analysis could be continued to identify the important variables (may it be SNPs or clinical variables) that contribute to lung cancer survival[296]. To obtain the above information the samples in the OOB are used and the most important variable is obtained from a measure called the VIMP (variable importance) by subtracting the value of the prediction error of the new ensemble from the old[296]. Therefore, variables that have a high VIMP measure across the sampled survival trees could be studied and used to further evaluate and improve survival of lung cancer patients.

Though the significant pathways from this analysis are explained in terms of their role in carcinogenesis, their significance in survival analysis indicate that the extent of alteration may have an effect on survival. The significant pathways can be used to develop targeted therapy. For instance, there are drugs available that are directed towards signalling pathways[355]. Other drugs that are designed around processes in cancer such as angiogenesis and anti-apoptosis are also available[356].

To the best of our knowledge, this was the first survival pathway analysis conducted in lung cancer. Due to it being computationally intensive and time consuming, it has only been applied to 18 pathways, but could have been used to identify significant pathways associated with lung cancer from all the pathways deposited in the KEGG database. Though KEGG database has been manually curated, other pathway databases such as GO could be used to collate information towards forming a more complete and accurate pathway (section 5.2.1).

The next step would therefore not only be to test all the KEGG and pathways from other databases for significance in lung cancer survival for all split rules, but also to replicate and validate the results using different datasets.

CHAPTER 6

CONCLUSION

The primary aim of the project was to evaluate the epidemiological and biological (genetic) factors associated with lung cancer, in the Liverpool population. Epidemiological analyses were conducted to identify significant comorbidity indices (Charlson comorbidity index (CCI) and Elixhauser Comorbidity Index (ECI)), affecting the incidence of lung cancer patients. Pertinent data were obtained from the Hospital Episode Statistics (HES) database, and both univariate and multivariate analyses using the Cox proportional hazard model were conducted. CCI and ECI were significant in the incidence analysis, suggesting that their use may contribute towards the identification of high-risk individuals. This is the first study to have used ECI and CCI indices to study the incidence of lung cancer.

As part of this study, a 5-year sex-specific incidence prediction model was developed and internally validated using a 10 fold cross validation (AUC_{male} = 0.73; AUC_{female} = 0.77). The model was developed using the Cox proportional hazard regression model using age at the start of the study, chronic pulmonary conditions and smoking pack years as covariates. A point-based risk estimate system was developed; this is the first of its kind for lung cancer. The model, though internally validated, requires further validation in another dataset.

SNPs that could increase the susceptibility of lung cancer were identified using a genome wide association analysis (GWAS) of lung cancer cases from Liverpool and a control dataset from the 1958 Birth Cohort (Chapter 3). Different genetic models were used to identify SNPs associated with lung cancer in the Liverpool population. Additive, dominant and genotypic inheritance models consistently identified SNPs within the genes *PRDM11*, *ZNF382* and *HMGA2*, whereas a recessive model identified SNPs in *ITIH2*.

This study was limited by the unavailability for adjustment in the model of important covariates (such as age and smoking pack years) for the control dataset. Therefore, the identified SNPs require validation and testing in a multivariate model after adjusting for important covariates. Also, this analysis could be taken forward by conducting pathway

analysis to identify significant biological pathways associated with lung cancer susceptibility.

Genome wide survival analysis was conducted on 185 LLP NSCLC cases (Chapter 4). Multivariable Cox proportional hazard regression analysis identified single nucleotide polymorphisms (SNPs) associated with cause-specific and overall analysis. These were rs10230420 (*WIPF3*), rs9949512 (*SALL3*) and rs2139133 (*MYO3B*) for cause specific analysis, and rs10230420 (*WIPF3*) for overall survival analysis.

Genome wide survival analyses were also carried out on early stage and advanced stage NSCLCs, separately. Multivariable Cox proportional regression analysis after adjusting for age at diagnosis, stage, cell type and smoking pack years identified SNPs associated with cause-specific and overall survival. Significant SNPs associated with cause-specific analysis of early stage cases were rs10230420 (*WIPF3*), rs3746619 (*MC3R*) and rs3827103 (*MC3R*). In advanced stage cases, significant SNPs were rs1868110 (*NEK10*) and rs2206779 (*AF357533*). For the overall survival analysis, significant SNPs were rs10230420 (*WIPF3*), rs2056533 (*ZBTB20*) and rs6708630 (*CYS1*) in early stage cases; and while rs1868110 (*NEK10*) and rs2206779 (*AF357533*) in advanced stage NSCLC cases. The identified SNPs require replication and validation in another dataset.

These SNPs will add to the currently known SNPs associated with lung cancer survival. The previously identified SNPs were not identified in more than one publication (Table 4.1a and Table 4.1b), suggesting little overlap, and therefore much potential for further discoveries about the genetics of survival in lung cancer.

This dataset (185 NSCLC cases) was also used to carry out a pathway analysis using the random survival forest technique with various split rules such as log-rank, log-rank with bivariate, conserve and log-rank score (Chapter 5). Eighteen pathways closely related to

lung cancer were selected to test their importance for lung cancer survival, for both cause-specific and overall survival outcomes.

This analysis identified pathways such as apoptosis, cell cycle, BER, NER and MMR for both cause specific ad overall survival analysis that were significant ($p \leq 0.05$) and accurate in outcome prediction (AUC ≥ 0.7). The flexibility of this analysis with regard to the various methodologies employing the different split rules makes it a robust technique in survival pathway analysis.

This research is the first pathway-based survival analysis in lung cancer. Though the analysis was conducted on 18 pathways, this technique of identifying significant pathways could be applied to all pathways in the KEGG database. The pathways identified have to be replicated and validated in a separate lung cancer population. Significant SNPs, genes and pathways identified through survival analysis have potential application in cancer therapy. As such, genetic profiling could be used in the development of personalised therapy and also to improve patient management, though much work still needs to be done in this field.

This PhD project has contributed to lung cancer research in more than one way. It has evaluated comorbidity indices in incidence and survival, developed and internally validated a 5-year sex-specific incidence model and developed a point based risk estimation system. Through GWAS, it has identified plausible SNPs that could increase the lung cancer risk and those that can predict the occurrence of lung cancer and overall survival of lung cancer patients. Significant pathways associated with cause-specific and overall survival in lung cancer were also identified, using different split rules within the random survival forest framework.

REFERENCES

1. Ferlay, J., et al., *Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008.* International Journal of Cancer, 2010. **127**(12): p. 2893-2917.
2. Network, M.a.C.C., *UK Lung Cancer Coalition Commissioning Communications Toolkit Supporting clinicians to engage with and strengthen lung cancer commissioning.*
3. Lozano, R., et al., *Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.* Lancet, 2012. **380**(9859): p. 2095-128.
4. Indicators, C.o.C.H., *Compendium of Clinical & Health Indicators.* 2008.
5. Gabrielson, E., *Worldwide trends in lung cancer pathology.* Respirology, 2006. **11**(5): p. 533-8.
6. Sun, S., J.H. Schiller, and A.F. Gazdar, *Lung cancer in never smokers - a different disease.* Nature Reviews Cancer, 2007. **7**: p. 778-790.
7. Paone, G., et al., *DISCRIMINANT-ANALYSIS ON SMALL-CELL LUNG-CANCER AND NONSMALL CELL LUNG-CANCER BY MEANS OF NSE AND CYFRA-21.1.* European Respiratory Journal, 1995. **8**(7): p. 1136-1140.
8. Travis, W.D., E. Brambilla, and G.J. Riely, *New pathologic classification of lung cancer: relevance for clinical practice and clinical trials.* J Clin Oncol, 2013. **31**(8): p. 992-1001.
9. Dela Cruz, C.S., L.T. Tanoue, and R.A. Matthay, *Lung cancer: epidemiology, etiology, and prevention.* Clin Chest Med, 2011. **32**(4): p. 605-44.
10. Walser, T., et al., *Smoking and lung cancer: the role of inflammation.* Proc Am Thorac Soc, 2008. **5**(8): p. 811-5.
11. Engels, E.A., *Inflammation in the development of lung cancer: epidemiological evidence.* Expert Rev Anticancer Ther, 2008. **8**(4): p. 605-15.
12. Rooney, C. and T. Sethi, *The Epithelial Cell and Lung Cancer: The Link between Chronic Obstructive Pulmonary Disease and Lung Cancer.* Respiration, 2011. **81**(2): p. 89-104.
13. Yang, I.A., et al., *Common pathogenic mechanisms and pathways in the development of COPD and lung cancer.* Expert Opinion on Therapeutic Targets, 2011. **15**(4): p. 439-456.
14. Yokota, J., K. Shiraishi, and T. Kohno, *Genetic basis for susceptibility to lung cancer: Recent progress and future directions.* Adv Cancer Res, 2010. **109**: p. 51-72.
15. Auerbach, O., et al., *Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer.* N Engl J Med, 1961. **265**: p. 253-67.
16. Lee, G., T.C. Walser, and S.M. Dubinett, *Chronic inflammation, chronic obstructive pulmonary disease, and lung cancer.* Curr Opin Pulm Med, 2009. **15**(4): p. 303-7.
17. Larsen, J.E. and J.D. Minna, *Molecular biology of lung cancer: clinical implications.* Clin Chest Med, 2011. **32**(4): p. 703-40.
18. Brenner, D.R., et al., *Previous lung diseases and lung cancer risk: a pooled analysis from the International Lung Cancer Consortium.* Am J Epidemiol, 2012. **176**(7): p. 573-85.
19. Hosseini, M., et al., *Environmental risk factors for lung cancer in Iran: a case-control study.* International Journal of Epidemiology, 2009. **38**(4): p. 989-996.
20. Brenner, D.R., et al., *Lung cancer risk in never-smokers: a population-based case-control study of epidemiologic risk factors.* BMC Cancer, 2010. **10**: p. 285.
21. Gao, Y., et al., *Family history of cancer and nonmalignant lung diseases as risk factors for lung cancer.* Int J Cancer, 2009. **125**(1): p. 146-52.

22. Mahabir, S., et al., *Dietary magnesium and DNA repair capacity as risk factors for lung cancer.* Carcinogenesis, 2008. **29**(5): p. 949-956.
23. Mahabir, S., et al., *Dietary boron and hormone replacement therapy as risk factors for lung cancer in women.* American Journal of Epidemiology, 2008. **167**(9): p. 1070-1080.
24. Neuberger, J.S., et al., *Risk factors for lung cancer in Iowa women: Implications for prevention.* Cancer Detection and Prevention, 2006. **30**(2): p. 158-167.
25. Kreuzer, M., et al., *Hormonal factors and risk of lung cancer among women?* International Journal of Epidemiology, 2003. **32**(2): p. 263-271.
26. Takezaki, T., et al., *Dietary factors and lung cancer risk in Japanese: with special reference to fish consumption and adenocarcinomas.* British Journal of Cancer, 2001. **84**(9): p. 1199-1206.
27. Martin, J.C., et al., *Occupational risk factors for lung cancer in the French electricity and gas industry - A case-control survey nested in a cohort of active employees.* American Journal of Epidemiology, 2000. **151**(9): p. 902-912.
28. Straif, K., et al., *Occupational risk factors for mortality from stomach and lung cancer among rubber workers: an analysis using internal controls and refined exposure assessment.* International Journal of Epidemiology, 1999. **28**(6): p. 1037-1043.
29. Droste, J.H.J., et al., *Occupational risk factors of lung cancer: a hospital based case-control study.* Occupational and Environmental Medicine, 1999. **56**(5): p. 322-327.
30. Jockel, K.H., et al., *Occupational risk factors for lung cancer: a case-control study in West Germany.* International Journal of Epidemiology, 1998. **27**(4): p. 549-560.
31. Ko, Y.C., et al., *Risk factors for primary lung cancer among non-smoking women in Taiwan.* International Journal of Epidemiology, 1997. **26**(1): p. 24-31.
32. Benhamou, S., E. Benhamou, and R. Flamant, *Occupational Risk-Factors of Lung-Cancer in a French Case-Control Study.* British Journal of Industrial Medicine, 1988. **45**(4): p. 231-233.
33. Buiatti, E., et al., *A Case Control Study of Lung-Cancer in Florence, Italy .1. Occupational Risk-Factors.* Journal of Epidemiology and Community Health, 1985. **39**(3): p. 244-250.
34. Wingo, P.A., et al., *Long-term trends in cancer mortality in the United States, 1930-1998.* Cancer, 2003. **97**(12 Suppl): p. 3133-275.
35. Wingo, P.A., et al., *Annual report to the nation on the status of cancer, 1973-1996, with a special section on lung cancer and tobacco smoking.* Journal of the National Cancer Institute, 1999. **91**(8): p. 675-690.
36. Weir, H.K., et al., *Annual report to the nation on the status of cancer, 1975-2000, featuring the uses of surveillance data for cancer prevention and control.* J Natl Cancer Inst, 2003. **95**(17): p. 1276-99.
37. Thomas, L., L.A. Doyle, and M.J. Edelman, *Lung cancer in women: emerging differences in epidemiology, biology, and therapy.* Chest, 2005. **128**(1): p. 370-81.
38. Jemal, A., et al., *Global cancer statistics.* CA Cancer J Clin, 2011. **61**(2): p. 69-90.
39. Ryberg, D., et al., *Different susceptibility to smoking-induced DNA damage among male and female lung cancer patients.* Cancer Res, 1994. **54**(22): p. 5801-3.
40. Taioli, E. and E.L. Wynder, *Re: Endocrine factors and adenocarcinoma of the lung in women.* J Natl Cancer Inst, 1994. **86**(11): p. 869-70.
41. Lissowska, J., et al., *Family history and lung cancer risk: international multicentre case-control study in Eastern and Central Europe and meta-analyses.* Cancer Causes Control, 2010. **21**(7): p. 1091-104.
42. Tokuhata, G.K. and A.M. Lilienfeld, *Familial aggregation of lung cancer in humans.* J Natl Cancer Inst, 1963. **30**: p. 289-312.

43. Takemiya, M., et al., *Bloom's syndrome with porokeratosis of Mibelli and multiple cancers of the skin, lung and colon.* Clin Genet, 1987. **31**(1): p. 35-44.
44. Yamanaka, A., et al., *Lung cancer associated with Werner's syndrome: a case report and review of the literature.* Jpn J Clin Oncol, 1997. **27**(6): p. 415-8.
45. Matakidou, A., T. Eisen, and R.S. Houlston, *Systematic review of the relationship between family history and lung cancer risk.* Br J Cancer, 2005. **93**(7): p. 825-833.
46. Bailey-Wilson, J.E., et al., *A major lung cancer susceptibility locus maps to chromosome 6q23-25.* American journal of human genetics, 2004. **75**(3): p. 460-474.
47. You, M., et al., *Fine Mapping of Chromosome 6q23-25 Region in Familial Lung Cancer Families Reveals RGS17 as a Likely Candidate Gene.* Clinical Cancer Research, 2009. **15**(8): p. 2666-2674.
48. Spitz, M.R., et al., *A risk model for prediction of lung cancer.* J Natl Cancer Inst, 2007. **99**(9): p. 715-26.
49. Spitz, M.R., et al., *An expanded risk prediction model for lung cancer.* Cancer Prev Res (Phila), 2008. **1**(4): p. 250-4.
50. Cassidy, A., et al., *The LLP risk model: an individual risk prediction model for lung cancer.* Br J Cancer, 2008. **98**(2): p. 270-6.
51. Smith, C.J., et al., *"IARC Group 2B carcinogens" reported in cigarette mainstream smoke.* Food Chem Toxicol, 2001. **39**(2): p. 183-205.
52. Smith, C.J., et al., *"IARC group 2B Carcinogens" reported in cigarette mainstream smoke.* Food Chem Toxicol, 2000. **38**(9): p. 825-48.
53. Hoffmann, D., et al., *A study of tobacco carcinogenesis. LI. Relative potencies of tobacco-specific N-nitrosamines as inducers of lung tumours in A/J mice.* Cancer Lett, 1993. **71**(1-3): p. 25-30.
54. Belinsky, S.A., et al., *Relationship between the formation of promutagenic adducts and the activation of the K-ras protooncogene in lung tumors from A/J mice treated with nitrosamines.* Cancer Res, 1989. **49**(19): p. 5305-11.
55. Rodenhuis, S. and R.J. Slebos, *Clinical significance of ras oncogene activation in human lung cancer.* Cancer Res, 1992. **52**(9 Suppl): p. 2665s-2669s.
56. Westra, W.H., et al., *K-ras oncogene activation in lung adenocarcinomas from former smokers. Evidence that K-ras mutations are an early and irreversible event in the development of adenocarcinoma of the lung.* Cancer, 1993. **72**(2): p. 432-8.
57. Denissenko, M.F., et al., *Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53.* Science, 1996. **274**(5286): p. 430-2.
58. Arcavi, L. and N.L. Benowitz, *Cigarette smoking and infection.* Archives of Internal Medicine, 2004. **164**(20): p. 2206-2216.
59. Cihak, R.W., *RADIATION AND LUNG CANCER.* Human Pathology, 1971. **2**(4): p. 525-528.
60. Grosche, B., et al., *Lung cancer risk among German male uranium miners: a cohort study, 1946-1998.* Br J Cancer, 2006. **95**(9): p. 1280-7.
61. Wagoner, J.K., et al., *Radiation as the Cause of Lung Cancer among Uranium Miners.* N Engl J Med, 1965. **273**: p. 181-8.
62. Lubin, J.H. and J.D. Boice, Jr., *Lung cancer risk from residential radon: meta-analysis of eight epidemiologic studies.* J Natl Cancer Inst, 1997. **89**(1): p. 49-57.
63. Hei, T.K., et al., *Malignant transformation of human bronchial epithelial cells by radon-simulated alpha-particles.* Carcinogenesis, 1994. **15**(3): p. 431-7.
64. Doll, R., *Mortality from lung cancer in asbestos workers.* Br J Ind Med, 1955. **12**(2): p. 81-6.
65. Suvatne, J. and R.F. Browning, *Asbestos and Lung Cancer.* Dm Disease-a-Month, 2011. **57**(1): p. 55-68.

66. Weiss, W., *Asbestosis: a marker for the increased risk of lung cancer among workers exposed to asbestos.* Chest, 1999. **115**(2): p. 536-49.
67. Churg, A. and B. Wiggs, *Fiber size and number in amphibole asbestos-induced mesothelioma.* Am J Pathol, 1984. **115**(3): p. 437-42.
68. Mossman, B.T. and A. Churg, *Mechanisms in the pathogenesis of asbestosis and silicosis.* Am J Respir Crit Care Med, 1998. **157**(5 Pt 1): p. 1666-80.
69. Jones, R.N., J.M. Hughes, and H. Weill, *Asbestos exposure, asbestosis, and asbestos-attributable lung cancer.* Thorax, 1996. **51 Suppl 2**: p. S9-15.
70. Hughes, J.M. and H. Weill, *Asbestosis as a precursor of asbestos related lung cancer: results of a prospective mortality study.* Br J Ind Med, 1991. **48**(4): p. 229-33.
71. Wu, C.Y., et al., *Pulmonary Tuberculosis Increases the Risk of Lung Cancer.* Cancer, 2011. **117**(3): p. 618-624.
72. Engels, E.A., et al., *Tuberculosis and subsequent risk of lung cancer in Xuanwei, China.* International Journal of Cancer, 2009. **124**(5): p. 1183-1187.
73. Zhan, P., et al., *Chlamydia pneumoniae infection and lung cancer risk: A meta-analysis.* European Journal of Cancer, 2011. **47**(5): p. 742-747.
74. Koshiol, J., et al., *Lower Risk of Lung Cancer after Multiple Pneumonia Diagnoses.* Cancer Epidemiology Biomarkers & Prevention, 2010. **19**(3): p. 716-721.
75. Adcock, I.M., G. Caramori, and P.J. Barnes, *Chronic Obstructive Pulmonary Disease and Lung Cancer: New Molecular Insights.* Respiration, 2011. **81**(4): p. 265-284.
76. Kishi, K., et al., *The correlation of emphysema or airway obstruction with the risk of lung cancer: a matched case-controlled study.* European Respiratory Journal, 2002. **19**(6): p. 1093-1098.
77. Wilson, D.O., et al., *The Pittsburgh Lung Screening Study (PLuSS) Outcomes within 3 Years of a First Computed Tomography Scan.* American Journal of Respiratory and Critical Care Medicine, 2008. **178**(9): p. 956-961.
78. Mao, Y., et al., *Socioeconomic status and lung cancer risk in Canada.* International Journal of Epidemiology, 2001. **30**(4): p. 809-817.
79. Vanloon, A.J.M., R.A. Goldbohm, and P.A. Vandenbrandt, *LUNG-CANCER - IS THERE AN ASSOCIATION WITH SOCIOECONOMIC-STATUS IN THE NETHERLANDS.* Journal of Epidemiology and Community Health, 1995. **49**(1): p. 65-69.
80. Shack, L., et al., *Variation in incidence of breast, lung and cervical cancer and malignant melanoma of skin by socioeconomic group in England.* BMC Cancer, 2008. **8**: p. 271.
81. Osada, H. and T. Takahashi, *Genetic alterations of multiple tumor suppressors and oncogenes in the carcinogenesis and progression of lung cancer.* Oncogene, 2002. **21**(48): p. 7421-34.
82. Risch, A. and C. Plass, *Lung cancer epigenetics and genetics.* International Journal of Cancer, 2008. **123**(1): p. 1-7.
83. Knudson, A.G., Jr., *Mutation and cancer: statistical study of retinoblastoma.* Proc Natl Acad Sci U S A, 1971. **68**(4): p. 820-3.
84. Kohno, T. and J. Yokota, *How many tumor suppressor genes are involved in human lung carcinogenesis?* Carcinogenesis, 1999. **20**(8): p. 1403-10.
85. Sato, M., et al., *A translational view of the molecular pathogenesis of lung cancer.* Journal of Thoracic Oncology, 2007. **2**: p. 327-343.
86. Salgia, R. and A.T. Skarin, *Molecular abnormalities in lung cancer.* J Clin Oncol, 1998. **16**(3): p. 1207-17.
87. Sekido, Y., K.M. Fong, and J.D. Minna, *Molecular genetics of lung cancer.* Annu Rev Med, 2003. **54**: p. 73-87.

88. Dong, L.M., et al., *Genetic susceptibility to cancer: the role of polymorphisms in candidate genes.* JAMA, 2008. **299**(20): p. 2423-36.
89. Ye, Z., et al., *Five glutathione s-transferase gene variants in 23,452 cases of lung cancer and 30,397 controls: meta-analysis of 130 studies.* PLoS Med, 2006. **3**(4): p. e91.
90. Li, H., et al., *The hOGG1 Ser326Cys polymorphism and lung cancer risk: a meta-analysis.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(7): p. 1739-45.
91. Wilkening, S., J.L. Bermejo, and K. Hemminki, *MDM2 SNP309 and cancer risk: a combined analysis.* Carcinogenesis, 2007. **28**(11): p. 2262-7.
92. Dai, S., et al., *P53 polymorphism and lung cancer susceptibility: a pooled analysis of 32 case-control studies.* Hum Genet, 2009. **125**(5-6): p. 633-8.
93. Esquela-Kerscher, A. and F.J. Slack, *Oncomirs - microRNAs with a role in cancer.* Nature Reviews Cancer, 2006. **6**(4): p. 259-269.
94. Izzotti, A., et al., *Relationships of microRNA expression in mouse lung with age and exposure to cigarette smoke and light.* Faseb Journal, 2009. **23**(9): p. 3243-3250.
95. Yu, L., et al., *Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers.* Int J Cancer, 2010. **127**(12): p. 2870-8.
96. Yu, S.L., et al., *MicroRNA signature predicts survival and relapse in lung cancer.* Cancer Cell, 2008. **13**(1): p. 48-57.
97. Sozzi, G., U. Pastorino, and C.M. Croce, *MicroRNAs and lung cancer: from markers to targets.* Cell Cycle, 2011. **10**(13): p. 2045-6.
98. Field, J.K., et al., *Prospects for population screening and diagnosis of lung cancer.* Lancet, 2013. **382**(9893): p. 732-41.
99. Statistics, O.f.N., *Cancer survival in England: patients diagnosed 2005-2009 and followed up to 2010.* London: Office for National Statistics, 2011.
100. Hirsch, F.R., et al., *Early detection of lung cancer: clinical perspectives of recent advances in biology and radiology.* Clin Cancer Res, 2001. **7**(1): p. 5-22.
101. Thunnissen, F., *Sputum examination for early detection of lung cancer.* Journal of Clinical Pathology, 2003. **56**(11): p. 805-810.
102. van't Westeinde, S.C. and R.J. van Klaveren, *Screening and Early Detection of Lung Cancer.* Cancer Journal, 2011. **17**(1): p. 3-10.
103. Baldwin, D.R., et al., *UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer.* Thorax, 2011. **66**(4): p. 308-13.
104. National Lung Screening Trial Research, T., et al., *Reduced lung-cancer mortality with low-dose computed tomographic screening.* N Engl J Med, 2011. **365**(5): p. 395-409.
105. Yasufuku, K., *Early Diagnosis of Lung Cancer.* Clinics in Chest Medicine, 2010. **31**(1): p. 39-47.
106. Patel, D., et al., *Attitudes to participation in a lung cancer screening trial: a qualitative study.* Thorax, 2012. **67**(5): p. 418-25.
107. International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
108. Ross, J.S. and M. Cronin, *Whole cancer genome sequencing by next-generation methods.* Am J Clin Pathol, 2011. **136**(4): p. 527-39.
109. Kiri, V.A., et al., *Recent trends in lung cancer and its association with COPD: an analysis using the UK GP Research Database.* Prim Care Respir J, 2010. **19**(1): p. 57-61.
110. Petty, R.D., et al., *Gene expression profiling in non-small cell lung cancer: From molecular mechanisms to clinical application.* Clinical Cancer Research, 2004. **10**(10): p. 3237-3248.

111. Amos, C.I., et al., *Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1.* Nature Genetics, 2008. **40**(5): p. 616-622.
112. Thorgeirsson, T.E., et al., *A variant associated with nicotine dependence, lung cancer and peripheral arterial disease.* Nature, 2008. **452**(7187): p. 638-U9.
113. Hung, R.J., et al., *A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25.* Nature, 2008. **452**(7187): p. 633-7.
114. Amos, C.I., et al., *Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1.* Nat Genet, 2008. **40**(5): p. 616-22.
115. Brennan, P., P. Hainaut, and P. Boffetta, *Genetics of lung-cancer susceptibility.* Lancet Oncology, 2011. **12**(4): p. 399-408.
116. Taylor, J.M., D.P. Ankerst, and R.R. Andridge, *Validation of biomarker-based risk prediction models.* Clin Cancer Res, 2008. **14**(19): p. 5977-83.
117. Gail, M.H., et al., *Projecting individualized probabilities of developing breast cancer for white females who are being examined annually.* J Natl Cancer Inst, 1989. **81**(24): p. 1879-86.
118. Selvachandran, S.N., et al., *Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study.* Lancet, 2002. **360**(9329): p. 278-83.
119. Thompson, I.M., et al., *Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial.* J Natl Cancer Inst, 2006. **98**(8): p. 529-34.
120. Kattan, M.W., et al., *A preoperative nomogram for disease recurrence following radical prostatectomy for prostate cancer.* J Natl Cancer Inst, 1998. **90**(10): p. 766-71.
121. Skates, S.J., D.K. Pauler, and I.J. Jacobs, *Screening based on the risk of cancer calculation from Bayesian hierarchical changepoint and mixture models of longitudinal markers.* Journal of the American Statistical Association, 2001. **96**: p. 429-439.
122. Paik, S., et al., *A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* N Engl J Med, 2004. **351**(27): p. 2817-26.
123. Etzel, C.J., et al., *Development and validation of a lung cancer risk prediction model for African-Americans.* Cancer Prev Res (Phila), 2008. **1**(4): p. 255-65.
124. Bach, P.B., et al., *Variations in Lung Cancer Risk Among Smokers.* Journal of the National Cancer Institute, 2003. **95**(6): p. 470-478.
125. ONS, *Office for National Statistics, Cancer registrations in England* 2010.
126. de Groot, V., et al., *How to measure comorbidity: a critical review of available methods.* Journal of Clinical Epidemiology, 2003. **56**(3): p. 221-229.
127. Geraci, J.M., et al., *Comorbid disease and cancer: The need for more relevant conceptual models in health services research.* Journal of Clinical Oncology, 2005. **23**(30): p. 7399-7404.
128. Ogle, K.S., et al., *Cancer and comorbidity - Redefining chronic diseases.* Cancer, 2000. **88**(3): p. 653-663.
129. Shieh, S.H., et al., *Decreased survival among lung cancer patients with co-morbid tuberculosis and diabetes.* BMC Cancer, 2012. **12**: p. 174.
130. Smith, L., et al., *Body mass index and risk of lung cancer among never, former, and current smokers.* J Natl Cancer Inst, 2012. **104**(10): p. 778-89.
131. Punturieri, A., et al., *Lung cancer and chronic obstructive pulmonary disease: needs and opportunities for integrated research.* J Natl Cancer Inst, 2009. **101**(8): p. 554-9.
132. Mannino, D.M., et al., *Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD.* European Respiratory Journal, 2008. **32**(4): p. 962-969.

133. Hsieh, M.C., et al., *The influence of type 2 diabetes and glucose-lowering therapies on cancer risk in the Taiwanese.* Exp Diabetes Res, 2012. **2012**: p. 413782.
134. Luo, J., et al., *Diabetes and lung cancer among postmenopausal women.* Diabetes Care, 2012. **35**(7): p. 1485-91.
135. Maddams, J., M. Utley, and H. Moller, *A person-time analysis of hospital activity among cancer survivors in England.* Br J Cancer, 2011. **105 Suppl 1**: p. S38-45.
136. Millett, E.R., et al., *Incidence of community-acquired lower respiratory tract infections and pneumonia among older adults in the United Kingdom: a population-based study.* PLoS One, 2013. **8**(9): p. e75131.
137. Hippisley-Cox, J. and C. Coupland, *Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score.* BMJ Open, 2013. **3**(8): p. e003482.
138. Tsang, C., et al., *Cancer diagnosed by emergency admission in England: an observational study using the general practice research database.* BMC Health Serv Res, 2013. **13**: p. 308.
139. Wright, F.L., et al., *Vascular disease in women: comparison of diagnoses in hospital episode statistics and general practice records in England.* BMC Med Res Methodol, 2012. **12**: p. 161.
140. Aylin, P., A. Bottle, and A. Majeed, *Use of administrative data or clinical databases as predictors of risk of death in hospital: comparison of models.* BMJ, 2007. **334**(7602): p. 1044.
141. Extermann, M., *Measuring comorbidity in older cancer patients.* European Journal of Cancer, 2000. **36**(4): p. 453-471.
142. Yancik, R., et al., *Cancer and comorbidity in older patients: A descriptive profile.* Annals of Epidemiology, 1996. **6**(5): p. 399-412.
143. Janssen-Heijnen, M.L., et al., *Comorbidity in older surgical cancer patients: influence on patient care and outcome.* Eur J Cancer, 2007. **43**(15): p. 2179-93.
144. Janssen-Heijnen, M.L., et al., *Effect of comorbidity on the treatment and prognosis of elderly patients with non-small cell lung cancer.* Thorax, 2004. **59**(7): p. 602-7.
145. Jorgensen, T.L., et al., *Comorbidity in elderly cancer patients in relation to overall and cancer-specific mortality.* Br J Cancer, 2012. **106**(7): p. 1353-60.
146. Ganti, A.K., et al., *Predictive Ability of Charlson Comorbidity Index on Outcomes From Lung Cancer.* Am J Clin Oncol, 2011.
147. Fleming, S.T., et al., *A comprehensive prognostic index to predict survival based on multiple comorbidities: a focus on breast cancer.* Med Care, 1999. **37**(6): p. 601-14.
148. Pujol, J.L., et al., *A new simplified comorbidity score as a prognostic factor in non-small-cell lung cancer patients: description and comparison with the Charlson's index.* British Journal of Cancer, 2005. **93**(10): p. 1098-1105.
149. Elixhauser, A., et al., *Comorbidity measures for use with administrative data.* Med Care, 1998. **36**(1): p. 8-27.
150. Lieffers, J.R., et al., *A comparison of charlson and elixhauser comorbidity measures to predict colorectal cancer survival using administrative health data.* Cancer, 2010.
151. Charlson, M.E., et al., *A new method of classifying prognostic comorbidity in longitudinal studies: development and validation.* J Chronic Dis, 1987. **40**(5): p. 373-83.
152. Birim, O., A.P. Kappetein, and A.J. Bogers, *Charlson comorbidity index as a predictor of long-term outcome after surgery for nonsmall cell lung cancer.* Eur J Cardiothorac Surg, 2005. **28**(5): p. 759-62.
153. Wang, T.J., et al., *Multiple Biomarkers for the Prediction of First Major Cardiovascular Events and Death.* New England Journal of Medicine, 2006. **355**(25): p. 2631-2639.

154. Wang, C.Y., et al., *Comparison of Charlson comorbidity index and Kaplan-Feinstein index in patients with stage I lung cancer after surgical resection.* Eur J Cardiothorac Surg, 2007. **32**(6): p. 877-81.
155. Do, S.Y., D.A. Bush, and J.D. Slater, *Comorbidity-adjusted survival in early stage lung cancer patients treated with hypofractionated proton therapy.* J Oncol, 2010. **2010**: p. 251208.
156. Firat, S., et al., *Comorbidity and KPS are independent prognostic factors in stage I non-small-cell lung cancer.* Int J Radiat Oncol Biol Phys, 2002. **52**(4): p. 1047-57.
157. Sanchez, P.G., et al., *Lobectomy for treating bronchial carcinoma: analysis of comorbidities and their impact on postoperative morbidity and mortality.* J Bras Pneumol, 2006. **32**(6): p. 495-504.
158. Liu, C.T., et al., *Impact of comorbidity on survival for locally advanced head and neck cancer patients treated by radiotherapy or radiotherapy plus chemotherapy.* Chang Gung Med J, 2010. **33**(3): p. 283-91.
159. Singh, B., et al., *Validation of the Charlson comorbidity index in patients with head and neck cancer: a multi-institutional study.* Laryngoscope, 1997. **107**(11 Pt 1): p. 1469-75.
160. Breccia, M., et al., *Charlson comorbidity index and adult comorbidity evaluation-27 might predict compliance and development of pleural effusions in elderly chronic myeloid leukemia patients treated with dasatinib after resistance/intolerance to imatinib.* Haematologica, 2011.
161. Hines, R.B., et al., *Predictive capacity of three comorbidity indices in estimating mortality after surgery for colon cancer.* J Clin Oncol, 2009. **27**(26): p. 4339-45.
162. Koppie, T.M., et al., *Age-adjusted Charlson comorbidity score is associated with treatment decisions and clinical outcomes for patients undergoing radical cystectomy for bladder cancer.* Cancer, 2008. **112**(11): p. 2384-92.
163. Miller, D.C., et al., *The impact of co-morbid disease on cancer control and survival following radical cystectomy.* J Urol, 2003. **169**(1): p. 105-9.
164. Gore, J.L., et al., *Use of radical cystectomy for patients with invasive bladder cancer.* J Natl Cancer Inst, 2010. **102**(11): p. 802-11.
165. Fisher, M.B., et al., *Cardiac history and risk of post-cystectomy cardiac complications.* Urology, 2009. **74**(5): p. 1085-9.
166. Gettman, M.T., et al., *Charlson co-morbidity index as a predictor of outcome after surgery for renal cell carcinoma with renal vein, vena cava or right atrium extension.* J Urol, 2003. **169**(4): p. 1282-6.
167. Tetsche, M.S., et al., *The impact of comorbidity and stage on ovarian cancer mortality: a nationwide Danish cohort study.* BMC Cancer, 2008. **8**: p. 31.
168. Wahlgren, T., et al., *Use of the Charlson Combined Comorbidity Index To Predict Postradiotherapy Quality of Life for Prostate Cancer Patients.* Int J Radiat Oncol Biol Phys, 2010.
169. Kastner, C., et al., *The Charlson comorbidity score: a superior comorbidity assessment tool for the prostate cancer multidisciplinary meeting.* Prostate Cancer Prostatic Dis, 2006. **9**(3): p. 270-4.
170. Alibhai, S.M., et al., *Is there an optimal comorbidity index for prostate cancer?* Cancer, 2008. **112**(5): p. 1043-50.
171. Pujol, J.L., et al., *Quality of life and comorbidity score as prognostic determinants in non-small-cell lung cancer patients.* Annals of Oncology, 2008. **19**(8): p. 1458-1464.
172. Moro-Sibilot, D., et al., *Comorbidities and Charlson score in resected stage I nonsmall cell lung cancer.* Eur Respir J, 2005. **26**(3): p. 480-6.

173. Birim, O., et al., *Validation of the Charlson comorbidity index in patients with operated primary non-small cell lung cancer.* Eur J Cardiothorac Surg, 2003. **23**(1): p. 30-4.
174. Froehner, M., et al., *Comparison of the American Society of Anesthesiologists Physical Status classification with the Charlson score as predictors of survival after radical prostatectomy.* Urology, 2003. **62**(4): p. 698-701.
175. Kleinbaum, D.G., M. Klein, and SpringerLink, *Survival analysis : a self-learning text*. 3rd ed. Statistics for biology and health,1431-8776. 2012, New York: Springer. xv, 700 p.
176. Sinha, S., et al., *Epidemiological study of provision of cholecystectomy in England from 2000 to 2009: retrospective analysis of Hospital Episode Statistics.* Surg Endosc, 2012.
177. Quan, H., et al., *Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data.* Med Care, 2005. **43**(11): p. 1130-9.
178. Etzel, C.J. and P.B. Bach, *Estimating Individual Risk for Lung Cancer.* Seminars in Respiratory and Critical Care Medicine, 2011. **32**(1): p. 3-9.
179. StataCorp, *Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.* 2011.
180. Ording, A.G., et al., *Hospital recorded morbidity and breast cancer incidence: a nationwide population-based case-control study.* PLoS One, 2012. **7**(10): p. e47329.
181. Cassidy, A., et al., *Lung cancer risk prediction: A tool for early detection.* International Journal of Cancer, 2007. **120**(1): p. 1-6.
182. Cronin, K.A., et al., *Validation of a model of lung cancer risk prediction among smokers.* J Natl Cancer Inst, 2006. **98**(9): p. 637-40.
183. D'Amelio, A.M., et al., *Comparison of discriminatory power and accuracy of three lung cancer risk models.* British Journal of Cancer, 2010. **103**(3): p. 423-429.
184. Raji, O.Y., et al., *Predictive Accuracy of the Liverpool Lung Project Risk Model for Stratifying Patients for Computed Tomography Screening for Lung Cancer A Case-Control and Cohort Validation Study.* Annals of Internal Medicine, 2012. **157**(4): p. 242-+.
185. Tammemagi, C.M., et al., *Lung cancer risk prediction: Prostate, Lung, Colorectal And Ovarian Cancer Screening Trial models and validation.* J Natl Cancer Inst, 2011. **103**(13): p. 1058-68.
186. Sullivan, L.M., J.M. Massaro, and R.B. D'Agostino, Sr., *Presentation of multivariate data for clinical use: The Framingham Study risk score functions.* Stat Med, 2004. **23**(10): p. 1631-60.
187. D'Agostino, R.B., Sr., et al., *General cardiovascular risk profile for use in primary care: the Framingham Heart Study.* Circulation, 2008. **117**(6): p. 743-53.
188. Cook, N.R., *Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve.* Clin Chem, 2008. **54**(1): p. 17-23.
189. Pencina, M.J., et al., *Predicting the 30-year risk of cardiovascular disease: the framingham heart study.* Circulation, 2009. **119**(24): p. 3078-84.
190. van Rens, M.T., et al., *Prognostic assessment of 2,361 patients who underwent pulmonary resection for non-small cell lung cancer, stage I, II, and IIIA.* Chest, 2000. **117**(2): p. 374-9.
191. Park, S., et al., *Individualized risk prediction model for lung cancer in Korean men.* PLoS One, 2013. **8**(2): p. e54823.
192. Smith, A.V., *Genetic analysis: moving between linkage and association.* Cold Spring Harb Protoc, 2012. **2012**(2): p. 174-82.
193. McKay, J.D., et al., *Lung cancer susceptibility locus at 5p15.33.* Nature Genetics, 2008. **40**(12): p. 1404-1406.

194. Power, C. and J. Elliott, *Cohort profile: 1958 British birth cohort (National Child Development Study).* Int J Epidemiol, 2006. **35**(1): p. 34-41.
195. Sanders, B.M., et al., *Non-ocular cancer in relatives of retinoblastoma patients.* Br J Cancer, 1989. **60**(3): p. 358-65.
196. Swift, M. and C. Chase, *Cancer in families with xeroderma pigmentosum.* J Natl Cancer Inst, 1979. **62**(6): p. 1415-21.
197. Hwang, S.J., et al., *Lung cancer risk in germline p53 mutation carriers: association between an inherited cancer predisposition, cigarette smoking, and cancer risk.* Hum Genet, 2003. **113**(3): p. 238-43.
198. Alexandrov, K., et al., *CYP1A1 and GSTM1 genotypes affect benzo[a]pyrene DNA adducts in smokers' lung: comparison with aromatic/hydrophobic adduct formation.* Carcinogenesis, 2002. **23**(12): p. 1969-77.
199. Le Calvez, F., et al., *TP53 and KRAS mutation load and types in lung cancers in relation to tobacco smoke: distinct patterns in never, former, and current smokers.* Cancer Res, 2005. **65**(12): p. 5076-83.
200. Schwartz, A.G., et al., *The molecular epidemiology of lung cancer.* Carcinogenesis, 2007. **28**: p. 507-518.
201. International HapMap, C., et al., *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.
202. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
203. Landi, M.T., et al., *A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma.* The American Journal of Human Genetics, 2009. **85**(5): p. 679-691.
204. Paige, A.J., *Redefining tumour suppressor genes: exceptions to the two-hit hypothesis.* Cell Mol Life Sci, 2003. **60**(10): p. 2147-63.
205. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
206. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
207. Chung, C.C., et al., *Genome-wide association studies in cancer--current and future directions.* Carcinogenesis, 2010. **31**(1): p. 111-20.
208. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.
209. Kim, S.Y., et al., *Estimation of allele frequency and association mapping using next-generation sequencing data.* BMC Bioinformatics, 2011. **12**: p. 231.
210. Cordell, H.J. and D.G. Clayton, *Genetic association studies.* Lancet, 2005. **366**(9491): p. 1121-31.
211. Klein, R.J., *Power analysis for genome-wide association studies.* BMC Genet, 2007. **8**: p. 58.
212. Seo, S., et al., *Functional analysis of deep intronic SNP rs13438494 in intron 24 of PCLO gene.* PLoS One, 2013. **8**(10): p. e76960.
213. Gray, I.C., D.A. Campbell, and N.K. Spurr, *Single nucleotide polymorphisms as tools in human genetics.* Hum Mol Genet, 2000. **9**(16): p. 2403-8.
214. Balding, D.J., *A tutorial on statistical methods for population association studies.* Nat Rev Genet, 2006. **7**(10): p. 781-91.
215. Yamamoto, K., et al., *A novel gene, CRR9, which was up-regulated in CDDP-resistant ovarian tumor cell line, was associated with apoptosis.* Biochem Biophys Res Commun, 2001. **280**(4): p. 1148-54.

216. Shiina, T., et al., *Genomic anatomy of a premier major histocompatibility complex paralogous region on chromosome 1q21-q22.* Genome Res, 2001. **11**(5): p. 789-802.
217. Crawley, M.J., *Statistics : an introduction using R*. 2005, Chichester. Chichester, West Sussex: John Wiley, John Wiley & Sons. xiii, 327 p.
218. Wellcome Trust Case Control, C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
219. Wacholder, S., et al., *Assessing the probability that a positive report is false: An approach for molecular epidemiology studies.* Journal of the National Cancer Institute, 2004. **96**(6): p. 434-442.
220. Falconer, D.S. and T.F.C. Mackay, *Introduction to quantitative genetics*. 4th ed. 1996, Harlow: Longman. xv,464p.
221. Wang, J. and S. Shete, *Testing departure from Hardy-Weinberg proportions.* Methods Mol Biol, 2012. **850**: p. 77-102.
222. Guo, S.W. and E.A. Thompson, *Performing the exact test of Hardy-Weinberg proportion for multiple alleles.* Biometrics, 1992. **48**(2): p. 361-72.
223. Anderson, C.A., et al., *Data quality control in genetic case-control association studies.* Nat Protoc, 2010. **5**(9): p. 1564-73.
224. Sabatti, C. and N. Risch, *Homozygosity and linkage disequilibrium.* Genetics, 2002. **160**(4): p. 1707-19.
225. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
226. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.
227. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
228. Field, J.K., et al., *The Liverpool Lung Project research protocol.* Int J Oncol, 2005. **27**(6): p. 1633-45.
229. R, *R: A Language and Environment for Statistical Computing.* R Develoment Core Team, 2010.
230. Graffelman, J., *calibrate: Calibration of Scatterplot and Biplot Axes.* 2010.
231. Scott, M., *NCBI2R: NCBI2R-An R package to navigate and annotate genes and SNPs.* 2010.
232. Swinton, J., *Vennerable: Venn and Euler area-proportional diagrams.* 2009.
233. Wang, Y., et al., *Common 5p15.33 and 6p21.33 variants influence lung cancer risk.* Nat Genet, 2008. **40**(12): p. 1407-9.
234. Landi, M.T., et al., *A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma.* American journal of human genetics, 2009. **85**(5): p. 679-691.
235. Fog, C.K., G.G. Galli, and A.H. Lund, *PRDM proteins: important players in differentiation and disease.* Bioessays, 2012. **34**(1): p. 50-60.
236. Di Zazzo, E., et al., *PRDM Proteins: Molecular Mechanisms in Signal Transduction and Transcriptional Regulation.* Biology, 2013. **2**(1): p. 107-141.
237. Hohenauer, T. and A.W. Moore, *The Prdm family: expanding roles in stem cells and development.* Development, 2012. **139**(13): p. 2267-82.
238. Jiang, G.L. and S. Huang, *The yin-yang of PR-domain family genes in tumorigenesis.* Histol Histopathol, 2000. **15**(1): p. 109-17.
239. Porcu, E., et al., *A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function.* PLoS Genet, 2013. **9**(2): p. e1003266.

240. Chen, C., et al., *Next-generation-sequencing-based risk stratification and identification of new genes involved in structural and sequence variations in near haploid lymphoblastic leukemia.* Genes Chromosomes Cancer, 2013. **52**(6): p. 564-79.
241. Cheng, Y., et al., *KRAB zinc finger protein ZNF382 is a proapoptotic tumor suppressor that represses multiple oncogenes and is commonly silenced in multiple carcinomas.* Cancer Res, 2010. **70**(16): p. 6516-26.
242. Greco, S.A., et al., *Thrombospondin-4 is a putative tumour-suppressor gene in colorectal cancer that exhibits age-related methylation.* BMC Cancer, 2010. **10**: p. 494.
243. Di Cello, F., et al., *HMGA2 participates in transformation in human lung cancer.* Mol Cancer Res, 2008. **6**(5): p. 743-50.
244. Meyer, B., et al., *HMGA2 overexpression in non-small cell lung cancer.* Mol Carcinog, 2007. **46**(7): p. 503-11.
245. Soler Artigas, M., et al., *Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function.* Nature Genetics, 2011. **43**(11): p. 1082-U70.
246. Quaye, L., et al., *Tagging single-nucleotide polymorphisms in candidate oncogenes and susceptibility to ovarian cancer.* Br J Cancer, 2009. **100**(6): p. 993-1001.
247. Hamm, A., et al., *Frequent expression loss of Inter-alpha-trypsin inhibitor heavy chain (ITIH) genes in multiple human solid tumors: a systematic expression analysis.* BMC Cancer, 2008. **8**: p. 25.
248. Unoki, M., et al., *UHRF1 is a novel diagnostic marker of lung cancer.* Br J Cancer, 2010. **103**(2): p. 217-22.
249. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Res, 2012. **40**(Database issue): p. D130-5.
250. NICE, *Lung Cancer. The diagnosis and treatment of lung cancer.* National Institute for Clinical Excellence, 2005.
251. Sato, Y., et al., *Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel.* J Thorac Oncol, 2011. **6**(1): p. 132-8.
252. Hu, L., et al., *Genome-wide association study of prognosis in advanced non-small cell lung cancer patients receiving platinum-based chemotherapy.* Clin Cancer Res, 2012. **18**(19): p. 5507-14.
253. Wu, X., et al., *Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy.* J Natl Cancer Inst, 2011. **103**(10): p. 817-25.
254. Tan, X.L., et al., *Genetic variation predicting cisplatin cytotoxicity associated with overall survival in lung cancer patients receiving platinum-based chemotherapy.* Clin Cancer Res, 2011. **17**(17): p. 5801-11.
255. Lee, Y., et al., *Prognostic implications of genetic variants in advanced non-small cell lung cancer: a genome-wide association study.* Carcinogenesis, 2012.
256. Wu, C., F. Li, and S.C. Jiao, *Prognostic factors for survival of patients with extensive stage small cell lung cancer--a retrospective single institution analysis.* Asian Pac J Cancer Prev, 2012. **13**(10): p. 4959-62.
257. Yang, X.W., et al., *Analysis of the relationships between clinicopathologic factors and survival in gallbladder cancer following surgical resection with curative intent.* PLoS One, 2012. **7**(12): p. e51513.

258. Huh, J.W., et al., *Factors predicting long-term survival in colorectal cancer patients with a normal preoperative serum level of carcinoembryonic antigen.* J Cancer Res Clin Oncol, 2013.
259. Laohavinij, S., K. Ruikchuchit, and J. Maneechavakajorn, *Survival and prognostic factors of stage I-III breast cancer.* J Med Assoc Thai, 2013. **96 Suppl 3**: p. S23-34.
260. Noorkojuri, H., et al., *Application of smoothing methods for determining of the effecting factors on the survival rate of gastric cancer patients.* Iran Red Crescent Med J, 2013. **15**(2): p. 166-72.
261. Huang, Y.T., et al., *Genome-wide analysis of survival in early-stage non-small-cell lung cancer.* J Clin Oncol, 2009. **27**(16): p. 2660-7.
262. Niu, N.S., D.J.; Abo, R.P.; Kalari, K.; , *Genetic association with overall survival of taxane-treated lung cancer patients - A genome-wide association study in human lymphoblastoid cell lines followed by a clinical association study.* BMC Cancer 2012. **12**: p. 422.
263. Chen, X.D., et al., *ANKRD7 and CYTL1 are novel risk genes for alcohol drinking behavior.* Chin Med J (Engl), 2012. **125**(6): p. 1127-34.
264. Huang, Y.T., et al., *Genome-Wide Analysis of Survival in Early-Stage Non-Small-Cell Lung Cancer.* Journal of Clinical Oncology, 2009. **27**(16): p. 2660-2667.
265. Wang, K.S., X.F. Liu, and N. Aragam, *A genome-wide meta-analysis identifies novel loci associated with schizophrenia and bipolar disorder.* Schizophr Res, 2010. **124**(1-3): p. 192-9.
266. Yoshioka, N., et al., *POU6F1 is the transcription factor that might be involved in cell proliferation of clear cell adenocarcinoma of the ovary.* Human Cell, 2009. **22**(4): p. 94-100.
267. Lawless, J.F., *Statistical models and methods for lifetime data*. 2nd ed. Wiley series in probability and statistics. 2003, Hoboken, N.J. ; [Great Britain]: Wiley-Interscience. xx, 630 p.
268. Lawless, J.F., *Statistical models and methods for lifetime data*. Wiley series in probability and mathematical statistics,0271-6356. 1982, New York ; Chichester: Wiley. xi,580p.
269. Sparling, Y.H., et al., *Parametric survival models for interval-censored data with time-dependent covariates.* Biostatistics, 2006. **7**(4): p. 599-614.
270. Rich, J.T., et al., *A practical guide to understanding Kaplan-Meier curves.* Otolaryngol Head Neck Surg, 2010. **143**(3): p. 331-6.
271. Bland, J.M. and D.G. Altman, *Survival probabilities (the Kaplan-Meier method).* BMJ, 1998. **317**(7172): p. 1572.
272. Jager, K.J., et al., *The analysis of survival data: the Kaplan-Meier method.* Kidney Int, 2008. **74**(5): p. 560-5.
273. Layton, D.M., *Understanding kaplan-meier and survival statistics.* Int J Prosthodont, 2013. **26**(3): p. 218-26.
274. Therneau, T.M. and P.M. Grambsch, *Modeling survival data : extending the Cox model*. Statistics for biology and health. 2000, New York ; London: Springer. xiii, 350 p.
275. Therneau, T., *A Package for Survival Analysis in S. R package version 2.37-2, http://CRAN.R-project.org/package=survival. .* 2012.
276. Dalgaard, P. and I. ebrary, *Introductory statistics with R [electronic resource]*. Statistics and computing. 2002, New York, New York, NY: Springer, Springer-Verlag New York. 1 online resource (xv, 267 p.).
277. Hou, X., et al., *Cystin, a novel cilia-associated protein, is disrupted in the cpk mouse model of polycystic kidney disease.* J Clin Invest, 2002. **109**(4): p. 533-40.

278. Moniz, L.S. and V. Stambolic, *Nek10 mediates G2/M cell cycle arrest and MEK autoactivation in response to UV irradiation.* Mol Cell Biol, 2011. **31**(1): p. 30-42.
279. UniProt, C., *Reorganizing the protein space at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 2012. **40**(Database issue): p. D71-5.
280. Zhang, W., et al., *Identification and characterization of DPZF, a novel human BTB/POZ zinc finger protein sharing homology to BCL-6.* Biochem Biophys Res Commun, 2001. **282**(4): p. 1067-73.
281. Greenman, C., et al., *Patterns of somatic mutation in human cancer genomes.* Nature, 2007. **446**(7132): p. 153-8.
282. Suetsugu, S., et al., *Male-specific sterility caused by the loss of CR16.* Genes Cells, 2007. **12**(6): p. 721-33.
283. Ahmed, S., et al., *Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2.* Nat Genet, 2009. **41**(5): p. 585-90.
284. Davies, H., et al., *Somatic mutations of the protein kinase gene family in human lung cancer.* Cancer Res, 2005. **65**(17): p. 7591-5.
285. Govindan, R., et al., *Genomic landscape of non-small cell lung cancer in smokers and never-smokers.* Cell, 2012. **150**(6): p. 1121-34.
286. Tian, X.Q., et al., *Epigenetic silencing of LRRC3B in colorectal cancer.* Scand J Gastroenterol, 2009. **44**(1): p. 79-84.
287. Kim, M., et al., *LRRC3B, encoding a leucine-rich repeat-containing protein, is a putative tumor suppressor gene in gastric cancer.* Cancer Res, 2008. **68**(17): p. 7147-55.
288. Wang, Q., et al., *Zinc finger protein ZBTB20 expression is increased in hepatocellular carcinoma and associated with poor prognosis.* BMC Cancer, 2011. **11**: p. 271.
289. Lee, Y.S., L.K. Poh, and K.Y. Loke, *A novel melanocortin 3 receptor gene (MC3R) mutation associated with severe obesity.* J Clin Endocrinol Metab, 2002. **87**(3): p. 1423-6.
290. Dose, A.C. and B. Burnside, *A class III myosin expressed in the retina is a potential candidate for Bardet-Biedl syndrome.* Genomics, 2002. **79**(5): p. 621-4.
291. Yang, X.X., et al., *Aberrant methylation and downregulation of sall3 in human hepatocellular carcinoma.* World J Gastroenterol, 2012. **18**(21): p. 2719-26.
292. Schadt, E.E., *Molecular networks as sensors and drivers of common human diseases.* Nature, 2009. **461**(7261): p. 218-23.
293. Brennan, P., et al., *Lung cancer susceptibility locus at 5p15.33.* Nat Genet, 2008. **40**(12): p. 1404-1406.
294. Brennan, P., et al., *Replication of Lung Cancer Susceptibility Loci at Chromosomes 15q25, 5p15, and 6p21: A Pooled Analysis From the International Lung Cancer Consortium.* Journal of the National Cancer Institute, 2010. **102**(13): p. 959-971.
295. Ramanan, V.K., et al., *Pathway analysis of genomic data: concepts, methods, and prospects for future development.* Trends Genet, 2012. **28**(7): p. 323-32.
296. Pang, H., D. Datta, and H. Zhao, *Pathway analysis using random forests with bivariate node-split for survival outcomes.* Bioinformatics, 2010. **26**(2): p. 250-8.
297. Pang, H., M. Hauser, and S. Minvielle, *Pathway-based identification of SNPs predictive of survival.* Eur J Hum Genet, 2011. **19**(6): p. 704-9.
298. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
299. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
300. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32**(Database issue): p. D277-80.

301. Camon, E.B., et al., *An evaluation of GO annotation retrieval for BioCreAtIvE and GOA.* BMC Bioinformatics, 2005. **6 Suppl 1**: p. S17.
302. Khatri, P. and S. Draghici, *Ontological analysis of gene expression data: current tools, limitations, and open problems.* Bioinformatics, 2005. **21**(18): p. 3587-95.
303. Kanehisa, M., et al., *KEGG for representation and analysis of molecular networks involving diseases and drugs.* Nucleic Acids Res, 2010. **38**(Database issue): p. D355-60.
304. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.
305. Kanehisa, M., *Molecular network analysis of diseases and drugs in KEGG.* Methods Mol Biol, 2013. **939**: p. 263-75.
306. Kanehisa, M., et al., *The KEGG databases at GenomeNet.* Nucleic Acids Res, 2002. **30**(1): p. 42-6.
307. O'Dushlaine, C., et al., *Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility.* Mol Psychiatry, 2011. **16**(3): p. 286-92.
308. Wang, K., M. Li, and M. Bucan, *Pathway-based approaches for analysis of genomewide association studies.* Am J Hum Genet, 2007. **81**(6): p. 1278-83.
309. Wang, K., et al., *Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease.* Am J Hum Genet, 2009. **84**(3): p. 399-405.
310. Khatri, P., M. Sirota, and A.J. Butte, *Ten years of pathway analysis: current approaches and outstanding challenges.* PLoS Comput Biol, 2012. **8**(2): p. e1002375.
311. Elbers, C.C., et al., *Using Genome-Wide Pathway Analysis to Unravel the Etiology of Complex Diseases.* Genetic Epidemiology, 2009. **33**(5): p. 419-431.
312. Holmans, P., *Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits.* Adv Genet, 2010. **72**: p. 141-79.
313. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
314. Torkamani, A., E.J. Topol, and N.J. Schork, *Pathway analysis of seven common diseases assessed by genome-wide association.* Genomics, 2008. **92**(5): p. 265-72.
315. Askland, K., C. Read, and J. Moore, *Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission.* Hum Genet, 2009. **125**(1): p. 63-79.
316. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
317. Fehringer, G., et al., *Comparison of pathway analysis approaches using lung cancer GWAS data sets.* PLoS One, 2012. **7**(2): p. e31816.
318. Chung, R.H. and Y.E. Chen, *A two-stage random forest-based pathway analysis method.* PLoS One, 2012. **7**(5): p. e36662.
319. Lee, D., et al., *Pathway-based analysis using genome-wide association data from a korean non-small cell lung cancer study.* PLoS One, 2013. **8**(6): p. e65396.
320. Zhang, R., et al., *Pathway analysis for genome-wide association study of lung cancer in Han Chinese population.* PLoS One, 2013. **8**(3): p. e57763.
321. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.
322. Browning, B.L. and S.R. Browning, *A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.* Am J Hum Genet, 2009. **84**(2): p. 210-23.

323. Ishwaran H, K.U., *Random survival forests for R.* Rnews, 2007(7): p. 25-31.
324. Kanehisa, M., et al., *KEGG for integration and interpretation of large-scale molecular data sets.* Nucleic Acids Res, 2012. **40**(Database issue): p. D109-14.
325. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
326. Saccone, S.F., et al., *New tools and methods for direct programmatic access to the dbSNP relational database.* Nucleic Acids Res, 2011. **39**(Database issue): p. D901-7.
327. Pang, H., *brsf: Random Survival Forest with bivariate split.* R package version 3.5.1. http://www.duke.edu/, 2009.
328. Therneau, T., *A Package for Survival Analysis in S.* R package version 2.37-7, <URL: http://CRAN.R-project.org/package=survival>. 2014.
329. Therneau, T.a.G., PM *Modeling Survival Data: Extending the Cox Model.* 2000.
330. Heagerty, P.a.S.-C., P, *survivalROC: Time-dependent ROC curve estimation from censored survival data.* R package version 1.0.3. http://CRAN.R-project.org/package=survivalROC, 2013.
331. Benjamini, Y., and Hochberg, Y, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* of the Royal Statistical Society Series B, 1995(57): p. 289–300.
332. Luojus, M.K., et al., *Sleep duration and incidence of lung cancer in ageing men.* BMC Public Health, 2014. **14**(1): p. 295.
333. Liu, Z.H., et al., *Interleukin 7 signaling prevents apoptosis by regulating bcl-2 and bax via the p53 pathway in human non-small cell lung cancer cells.* Int J Clin Exp Pathol, 2014. **7**(3): p. 870-81.
334. Zhao, W., et al., *Polymorphisms in the base excision repair pathway modulate prognosis of platinum-based chemotherapy in advanced non-small cell lung cancer.* Cancer Chemother Pharmacol, 2013. **71**(5): p. 1287-95.
335. Wan, J., et al., *Thoc1 inhibits cell growth via induction of cell cycle arrest and apoptosis in lung cancer cells.* Mol Med Rep, 2014.
336. Emmendoerffer, A., et al., *Role of inflammation in chemical-induced lung cancer.* Toxicol Lett, 2000. **112-113**: p. 185-91.
337. Erdel, M., et al., *Cell interactions and motility in human lung tumor cell lines HS-24 and SB-3 under the influence of extracellular matrix components and proteinase inhibitors.* Anticancer Res, 1992. **12**(2): p. 349-59.
338. Nikolos, F., et al., *ERbeta Regulates NSCLC Phenotypes by Controlling Oncogenic RAS Signaling.* Mol Cancer Res, 2014.
339. Ioacara, S., et al., *Cancer specific mortality in insulin-treated type 2 diabetes patients.* PLoS One, 2014. **9**(3): p. e93132.
340. Li, M., et al., *Expression of the mismatch repair gene hMLH1 is enhanced in non-small cell lung cancer with EGFR mutations.* PLoS One, 2013. **8**(10): p. e78500.
341. Setia, S. and S.N. Sanyal, *Nuclear factor kappa B: a pro-inflammatory, transcription factor-mediated signalling pathway in lung carcinogenesis and its inhibition by nonsteroidal anti-inflammatory drugs.* J Environ Pathol Toxicol Oncol, 2012. **31**(1): p. 27-37.
342. Warren, G.W. and A.K. Singh, *Nicotine and lung cancer.* J Carcinog, 2013. **12**: p. 1.
343. Garcia Campelo, M.R., et al., *Stem cell and lung cancer development: blaming the Wnt, Hh and Notch signalling pathway.* Clin Transl Oncol, 2011. **13**(2): p. 77-83.
344. Mei, C.R., et al., *DNA Repair Gene Polymorphisms in the Nucleotide Excision Repair Pathway and Lung Cancer Risk: A Meta-analysis.* Chin J Cancer Res, 2011. **23**(2): p. 79-91.

345. Warin, R.F., et al., *Induction of lung cancer cell apoptosis through a p53 pathway by [6]-shogaol and its cysteine-conjugated metabolite M2.* J Agric Food Chem, 2014. **62**(6): p. 1352-62.
346. Cai, W.K., et al., *Activation of histamine H4 receptors decreases epithelial-to-mesenchymal transition progress by inhibiting transforming growth factor-beta1 signalling pathway in non-small cell lung cancer.* Eur J Cancer, 2014. **50**(6): p. 1195-206.
347. Chatterjee, S., et al., *Tumor VEGF:VEGFR2 autocrine feed-forward loop triggers angiogenesis in lung cancer.* J Clin Invest, 2013. **123**(4): p. 1732-40.
348. Brambilla, C., et al., *Early detection of lung cancer: role of biomarkers.* Eur Respir J Suppl, 2003. **39**: p. 36s-44s.
349. Fong, K.M., et al., *Lung cancer. 9: Molecular biology of lung cancer: clinical implications.* Thorax, 2003. **58**(10): p. 892-900.
350. Pogribny, I.P. and F.A. Beland, *DNA methylome alterations in chemical carcinogenesis.* Cancer Lett, 2012.
351. Kiyohara, C. and K. Yoshimasu, *Genetic polymorphisms in the nucleotide excision repair pathway and lung cancer risk: a meta-analysis.* Int J Med Sci, 2007. **4**(2): p. 59-71.
352. Landi, S., et al., *DNA repair and cell cycle control genes and the risk of young-onset lung cancer.* Cancer Res, 2006. **66**(22): p. 11062-9.
353. Simon, G.R., R. Ismail-Khan, and G. Bepler, *Nuclear excision repair-based personalized therapy for non-small cell lung cancer: from hypothesis to reality.* Int J Biochem Cell Biol, 2007. **39**(7-8): p. 1318-28.
354. Hung, R.J., et al., *Genetic polymorphisms in the base excision repair pathway and cancer risk: a HuGE review.* Am J Epidemiol, 2005. **162**(10): p. 925-42.
355. Ray, M.R., D. Jablons, and B. He, *Lung cancer therapeutics that target signaling pathways: an update.* Expert Rev Respir Med, 2010. **4**(5): p. 631-45.
356. Zhang, Y. and J. He, *The development of targeted therapy in small cell lung cancer.* J Thorac Dis, 2013. **5**(4): p. 538-48.
357. Erin LeDell, Maya Petersen and Mark van der Laan (2013). cvAUC: Cross-Validated Area Under the ROC Curve Confidence Intervals. R package version 1.0-0. http://CRAN.R-project.org/package=cvAUC.