Lung cancer in United Kingdom general practice and the possibility of developing an early-warning score

Barbara Iyen-Omofoman, MBBS, MPH

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

November 2012

Abstract

Background

Lung cancer has a dreadful prognosis and is the leading cause of cancer deaths in the world and in the UK. The UK survival rates are particularly poor when compared with survival in other countries in Europe. More than two-thirds of people with lung cancer in the UK are diagnosed at a late stage when curative treatment is no longer possible. Since lung cancer survival rates are higher with earlier diagnosis, there is need to diagnose cases earlier. This suggests a potential to examine and if possible, modify the care pathway for people with lung cancer to achieve earlier diagnosis.

Aim

The overall aim of this thesis was to explore the patient characteristics and interactions in primary care before the diagnosis of lung cancer, as a means of identifying the features that are predictive of lung cancer and the potential for earlier diagnosis. To achieve this aim, it was necessary to investigate and validate the use of lung cancer data from The Health Improvement Network.

Methods

The Health Improvement Network (THIN) database of United Kingdom general practice records, was used to identify and study the characteristics of cases of lung cancer in the UK. To ensure that THIN was a valid source of lung cancer information for research, a study was done to assess the completeness and representativeness of the lung cancer data in THIN by comparing the lung cancer patient characteristics, incidence and survival in THIN with the UK National Cancer Registry and the National Lung Cancer Audit Database. Experian's Mosaic

Public Sector [tm] variable linked into THIN database was then used to identify detailed profiles of the UK sectors of society where lung cancer incidence was highest as a means of exploring the potential of using this geo-demographic tool to facilitate disease ascertainment.

Two case-control datasets were developed from the database using the identified cases of lung cancer. The first dataset was matched on age, sex and general practice and it was used to carry out three studies in this thesis. The first study was a pilot study of methods to identify the socio-demographic and clinical features independently associated with lung cancer as well as to identify the timing of these clinical features before lung cancer was diagnosed. This was followed by two studies to examine separate hypotheses on the variation in lung cancer risk firstly between smokers of different socioeconomic status, then between smokers with and without a recorded history of depression, as socioeconomic deprivation and depression are both associated with increased prevalence of cigarette smoking.

The second case-control dataset was matched only on practice and this dataset expanded on the methods from the pilot study to identify the socio-demographic factors including age and sex, as well as the early clinical features that are predictive of lung cancer. This was followed by a study which used the identified predictors to develop and validate a risk-prediction model for lung cancer. The model validation was carried out using another dataset of patients in a more recent version of THIN with records spanning a time period after the last date of records for patients used for the earlier studies in the thesis.

Results

A study population of 12,135 patients with incident lung cancer were identified from the 1st of January 2000 to the 28th of July 2009. The overall incidence of

lung cancer, median survival and general lung cancer patient characteristics in THIN were similar to other national lung cancer databases - The National Lung Cancer Audit Data and the UK National Lung Cancer Registry data from the Office of National Statistics. Mosaic™ classifications identified wider variations in lung cancer incidence than existing markers of socioeconomic deprivation and therefore allowed more detailed classifications of the UK sectors of society where lung cancer incidence was highest. For example the incidence rate in Mosaic Public Sector™ type I50 (Cared-for pensioners) was 31.2 times higher (IRR 31.2; 95% CI 21.9-44.5) than the incidence rate in Mosaic Public Sector™ type B10 (Upscale new owners).

With regards to the risk of lung cancer among smokers from different socioeconomic groups, stratified analyses of the association between smoking and lung cancer by Townsend deprivation quintiles showed that the risks of lung cancer were similar in smokers of different socioeconomic status. Depression was associated with a 30% increased risk of lung cancer (odds ratio 1.30; 95% CI 1.24-1.38) which was completely explained by smoking. Cigarette smoking was more common and levels of consumption were higher among depressed compared to non-depressed individuals. Stratified analyses of the association between smoking and lung cancer by depression showed that there was no difference in lung cancer risk among depressed and non-depressed smokers.

Socio-demographic features - age, sex, socioeconomic status and smoking, increase in the frequency of general practice consultations as well as early records of presentation for symptoms of cough, haemoptysis, dyspnoea, weight loss, lower respiratory tract infections, non-specific chest infections, chest pain, hoarseness, upper respiratory tract infections and Chronic Obstructive Pulmonary Disease (COPD) were found to be independently associated with lung cancer 4 to 12 months before diagnosis. A risk prediction model was developed with these variables, and on validation using an independent THIN dataset of

1,826,293 patients, the model performed well with an area under the curve statistic of 0.88.

Conclusions

Routine electronic data in THIN are a valid source of lung cancer information for research. Mosaic™ identifies greater incidence differentials than standard area-level measures and as such could be used as a tool for public health programmes to ascertain future cases more effectively.

Neither socioeconomic deprivation nor a history of depression increases an individuals' vulnerability to the carcinogenic effects of cigarette smoke. The increase in lung cancer risk among more deprived individuals and those with depression is largely explained by the greater cigarette consumption by these groups of people. Smoking cessation interventions targeted to these groups of people are needed to reduce the lung cancer-related health inequalities associated with deprivation and depression.

A combination of patients' age, sex, socioeconomic characteristics, smoking status and early stage symptoms in general practice aid earlier identification of patients at increased risk of lung cancer. The model developed using these variables performed substantially better than the current NICE referral guidelines and all comparable models, being able to predict lung cancer early enough to make detection at a potentially curable stage feasible by allowing general practitioners to better risk-stratify their patients.

Acknowledgements

The work in this thesis was carried out under the supervision of Professor Richard Hubbard and Dr Laila Tata in the Division of Epidemiology and Public Health, University of Nottingham and the PhD studentship was funded by the Economic and Social Research Council.

I am extremely grateful to my supervisors for their guidance, commitment and encouragement throughout the period of this PhD, from the initial conception of the study to the final version of the thesis. Your support has been invaluable and I would not have reached this stage without you.

I would like to acknowledge and thank Chris Smith for his help and diligence in extracting all the data used for analyses in this thesis. I would also like to thank the staff at EPIC for providing access to the database used for the analyses in this thesis. Emma Bradley and Emily Sparks from Experian UK, were also of great help for their contributions to the studies on Mosaic in this thesis.

To the current and previous occupants of room C109 especially Ailsa Lyons and Ilze Bogdanovica who were there throughout my PhD tenure, I am grateful for your advice, support and friendship and for enduring two and a half years of my constant moaning. I am also particularly thankful to Lisa Szatkowski who was a great source of invaluable statistical help during the early stages of the PhD.

I would like to thank my parents and siblings for their support, encouragement, prayers and for instilling in me the belief that I can achieve anything that I believe and set out to do. I would like to thank my husband Austin and beautiful daughters Nicole and Biana for their love, support and for enduring the long days and nights when I have had to work while undertaking this project.

Lastly but by no means least, I am grateful to God for making all this possible.

List of peer reviewed published papers and conference presentations arising from this thesis

Iyen-Omofoman B, Hubbard RB, Smith CJ, Sparks E, Bradley E, Bourke A, Tata LJ. The distribution of lung cancer across sectors of society in the United Kingdom: A study using national primary care data. *BMC Public Health.* 2011;11(1):857.

Iyen-Omofoman B, Tata LJ, Baldwin DR, Smith CJ, Hubbard RB. Using Socio-demographic and early clinical features in general practice to identify people with lung cancer earlier. (Submitted and under review)

Iyen-Omofoman B, Tata LJ, Smith CJ, Hubbard RB. Depression, smoking and the risk of lung cancer in UK general population. (Submitted and under review)

Iyen-Omofoman B, Tata L, Hubbard R. Using a social marketing tool to identify sectors of the United Kingdom where lung cancer incidence is highest. *American Thoracic Society International Conference*, Denver, Colorado, 13-18 May 2011.

Iyen-Omofoman B, Tata L, Hubbard R. How long do patients in the UK get treated for non-specific respiratory symptoms by general practitioners before they are diagnosed with lung cancer? *American Thoracic Society International Conference*, Denver, Colorado, 13-18 May 2011.

Table of Contents

List of Tables

Table of Figures

Abbreviations

95% CI	95% Confidence Intervals
BMJ	British Medical Journal
c statistic	Concordance statistic
CA-125	Cancer Antigen 125
COPD	Chronic Obstructive Pulmonary Disease
CT	Computerised Tomography
EPIC	Epidemiology and Pharmacology Information Core
ETS	Environmental Tobacco Smoke
EUS-FNA	Endoscopic Ultrasound Guided Fine Needle Aspiration
GLS	General Lifestyle Survey
GP	General Practitioner
GPRD	General Practice Research Database
HSE	Health Survey for England
InPS	In Practice Systems
LRTI	Lower Respiratory Tract Infection
LUCADA	National Lung Cancer Audit Data
NCIN	National Cancer Intelligence Network
HE4	Human Epididymis Protein 4
NHS	National Health Service
NICE	National Institute of Health and Clinical Excellence
NSCLC	Non Small Cell Lung Cancer
ONS	Office for National Statistics
OR	Odds ratio
PCT	Primary Care Trust
QOF	Quality and Outcomes Framework
ROC	Receiver-Operating Characteristic Curve
SCLC	Small Cell Lung Cancer
SI	Symptom Index
TAC	Tobacco Advisory Council
THIN	The Health Improvement Network
UK	United Kingdom
URTI	Upper Respiratory Tract Infection
VAMP	Value Added Medical Products

Chapter 1. Introduction

Although the improvement of treatment and survival of people with lung cancer is of utmost priority among those in the field of lung cancer research in the UK, very few studies have explored the interaction between general practitioners (GPs) and patients who develop lung cancer before they are diagnosed. Using a computerised database of UK general practice records, this thesis aims to extensively investigate the GP-patient interaction in the period before lung cancer diagnosis, with a view to determining the possibility of developing a predictive score for lung cancer that could be used to aid earlier diagnosis of future cases.

This introductory chapter gives an overview of what is already known about the burden of lung cancer globally and in the UK in particular, the risk factors and other characteristics associated with lung cancer, the clinical presentation as well as current guidelines for diagnosing lung cancer in the UK. It also highlights the need to recognise lung cancer earlier in general practice, an overview of predictive scores with particular emphasis on existing scores for lung cancer and the gaps in the evidence. This will be followed by a rationale of the work in this thesis as well as detailed aims and objectives of the thesis.

1.1 Background

1.1.1 Definition of Lung cancer

Lung cancer is an epithelial tumour arising in the mucosa of the bronchi or more rarely, in the lung parenchyma[1] . These cancers may:

- expand into the airways and cause symptoms such as cough, haemoptysis, airway obstruction.
- invade locally within the thorax, leading to compression and invasion of the chest wall.
- spread through the hilar, mediastinal and supraclavicular nodes.
- metastasize through the blood, to other parts of the body, particularly to the brain, liver, adrenals and the axial skeleton.
- induce changes in the peripheral or central nervous system (paraneoplastic effects), causing symptoms such as anorexia and inappropriate hormone production.

1.1.2 International and national burden of lung cancer

Lung cancer is the most common cancer globally[2], with over 1.6 million cases diagnosed worldwide in 2008. In the western world, lung cancer is associated with a significant health burden[3] and is the leading cause of cancer deaths in Europe[4]. In 2008 in Europe, the number of diagnosed cases and deaths from lung cancer alone were an estimated 391,000 cases and 342,000 deaths respectively[4]. In the UK, lung cancer is the second most common cancer diagnosed[5] after breast cancer with 41,397 cases diagnosed in 2009 alone[6-9] (Figure 1.1). It is also the leading cause of cancer deaths in the UK, accounting for 22% of all cancer deaths and 6% of all deaths. In 2010, there were 34,874 deaths from lung cancer in the UK[10-12], which was over 7000 more deaths than

the 2nd and 3rd most common causes of cancer deaths, bowel cancer and breast cancer, combined (Figure 1.2).

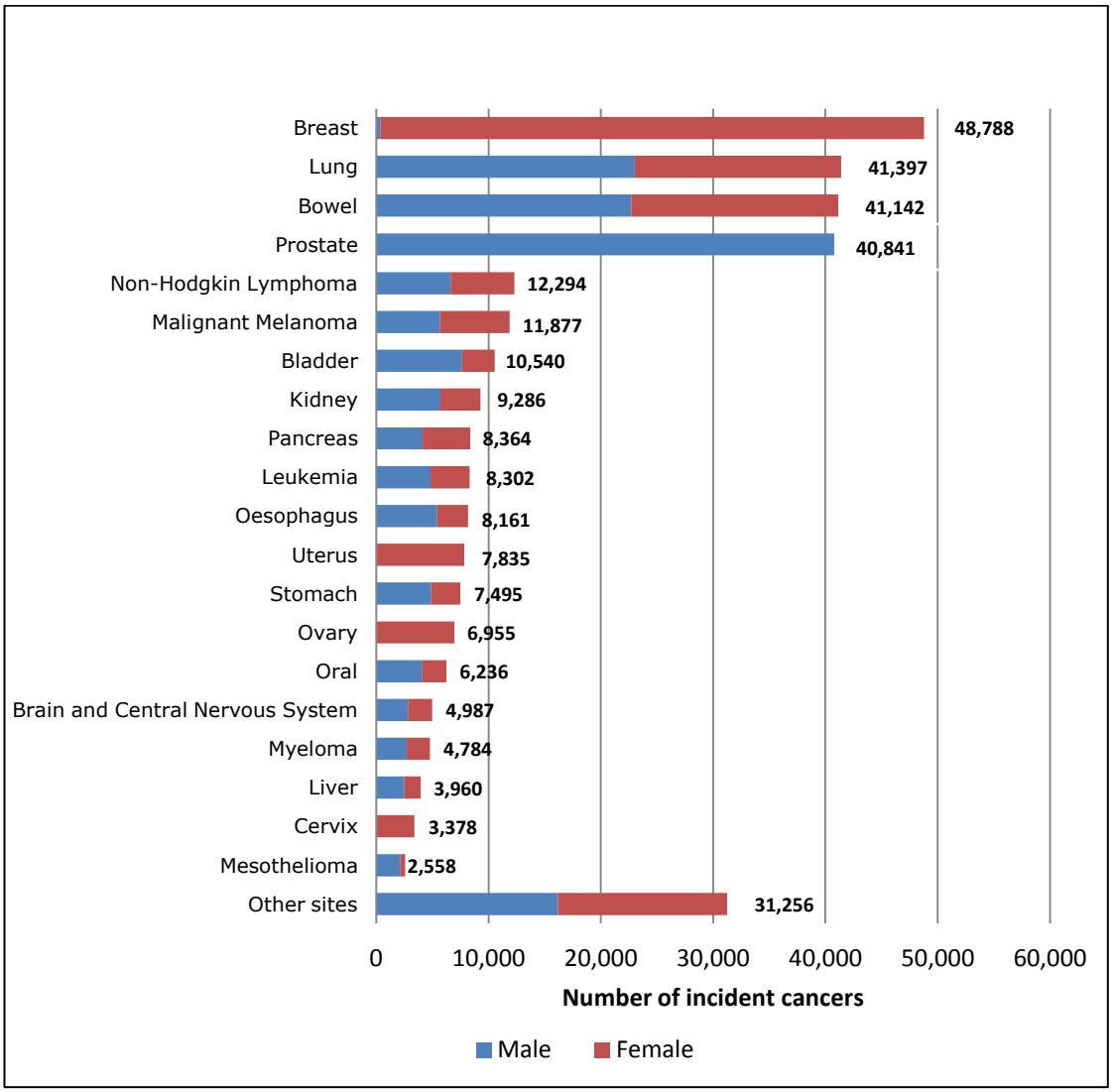


Figure 1.1 The 20 most commonly diagnosed cancers (excluding non-melanoma skin cancer) in the UK in 2009.
Based on data from Cancer Research UK[13]

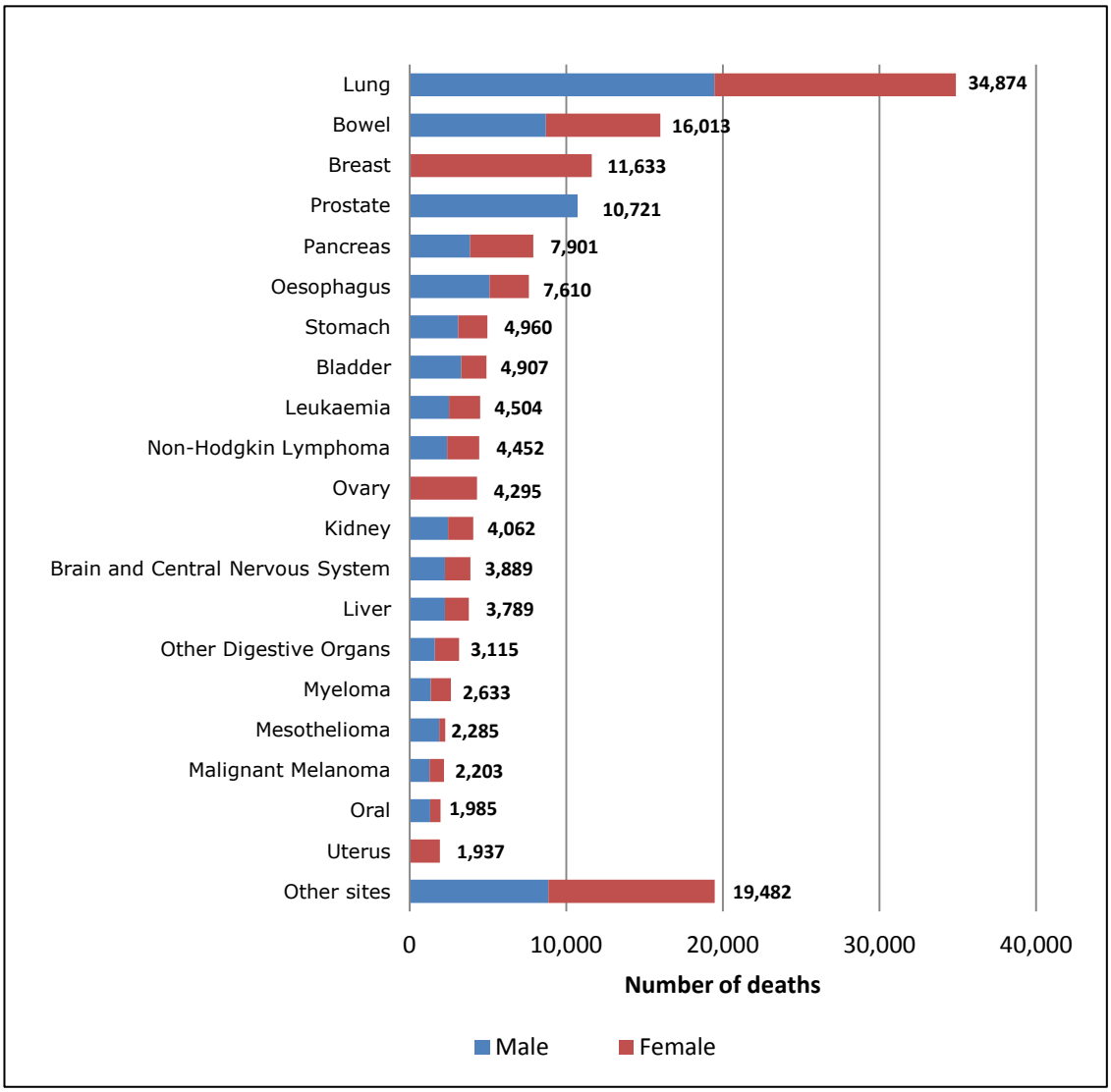


Figure 1.2 The 20 most common causes of cancer death in the UK in 2010.

Based on data from Cancer Research UK[14]

The survival rates for people with lung cancer are very low, thereby resulting in mortality rates that closely follow incidence[15]. Due to the poor prognosis for lung cancer, the survival outcome is one of the worst of any cancer. The latest data from the Office of National Statistics (ONS) show that for adults diagnosed in England during 2005 to 2009 and followed up to 2010, the one-year survival for lung cancer was 29% for men and 33% for women, falling to 8% and 9% respectively after five years[16].

Lung cancer survival rates have been shown to have great variation across Europe[17-20]. Evidence from data of lung cancer cases recorded in 44 population-based cancer registries in 17 countries showed that the highest relative age-standardized 1-year survival rates for lung cancer were approximately 40% in Finland, France, The Netherlands and Switzerland while they were relatively low for patients in the UK at 24%[17]. A recent study which compared lung cancer survival in England, Norway and Sweden, countries with a similar expenditure on healthcare, showed that across all categories of age and sex, 5-year survival after lung cancer diagnosis was lower in England compared to Norway and Sweden[18]. The difference in survival between England and the other 2 countries was much larger than the difference in survival between Norway and Sweden. Survival differences were most marked during the early period of follow-up after diagnosis and these diminished considerably with increasing years of follow-up (Figure 1.3).

The poor survival of patients with lung cancer in the UK has been suggested to be partly explained by a later stage at diagnosis among patients with lung cancer in the UK[18] and poor access to specialized care and treatment[17].

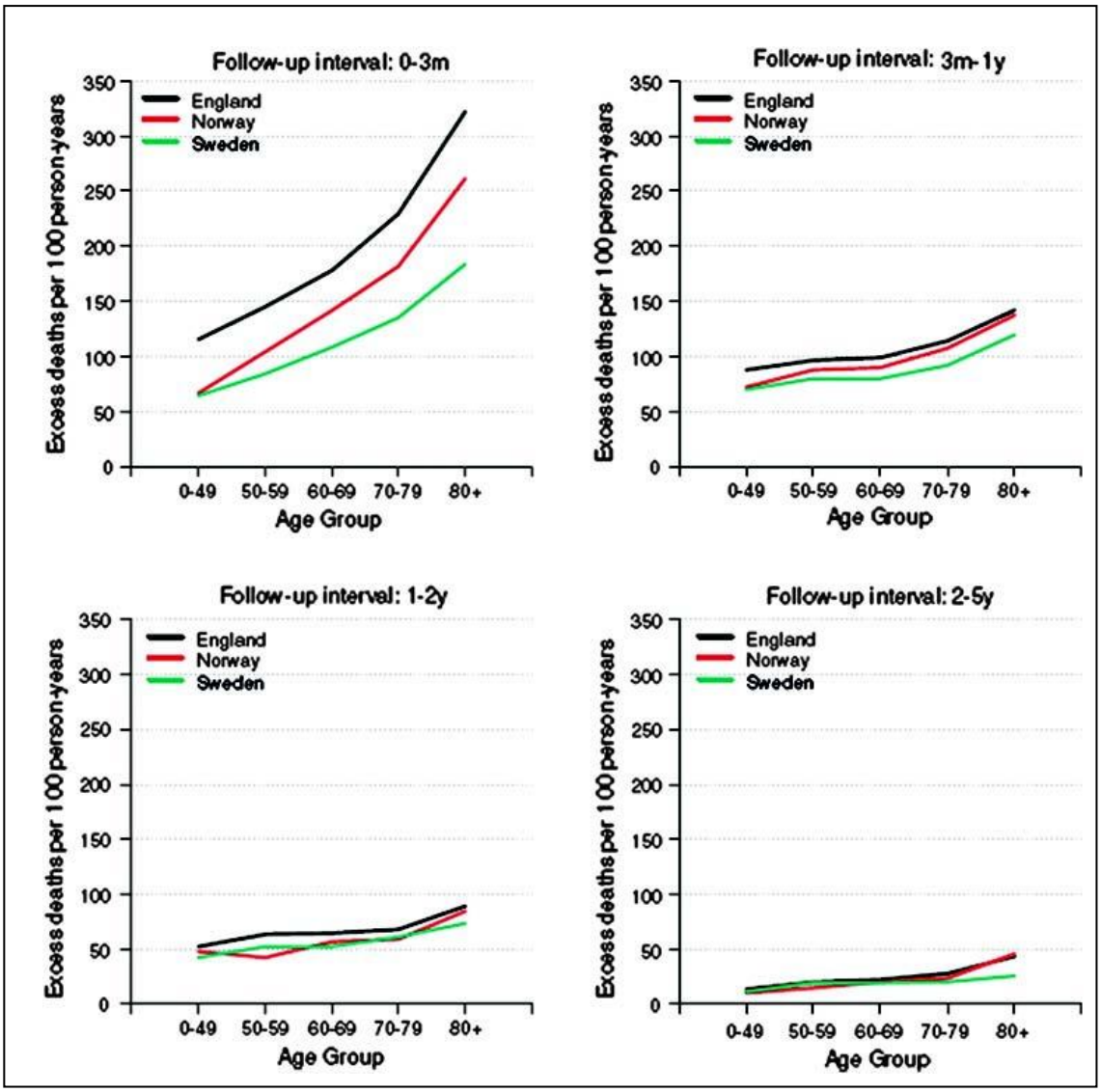


Figure 1.3 Excess deaths from lung cancer/ 100 person-years in England, Norway and Sweden, by age group and period of follow-up.

Adapted from Holmberg et al.[18]

1.1.3 Risk factors for lung cancer

1.1.3.1 Smoking

The first large-scale studies which examined the relationship between cigarette smoking and lung cancer[21-25] form the basis of some of the epidemiological studies in current use. This association was established over 60 years ago[23] and cigarette smoking has remained the most important risk factor for lung cancer[26]. Smoking has also been linked to other cancers including cancers of the oral cavity and pharynx, oesophageal cancer, stomach cancer and pancreatic cancer[26].

Cigarette smoking became popular in the United Kingdom in the early 20th century with the uptake occurring about 20 years earlier in men than women. Smoking prevalence peaked in the 1940s in men and in the late 1960s in women and has declined steadily since then[27]. Surveys by the Tobacco Advisory Council (TAC) in 1948 estimated that the prevalence of cigarette smoking among men and women in Britain were 65% and 41% respectively[28]. In 2010, the smoking prevalence among men and women in Britain as estimated by the General Lifestyle Survey (GLS) was 21% and 20% respectively[29]. Trends in the incidence of lung cancer largely reflects peoples' past smoking pattern with a latency period of more than 20 years[30].

In 2010 in the UK, 85% of the lung cancer in males and 80% of lung cancer in females were attributable to tobacco smoking[26]. Current smokers have a 15-fold higher risk of death from lung cancer compared with lifelong non smokers[31], however the risk reduces significantly in people who stop smoking before middle age[32]. There is a dose-response relationship between cigarette consumption and lung cancer risk[33], people who smoke fewer cigarettes daily for a longer duration have an increased risk of lung cancer compared with those who smoke more cigarettes for a shorter duration[34 35]. There is also an increase in the risk of lung

cancer following exposure to Environmental Tobacco Smoke (ETS)[36 37] and an estimated 14% to 15% of lung cancers among individuals who have never smoked are thought to be due to exposure to ETS[26].

1.1.3.2 Occupational carcinogens

Past exposure to occupational carcinogens have been shown to increase the risk of lung cancer[38-41]. The most commonly implicated occupational agent in lung cancer aetiology is asbestos and it increases the risk of lung cancer among smokers and non-smokers alike[42]. The incidence of lung cancer occurs as early as 5 to 9 years after first exposure to asbestos but the excess lung cancer risk continues to increase for up to 20 years or longer[42]. Other occupational agents that increase the risk of lung cancer include silica, diesel engine exhausts, mineral oils, paint dust (combined with solvents exposure) and arsenic. In the UK, an estimated 21% of male lung cancer cases and 4% of female lung cancer are associated with occupational exposures[43 44].

1.1.3.3 Radon

Radon is a noble gas produced from the decay of naturally occurring uranium that can accumulate indoors in buildings as well as in underground mines. It is carcinogenic to humans and has found to be strongly associated with lung cancer[45]. Exposure to Radon is much more likely to cause lung cancer in people who smoke but it is also the leading cause of lung cancer in non-smokers[46]. In the UK, 3.4% of the lung cancers are attributable to residential exposure to radon[47] and it has been estimated that 9% of lung cancer deaths in Europe are as a result of indoor radon exposure[48].

1.1.3.4 Family history of lung cancer

A history of lung cancer in a first-degree relative is associated with a two-fold increased risk of lung cancer regardless of smoking status[49] and suggests the possibility of a hereditary predisposition to lung cancer or shared environmental risk factor exposure by members of the same family. In individuals less than 60 years of age, there is a five-fold increase in lung cancer risk if they have a first degree relative who was diagnosed with lung cancer at less than 60 years[50].

1.1.3.5 Previous cancer treatment

The risk of lung cancer is significantly increased up to 25 years after treatment for Hodgkin's lymphoma[51]. Prior treatment with chemotherapy and radiotherapy contribute to the risk and this is further increased in smokers compared to non-smokers[52]. The risk of lung cancer has also been shown to increase after treatment for non-Hodgkin's lymphoma[53], breast cancer[54] and testicular cancer[55]. Following surgical resection of early-stage lung cancer, there is a 1% to 2% risk per patient per year of developing a second lung cancer[56]. Individuals who survive for 2 years or longer after treatment for small-cell lung cancer also have a 2% to 13% risk per patient per year of developing a second lung cancer.

1.1.3.6 Acquired lung diseases

Several acquired lung diseases may increase susceptibility to lung cancer[38] and these associations have been noted for obstructive lung diseases such as chronic obstructive pulmonary disease (COPD) and fibrotic lung diseases. Although lung cancer and COPD are both principally caused by cigarette smoking[57], surplus evidence suggest an increase in lung cancer incidence in individuals already

diagnosed with COPD[58] even after adjusting for the confounding effect of smoking[59 60]. A few studies had shown an inverse relationship between asthma and lung cancer[61 62]. However, evidence from several other studies suggest that after adjusting for the effect of cigarette smoking, a positive association exists between asthma and the risk of lung cancer[63-65]. An increased risk of lung cancer has also been found with interstitial lung diseases such as Idiopathic pulmonary fibrosis[66 67], certain pneumoconiosis and systemic sclerosis[68].

1.1.3.7 Other risk factors for lung cancer

Air pollution such as traffic-related air pollution, power plants and waste incinerator emissions[69], domestic air pollution from heating and cooking with solid fuels[70 71] are all associated with an increased risk of lung cancer.

Studies have shown that increased levels of physical activity reduce the risk of lung cancer[72] and this risk has been shown to reduce even in people who are current or ex-smokers[73 74].

Dietary factors are related to the risk of lung cancer and studies have suggested that a high intake of fruits and vegetables may reduce lung cancer risk in individuals[75-77]. A reduced lung cancer risk has been found with high dietary zinc and copper intakes[78]. A recent study showed that 9% of lung cancer cases in the UK in 2010 may be related to low intake of fruits[79].

There have been conflicting evidence on the association between alcohol consumption and lung cancer risk. A pooled analysis of data from seven prospective studies suggested a slightly greater risk of lung cancer in male never smokers who consume high quantities of alcohol[80] while other studies have failed to show an independent association between alcohol consumption and lung cancer risk[81 82]. A recently conducted meta-analysis which assessed this

association in never smokers however showed that alcohol consumption does not have an independent positive association with lung cancer[83].

1.1.4 Other factors associated with lung cancer

1.1.4.1 Lung cancer and age

Lung cancer is predominantly a disease of older adults, being rare in individuals less than 40 years[5]. Most cases are diagnosed over the age of 60 years and incidence rate peaks between 80-84 years[5]. In England, the median age of lung cancer diagnosis is 71 years[84].

1.1.4.2 Lung cancer and sex

Lung cancer is more common in males than females. In the 1950s, the ratio of lung cancer in males compared to females was 6:1, this difference has however narrowed considerably and the lung cancer ratio in males compared to females is now 1.3:1[85]. While the incidence rate is declining in men, it appears to be increasing in women and most likely reflects the much steeper decline in smoking prevalence over the years in males compared to females[85]. A few studies have examined the association between smoking and lung cancer risk in males and females, and although findings from some of the studies suggest that females may have a higher risk of lung cancer per cigarette smoked[86 87], others have failed to show any difference in the smoking-associated risk of lung cancer among the sexes[88 89]. Between 1993 and 2008, the UK male lung cancer incidence rates declined by about 30% whereas the rate increased by 11% in females[85]. In 2009, there were 18,492 men and 14,622 women diagnosed with lung cancer in England[6].

1.1.4.3 Lung cancer and socio-economic status

There is a socio-economic gradient in the incidence of lung cancer, being higher among individuals of lower socio-economic status based on various measures of socio-economic status[38 90-92]. Differential exposure to factors such as smoking[93 94], diet[93], occupational exposures[95] and educational attainment[90 95] have all been implicated to explain the differences in incidence between socioeconomic groups. However, these have not fully explained these differences. Evidence from a meta-analysis found an overall increase in lung cancer risk among individuals of lower socioeconomic status which persisted after pooling risk estimates from smoking-adjusted and smoking-unadjusted studies[90]. Using data of patients diagnosed with cancer from all eight English cancer registries in the UK, Shack et al[96] assigned patients to an index of multiple deprivation (IMD) based on their area of residence and then investigated the differences in the incidence of lung cancer and other cancers among different socioeconomic groups. Results of the study showed the highest incidence of disease among the most deprived patients such that lung cancer was 2.5 times higher among the most deprived men than the most affluent men. Outside the UK, a similar association between deprivation and lung cancer exists in other countries including Canada[97] and The Netherlands[98]. Whether the socioeconomic difference in lung cancer incidence is explained by factors other than smoking or if it is due to residual confounding, is currently unknown.

A study by Fidler et al.[99] assessed the distribution of saliva cotinine levels (a metabolite of nicotine and an indicator of daily nicotine consumption) among cigarette smokers in the Health Survey for England (HSE) and showed that cotinine levels were higher among individuals with lower social class and higher levels of deprivation than individuals with higher social class and lower deprivation levels, and this remained even after accounting for reported daily cigarette consumption[99]. Although cigarette smoking is known to be higher

among more deprived individuals[100], the findings from the study by Fidler et al.[99] suggest that there may be higher nicotine intake per cigarette and therefore higher levels of nicotine addiction among more deprived individuals. While this study may have been subjected to reporting bias in the reported number of daily cigarettes smoked as well as unmeasured confounding from factors such as exposure to environmental tobacco smoke, the likelihood that there are differences in the amount of cigarette smoke inhaled between deprived and non-deprived individuals implies that deprived individuals may have a higher susceptibility and therefore higher cigarette smoke-associated risk of lung cancer per cigarette smoked than less deprived individuals.

A commercial geodemographic classification system, Experian's Mosaic Public Sector ™ classification tool classifies postcodes in the UK into 11 groups and 61 types based on demographics, lifestyle, education and values[101]. Mosaic profiling is done at finer levels than available deprivation markers and it has been used to identify socioeconomic differentials in several health-related behaviours including smoking[102]. There have however been no studies yet on the socioeconomic differentials in lung cancer incidence using Mosaic tool.

1.1.4.4 Lung cancer and depression

Individuals with depression and other mental disorders such as phobias and obsessive compulsive disorders are twice as likely to smoke and are more likely to smoke more heavily than those without mental disorders[103]. It has also been suggested that depression alters the body's immune system possibly leading to an increase in the risk of immune-related disorders such as cancer[104]. There have been a few studies on the association between depression and lung cancer but evidence from these studies have been conflicting. Whereas one study has described an independent association between depression and an increased lung

cancer risk[105], others have either failed to show any increase in risk among depressed individuals[106] or have found an increased lung cancer risk only among depressed individuals who are smokers[107 108]. A meta-analysis which synthesized the evidence from several prospective, general population-based studies of depression and cancer risk found no statistically significant association between depression and subsequent lung cancer risk[109].

1.1.5 Clinical presentation/symptoms

More than 90% of people with lung cancer are symptomatic at presentation [110 111] and they present either with symptoms relating to the primary tumour, non-specific symptoms or specific symptoms from metastatic disease. Most patients present with cough, dyspnoea, fatigue, chest pain, loss of appetite, weight loss and haemoptysis before a diagnosis of lung cancer is made[112-114]. Cough is the most frequently reported symptom, being reported by more than half of patients[113 115] whereas haemoptysis is relatively uncommon prior to diagnosis[113 115]. Although other respiratory symptoms are more common than haemoptysis before diagnosis, they are yet more common in other benign conditions[116] and have positive predictive values for lung cancer of less than 2%. The positive predictive value for lung cancer with haemoptysis is 2.4% to 7.5% and is higher when accompanied by other symptoms[113 117]. Table 1.1 summarises prevalence of symptoms prior to lung cancer diagnosis as reported in different studies.

Table 1.1. Summary of studies on symptoms reported before lung cancer diagnosis

Study	Population studied	Age of patients	Symptoms reported before lung cancer diagnosis	Source of data
Lovgren, M et al.* 2008[118]	314 patients diagnosed with primary lung cancer in Sweden in 2003	38 - 92 years	Cough - 41.8% Dyspnoea - 32.3% Thoracic related pain - 17.7% Weight loss - 32.1% Fatigue - 25.7% Appetite loss - 12.9% Haemoptysis - 5.1% Hoarseness - 2.2% Neurological symptoms - 10.9%	First reported symptoms based on hospital physician's documentation in medical records
Bjerager, M et al.* 2006[114]	84 patients newly diagnosed with lung cancer in Denmark between 1 April and 31 May 2003, and 1 September and 31 September 2003	34 - 83 years	Cough - 31.5% Dyspnoea - 16.9% Fatigue - 10.8% Weight loss - 7.7% Thoracic pain - 5.4% Haemoptysis - 4.6% Shoulder pain - 3.1% Hoarseness - 0.8%	Telephone interview with patients' general practitioners
Hamilton, J et al. 2005[113]	247 primary lung cancers diagnosed between 1998 and 2002 in all 21 general practices in Exeter, UK	Over 40 years	Cough - 65% Haemoptysis - 20% Weight loss - 27% Loss of appetite - 19% Dyspnoea - 56% Chest or rib pain - 42% Fatigue - 35%	Anonymised photocopies of general practice records for 2 years before lung cancer diagnosis
Corner, J et al. 2005[112]	22 patients diagnosed with lung cancer at two cancer centres in the south and north of England	Not stated	Cough - 68% Fatigue - 68% Appetite change - 64% Chest pain - 64% Shortness of breath - 59% Sleep changes - 59% Weight loss - 50% Haemoptysis - 41%	Interview study of patients' accounts and hospital and primary care records
Buccheri, G et al. 2004[119]	1,277 consecutive lung cancer patients seen in a single institution, over 14 years, from January 1989 to October 2002, in Italy	32 - 90 years	Cough - 50.0% Dyspnoea - 33.9% Chest pain - 31.5% Haemoptysis - 29.8% Chest infection - 19.7% Systemic symptoms - 49.3%	Hospital database of prospectively built records of lung cancer patients

Koyi, H et al.* 2002[111]	365 patients newly diagnosed with lung cancer in Sweden between 1997 and 1999	23 - 96 years	Cough - 24.9% Dyspnoea - 15.1% Fatigue - 14.2% Pain in thorax - 4.9% Back pain - 3.8% Haemoptysis - 3.2% Hoarseness - 2.0% Neurological symptoms - 2.3%	Data collected from patients through questionnaires after referral to the hospital respiratory department
Cromartie, RS et al. 1980[115]	702 patients treated with lung cancer in Charleston, South Carolina between 1960 and 1970	10-year age category under 40 to 80 and over	Cough - 64.2% Weight loss - 55.3% Pain - 52.7% Sputum - 44.4% Haemoptysis - 28.3% Malaise - 26.5% Dizziness - 4.0%	Hospital records of patients at two hospitals in Charleston, South Carolina
Weiss, W et al. 1978[120]	33 newly diagnosed cases of lung cancer in the cohort of 6,027 men enrolled in the Philadelphia Pulmonary Neoplasm Research project	45 years and older	Expectoration - 52% Chronic cough - 42% Dyspnoea - 52% Heaviness in chest - 3% Haemoptysis - 3% Chest pain - 3% Hoarseness - 6%	Symptoms recorded six months before lung cancer detection. Records were taken during the six-monthly screening done for all volunteers to this study.

* study shows percentage of patients who had the symptom at first presentation (percentage either increased or decreased before specialist referral)

1.1.6 Histological classification of lung cancer

Histologically, lung cancer can be broadly classified into 2 types - small cell lung cancers (SCLC) which accounts for approximately 20% of all cases of lung cancer, and non-small cell lung cancers (NSCLC) accounting for the remaining 80% of lung cancers[121]. NSCLC are further subdivided into 3 subtypes: squamous cell cancer, adenocarcinomas and large cell lung cancer and these make up 35%, 27% and 10% of all lung cancers in the UK respectively[122]. The histological subtype of lung cancer determines its treatment and prognosis. NSCLC overall has better prognosis than SCLC. While the main treatment for SCLC is chemotherapy, the treatment options for NSCLC are surgery, radical radiotherapy and/or chemotherapy depending on the stage of the disease, the lung function adequacy and suitability of the patient for treatment[122]. Surgical resection however remains the treatment of choice for NSCLC[123] and the prognosis is good for patients with localised disease.

1.2 The importance of early lung cancer diagnosis

The outcome of lung cancer depends on the tumour stage at diagnosis[1] and is favourable with patients diagnosed at the early stages with tumours that can be treated with surgery[124] or radical radiotherapy. Reports of five-year survival after treatment of clinical stage I disease range from 38% to 76%[124 125] while five-year survival for patients with clinical stage IIIB and IV disease is between 1% and 7%[124]. A study which demonstrated a difference in survival between patients with localised early-stage lung cancer who were surgically treated and those who were untreated (due to patients' refusal) showed that the longer survival after surgical resection of early-stage tumours may not be attributable to lead-time and length-time bias[126]. There are currently no widely available screening tests

for lung cancer although several randomised controlled trials on the use of CT screening to detect the disease in the early asymptomatic stages are under way[127-130].

Delays in diagnosis and in receiving definitive treatment have been recognized as important factors in the overall outcome of lung cancer treatment[131] and survival rates have been found to be higher in patients whose disease was diagnosed earlier[132] and who were referred to a specialist earlier[119]. Evidence from a single-centre study of 29 lung cancer patients in the UK showed that following a delay of 18 to 131 days (median of 54 days) between diagnosis of lung cancer at the oncology clinic and radiotherapy planning, an increase in cross-sectional tumour size was noted on CT scans and 21% of potentially curable cancers became incurable[133] .

1.3 The problem of late diagnosis in the UK

In the UK almost all of the population are registered with general practitioners (GPs) who act as the gate-keepers to all specialised health care. Most patients diagnosed with lung cancer present initially with symptoms to their GP[134]. If the GP suspects a diagnosis of lung cancer, further investigations can be carried out with subsequent referral to a specialist if the investigations are abnormal or if there is diagnostic delay[134]. In order to diagnose lung cancer early, it is important that the GP recognizes patients who have symptoms of lung cancer and are at the same time, at potential risk of having the disease. This should then be followed up with a chest x-ray investigation and/or prompt referral to a chest physician.

In the UK, two-thirds of lung cancer patients get to specialist care when they have metastatic disease with evidence of spread to other organs[135] and curative

treatment is no longer possible. As a result, less than 20% of patients seen by specialists have potentially surgically resectable tumours[136], 17% undergo surgery[137] and about half of these will be alive for up to 5 years.

In an effort to reduce the time taken to diagnose cancer in the UK, the UK department of health produced a white paper in December 1997 titled "The new NHS - modern, dependable" in which all patients with suspected lung cancer were guaranteed prompt access to specialist services in a hospital within two weeks of an urgent GP referral[138]. This policy took effect for lung cancer in April 2000 and was clearly aimed at earlier detection of cancers by reducing referral and treatment delays[139]. These are delays from referral for further care or diagnostic investigation to being seen in secondary care and delays from being seen in secondary care to treatment respectively[139]. The National Awareness and Early Diagnosis Initiative (NAEDI) established more recently in 2008 as part of the UK government's strategy to improve cancer outcomes, have also set up programmes to increase public awareness of symptoms of lung cancer[140]. There has however been less focus on delays in other stages of the diagnostic process including primary care delay (from first presentation in primary care to referral or initiation of diagnostic investigations). Data from the 2002 National survey of NHS patients showed that patient and primary care delays contributed more significantly to the total diagnostic delay than referral and secondary care delays[139]. The median primary care delay in the UK is 51 days[141] whereas in Sweden, it is 28 days[118].

A study of all cancer diagnoses within a 2-year period at the Bradford hospitals NHS trust [142] showed that only 23% of patients diagnosed with lung cancer were referred urgently by their GPs. Others presented through other pathways such as non-urgent referrals, emergencies or referral from other clinics. In a more recent study in Exeter, 45% of patients with lung cancer were referred by their GP to hospital respiratory departments for specialist investigation while 23% were

admitted to hospital as emergencies, many of which will have been for respiratory infections[141].

The problem with early recognition of lung cancer by GPs however, is that most symptoms of lung cancer can be found in benign conditions and the benign causes of these symptoms are more common in general practice[113 116]. Although lung cancer is a relatively common disease, GPs only encounter one to two new cases every year making it even more difficult to identify patients early[134 135]. As a result, lung cancer diagnosis is only suspected in 50% of the actual cases seen by GPs[1 143]. The remaining are seen in hospital either as emergencies or following referral for other non-respiratory conditions. Data published by the National Cancer Intelligence Network (NCIN) in 2010 shows that in England, more than a third of lung cancers (38%) are diagnosed on acute admission following an emergency presentation[144].

1.4 Current guideline for lung cancer diagnosis in UK primary care

To reduce variation in the availability and quality of NHS treatments and care, a special health authority - the National Institute for Health and Clinical Excellence (NICE), was set up by the department of health in 1999[145]. One of the functions of NICE is to produce evidence-based guidelines on the most effective ways to diagnose, treat and prevent disease and ill health.

The UK Department of Health in 2000, published cancer referral guidelines to facilitate appropriate referral between primary and secondary care for patients whom a GP suspects may have cancer[146]. These guidelines have since been reviewed and updated by NICE in February 2005[147] and subsequently in April 2011[122].

The following excerpt from the NICE guidelines for lung cancer published in 2011[122], lists the criteria on which general practitioners should select patients for specialist referral as well as indications for chest radiography in primary care:

1) Urgent referral for a chest x-ray should be offered when a patient presents with:

- Haemoptysis or
- Any of the following unexplained or persistent (lasting more than 3 weeks) symptoms or signs:
    - Cough
    - Chest/shoulder pain
    - Dyspnoea
    - Weight loss
    - Chest signs
    - Hoarseness
    - Finger clubbing
    - Features suggestive of metastasis from a lung cancer (for example in brain, bone, liver or skin)
    - Cervical/supraclavicular lymphadenopathy

2) If a chest X-ray or chest computed tomography (CT) scan suggests lung cancer (including pleural effusion and slowly resolving consolidation), patients should be offered an urgent referral to a member of the lung cancer multidisciplinary team (MDT), usually a chest physician.

3) If the chest X-ray is normal but there is a high suspicion of lung cancer, patients should be offered urgent referral to a member of the lung cancer team MDT, usually the chest physician.

4) Patients should be offered an urgent referral to a member of the lung cancer MDT, usually the chest physician, while awaiting the result of a chest X-ray, if any of the following are present:
    - Persistent haemoptysis in smokers/ex-smokers older than 40 years
    - Signs of superior vena cava obstruction (face or neck swelling with fixed elevation of the jugular venous pressure)
    - Stridor

The UK department of health warrants that following an urgent GP referral, patients with suspected lung cancer should be provided prompt access to specialist services within two weeks[146].

1.5 Mesothelioma

Mesothelioma is a highly fatal cancer[148] that principally affects the pleura (lining of the lungs) and the peritoneum (lining of the abdominal cavity). Pleural mesothelioma makes up over 90% of cases of mesothelioma with a known first site[135]. Unlike lung cancer where the most important risk factor is smoking, mesothelioma has been linked with exposure to asbestos fibres[148-151]. It is a rare disease although its incidence has been increasing [148 149]. It has a long latency period, the median interval between asbestos exposure and development of the disease being 30 years[135]. It also has a poor prognosis with a median survival of 7 to 9 months[135].

In the early stages of mesothelioma, there are no symptoms. When symptoms present, they are similar to those of lung cancer. They are non-specific and include persistent cough, dyspnoea, voice hoarseness, chest pain, fatigue and weight loss [152]. As a result, the same guidelines for urgent referral of lung cancer

patients also cover patients with suspected mesothelioma. However, there is currently no real potential for the cure of mesothelioma.

As mentioned in the introductory section of this chapter, the principal aim of the research covered in this thesis is to extensively investigate the interaction between GPs and patients prior to the diagnosis of lung cancer. Since mesothelioma is not covered in the scope of this research, the work in this thesis has therefore been done using data of patients with lung cancer with the exclusion of patients with a known diagnosis of mesothelioma.

1.6 Risk prediction scores

Risk prediction scores also known as predictive tools or predictive models, are tools designed to estimate or predict the risk of a patient developing some future clinical event by combining two or more items of patient and disease characteristics[153]. The main aim of these tools is to aid clinical decision-making by doctors, by providing objective estimates of risk probability as a supplement to other relevant clinical information[154]. They therefore have a potential to improve clinician performance with active guidelines for preventative and active care[155]. These tools can also be used to select patients with an increased risk of disease, for therapeutic research. Because clinical risk prediction tools are designed to guide clinical practice, it is important that they are reliable and accurate[156].

A few published papers have compiled characteristics which clinical risk prediction tools should conform to if they are to be clinically useful. Among the features listed in an editorial by Grady et al.[156], it was noted that these tools should be developed using data from patients who are representative of the population for whom the score will eventually be used, they should be relatively

easy to incorporate into routine clinical practice and most importantly, risk prediction tools will only achieve improved clinical outcomes if the predicted outcome could be prevented or delayed with effective treatment[156].

In a publication in the BMJ by Wyatt, J et al.[157], while highlighting the need for risk prediction tools to show evidence of clinical credibility and ability to support with decisions to guide patient care, the authors stated that some prediction models predict outcomes that are not clinically relevant or they do not predict outcomes in enough time to inform clinical decisions. They also reiterated the fact that in applying risk prediction tools in practice, it should be easy for clinicians to obtain all the patient data required without expending undue resources[157].

On the issue of evaluating and validating risk prediction models, the important characteristics of model performance described by Freedman, A et al. are calibration, discrimination and accuracy[158]. The calibration of a model is assessed by comparing the observed number of events with the expected number of events. It is commonly evaluated using the goodness-of-fit or chi-square test[158]. Good calibration of a model is especially important in planning population-level interventions. Model discrimination on the other hand, is a measure of how well the model can separate those who do and do not have the outcome of interest[159]. Discrimination is often measured using the receiver operating characteristic curve (ROC) curve or concordance statistic (*c* statistic)[159], which for binary outcomes is identical to the area under the ROC curve[160]. Model discrimination is particularly useful in assessing tools used in classifying into groups with and without disease such as in diagnostic testing. Calibration and discrimination are the two major components used to measure the performance of prediction models[159]. Model accuracy scores which include the positive and negative predictive values, can be used to evaluate how well a model categorises

specific individuals[158]. These measures however even with good sensitivity and specificity, may be low particularly with rare outcomes.

Several risk prediction scores such as the Framingham score[161], QRISK[162] and ASSIGN[163] are used to predict patients' cardiovascular risk based on socio-demographic, clinical and lifestyle characteristics. Other assessment scores which have been developed to inform decision-making in clinical practice include the Finnish diabetes risk score[164] to identify individuals at high risk of type 2 diabetes as well as scores to assess the status of the central nervous system of patients such as the Glasgow-coma scale[165], APACHE III[166] and the simplified acute physiology score (SAPS II)[167].

1.6.1 Cancer risk prediction scores

Some cancer risk prediction models have also been developed to aid the identification of individuals at high risk of cancer who may benefit from targeted screening or other intervention, to aid clinical decision-making, to develop benefit-risk indices, to estimate the population burden, cost and impact of specific interventions[158].

Models to predict the risk of breast cancer in women were developed using known risk factors such as age, age at menarche, age at first live birth, oestrogen use, number of previous benign breast biopsies and family history of breast cancer or other reproductive cancers in a first degree relative[168-172]. Following the discovery in the mid 1990s of the BRCA1 and BRCA2 genes which were found to increase susceptibility to breast cancer, models to predict the likelihood that an individual carried any of these genes predisposing to breast cancer were developed[173-175].

Risk prediction tools have also been developed for colorectal cancer. A model was developed using a weighted numerical score which was derived from weighting of primary symptoms and symptom complexes, and comprehensive patient consultation questionnaires[176]. This model was shown to have a high sensitivity and specificity and therefore high accuracy in prioritising patients with colorectal symptoms following referral by their general practitioners using the current NHS guidelines[176]. Another qualitative index of colorectal cancer risk was developed using information on age and modifiable factors such as alcohol use, smoking status and body mass index (BMI) to define 10 risk groups[177]. More recently, an absolute risk prediction model for colorectal cancer was developed using data from two population-based case-control studies[178]. By combining the risk estimates from age and several risk and protective factors, the risk factors which were found to be related to colorectal cancer risk and therefore included in the model were age, cancer-negative sigmoidoscopy/colonoscopy in the last 10 years, history of polyp in the last 10 years, family history of colorectal cancer in first-degree relatives, use of aspirin and non steroidal anti-inflammatory drugs, cigarette smoking, BMI, vegetable consumption and leisure-time vigorous activity. Exposure to hormone-replacement therapy (HRT) and oestrogen exposure were additional risk factors in the model for women. This model was found to be well calibrated when validated in a large prospective cohort study and has been judged to be clinically useful[179].

Several clinically applicable tools have been designed to predict the risk of prostate cancer in individuals. The cancer of the Prostate Risk Index (CAPRI) model was developed to predict a patient's overall risk of prostate cancer at biopsy by including four variables: prostate-specific antigen (PSA), digital rectal examination (DRE), race and age[180]. Despite the high predictive capability of the CAPRI test for prostate cancer, another prostate cancer risk assessment tool was developed using prostate biopsy data from men who participated in the Prostate

Cancer Prevention Trial (PCPT)[181]. Following logistic regression modelling of several risk factors for prostate cancer, the variables which were found to be predictive of prostate cancer in this model were higher PSA level, race/ethnicity, family history of prostate cancer, age, abnormal DRE, and a previous prostate biopsy. The risk equation that was developed from this model has been used to develop a clinical prostate cancer risk calculator that can be used by physicians or patients[181]. Also to assess the risk of prostate cancer in individuals, a clinical nomogram was constructed by assessing all known risk factors for prostate cancer in a cross-sectional study of men who had a prostate biopsy as well as some volunteers with normal PSA levels[182]. Results showed that in addition to age, family history of prostate cancer, ethnicity, PSA and DRE, other variables which were important to consider were urinary voiding symptoms and the ratio of free:total PSA.

An epidemiologic-genetic risk assessment model to project the individualized probability of developing bladder cancer was developed using data from a large case-control study of White individuals in the United States[183]. Cases were patients with newly diagnosed and histologically confirmed bladder cancer while controls were healthy individuals who had no previous history of cancer and who had come to clinic for their annual health check-ups. Cases and controls in the study were matched by age, sex and race. By incorporating the epidemiologic risk factors: duration of smoking (pack-years smoked), past exposures to diesel fuels, aromatic amines, dry cleaning fluids, radioactive materials and arsenic and the genetic factor: mutagen sensitivity (a phenotypic marker), this prediction model aids the identification of populations at high risk of bladder cancer.

A risk model to identify individuals at high risk of melanoma was developed using information such as age, host characteristics and geographical area[184]. Another risk model exists which estimates an individual's risk of melanoma using self assessed risk factors such as sex, age, hair colour, density of freckles,

history of severe sunburns in childhood and adolescence, raised moles on the arms and history of non-melanoma skin cancer[185].

To improve diagnostic test performance for ovarian cancer in the early stages, a model which predicts based on a combination of any two of the three tests: the Symptom Index (SI), serum Human Epididymis protein 4 (HE4) test and Cancer Antigen 125 (CA-125) test, was found to be more highly discriminatory of ovarian cancer than previously developed ovarian cancer detection tools[186].

Another cancer for which risk prediction models have been developed, is pancreatic cancer. The first risk prediction model for familial pancreatic cancer was the PancPRO which uses a Mendelian risk prediction approach to provide the probability that an individual carries a mutation in a pancreatic cancer susceptibility gene[187]. This model was developed using the Bayesian modelling framework and apart from providing information on mutation carrier probability, it also provides the absolute risk of pancreatic cancer for specified age intervals[187]. Another prediction model was developed to stratify risk of pancreatic cancer in chronic pancreatitis patients with focal pancreatic mass lesions with prior negative endoscopic ultrasound guided fine needle aspiration (EUS-FNA)[188]. In developing this model, logistic regression modelling was used to test the association of different cancer predictors with pancreatic cancer in a cross-sectional study of 138 consecutive chronic pancreatitis patients with focal pancreatic mass lesions who attended one of three hospitals for an initial EUS-FNA. Based on findings from this model, the predictors of pancreatic cancer were age, mass location, mass number, direct bilirubin and cancer antigen 19-9 (CA 19-9)[188].

A few risk prediction models have also been developed for lung cancer[189-193], however these are described in more detail in the following section.

Although the cancer risk prediction scores described above have all been shown in their respective studies to be clinically useful, it is difficult to ascertain those that are routinely used in clinical practice and their actual usefulness in practice.

1.6.2 Lung cancer risk prediction scores

Several models have been developed to estimate the risk of lung cancer using individual baseline risk factors.

The Bach model was developed in 2003 to determine predictable variations in the risk of lung cancer among smokers[189]. The model was created using data from 18,172 individuals aged between 45 and 69 years who had a documented history of current or former smoking and who were enrolled in the Carotene and Retinol Efficacy Trial (CARET), a randomised trial of lung cancer prevention. Lung cancer predictors that were analysed in the development of this model were age, sex, prior history of asbestos exposure, smoking duration, average daily number of cigarettes smoked and duration of smoking abstinence for former smokers. The authors of this model did not consider other possible predictors of lung cancer such as history of chronic obstructive pulmonary disease (COPD), chest x-ray findings, exposure to second hand smoke, radon exposure and type of asbestos exposure because these were not recorded in the CARET study. Using Cox proportional hazards regression modelling, the associations between the predictors of lung cancer and a diagnosis of lung cancer or death in the absence of a diagnosis, were estimated and used to derive 1-year risk models for the prediction of lung cancer. The calibration of the models were assessed by comparing the observed and predicted rates of lung cancer across different risk cut-offs and this was validated by assessing the extent to which the model could predict cancer in an independent CARET study site. To determine variation in lung cancer risk among smokers, the predicted 10-year lung cancer risk among

55 to 74 year old current or former smokers who were enrolled in an ongoing low dose CT trial were examined. In the one year risk model to predict lung cancer diagnosis, significant predictors of lung cancer among current or former smokers were duration of smoking, average number of cigarettes smoked daily, duration of abstinence, age and history of asbestos exposure. With a concordance index of 0.72 on comparing observed and predicted rates of lung cancer, this model was found to be internally valid and well calibrated. Validation of the model in an independent CARET study site showed that the observed rates of lung cancer were closely matched with that predicted by the model. A major limitation of this model however results from the fact that it was derived using data of participants enrolled in a clinical trial of lung cancer prediction. Also, since all the participants were current or former smokers, this model is only applicable to smokers, a subset of individuals at risk of lung cancer.

An absolute risk prediction model for lung cancer, The Spitz model[190] published in 2007, extended the work of Bach et al. and incorporated additional risk factors apart from smoking and asbestos exposure, in the development of the risk prediction model. The model was derived using epidemiologic data from a large case-control study of 1,851 newly diagnosed, histologically confirmed cases of lung cancer and 2,001 healthy controls, matched by age, sex, ethnicity and smoking status to the cases. Information was collected on smoking history (including exposure to environmental tobacco smoke (ETS)), age at smoking cessation for former smokers, family history of any cancer and of smoking-related cancers in first-degree relatives, exposure to wood dust, asbestos exposure, previous history of respiratory disease and hay fever, and then logistic regression models were constructed separately for never, former and current smokers. Multivariable logistic regression analyses were used to construct the final risk models to determine variables that were predictive of lung cancer. Variables which were found to be associated with lung cancer in never smokers

were exposure to ETS and family history of any cancer. In current and former smokers, lung cancer was associated with dust exposure, no previous history of hay fever, previous history of emphysema, family history of any cancer or tobacco-related cancers, smoking intensity and age at smoking cessation (in former smokers). Lung cancer was also found to be associated with exposure to asbestos in current but not former smokers. On validating the risk models, the concordance statistics in validation sets for the never, former and current smokers were 0.57, 0.63 and 0.58 respectively. Overall, the discriminatory accuracy of this model was found to be modest but was consistent with those from other risk-prediction models. A drawback of the model however, is that cases and controls were frequently matched on smoking status therefore affecting the importance of smoking as a risk factor. Also, the model was derived using data from non-Hispanic whites which limits its generalisability to other ethnic groups.

To compensate for the modest precision of the Spitz model, an expanded Spitz model[194] which incorporated select markers of DNA repair capacity was developed and published in 2008. This model was developed using assay data from 725 lung cancer cases and 615 controls - a subset of cases and controls from the original analysis. All the cases and controls included in this analysis were current or former smokers. Multivariable modelling were carried out using the variables in the original Spitz model with the addition of the biomarker assays. Comparison with the original Spitz model showed an improvement in the discrimination of the expanded Spitz model, with concordance statistics of 0.70 and 0.73 for former and current smokers respectively. The authors of this model however cautioned that the biomarker assays were time consuming and require some level of technical expertise. The model may therefore be applied in a controlled academic setting but it is not feasible to implement in the general population.

Another lung cancer risk prediction model - The Liverpool Lung Project (LLP) risk model[191], was developed to project the individual 5-year absolute risk of developing lung cancer using data from a case-control study of lung cancer in Liverpool, UK. Information on socioeconomic and demographic characteristics, medical history, family history of cancer, history of tobacco consumption and lifetime occupational history were collected from all 579 lung cancer cases aged between 20 and 80 years of age, and 1157 controls who were matched by age and sex with the cases. Conditional logistic regression models were constructed to identify the variables that were associated with lung cancer in multivariate analysis. Variables which were found to be associated and therefore included in the risk-prediction model were individuals' age, sex, duration of smoking, family history of lung cancer, occupational exposure to asbestos, prior history of pneumonia and prior diagnosis of any cancer other than lung cancer. Although the authors had yet to validate the model using independent data, assessment in the case-control dataset showed good discrimination between cases and controls.

Compared to the Bach and Spitz models, the LLP model has been found to correctly identify a higher proportion of lung cancer patients but it also has a much higher rate of false positives and therefore falsely identifies more individuals who have low risk of lung cancer than the previous two models[195]. At a cut-off value to capture 62% of cases of lung cancer, the LLP model falsely identifies 30% of non-lung cancer controls. Despite the shortcomings of this model, it is currently being employed to identify individuals at risk of lung cancer for the UK lung screen (UKLS) trial of low dose CT screening for lung cancer[128].

Lung cancer risk prediction models were also developed in 2011 using prospective data from 55 to 74 year old men and women enrolled in the Prostate, Lung, Colorectal and Ovarian Cancer screening Trial (PLCO) - a randomised clinical trial designed to study the effect of screening modalities on

cancer mortality rates[193]. Four annual chest radiographs were done for subjects in the screening arm of the study while other subjects in the control arm were given regular care as recommended by their physicians. Risk prediction models were developed using data from control subjects who were cancer-free at the time of entry into the study. One model was developed using data from 70,962 control subjects and another was developed using a sub-cohort of 38,254 control subjects who were ever smokers. The models were validated with 44,223 subjects who were in the intervention arm of the PLCO trial. Potential predictors of lung cancer which were analysed in the development of these models were age, socioeconomic status (education), race, sex, family history of lung cancer, body mass index, history of COPD, history of chest x-ray in the past 3 years and smoking history (including smoking intensity, quit time for former smokers, and pack-years smoked). In the first model that was developed using all eligible control subjects, the variables that were associated with lung cancer were age, educational attainment, BMI, family history of lung cancer, history of COPD, history of chest x-ray in the past 3 years, current smoking status, pack-years smoked and smoking duration. In the second model based on only the ever-smokers in the control arm of the study, the variables that were associated with lung cancer were age, pack-years and quit-time. On validation of the models using subjects in the intervention arm of the study, both models demonstrated high discrimination and calibration. The authors however acknowledged that several potentially useful predictors which had been included in previous models (for example, exposure to occupational carcinogens and history of adult pneumonia) were not included in the model. Also, the external validation sample came from the same PLCO referent screening trial population from which the prediction models were developed. The models may therefore not be generalisable and the discrimination may not be as good, when applied to other populations.

All the lung cancer risk models discussed thus far have estimated the risk of lung cancer using individual baseline risk factors. In a case-control study of 3,197 patients with lung cancer and 1,703 cancer-free controls, the discriminatory power of the Bach, Spitz and LLP models were assessed and in this study, the positive predictive values of the models were found to be high overall (>75%), indicating that they had a high probability of accurately categorising affected participants[195]. However, all three models had relatively low negative predictive values (between 45% and 56%) and therefore had a moderately low probability of accurately categorising unaffected participants in the study. The need to identify other important risk factors (than smoking) that have a different distribution in lung cancer patients compared to those who will not develop lung cancer were suggested as a means of improving the discriminatory power of these models[195].

A population-based case-control study using data from all 21 general practices in Exeter, UK, showed that several symptoms were independently associated with lung cancer up to 180 days before diagnosis[113]. However, if lung cancer risk were to be predicted using only alarm symptoms without the inclusion of other baseline risk factors, more than 75% of cases would be excluded[117].

A recently developed risk algorithm took account of baseline risk factors and symptoms in primary care as a means of identifying patients at high risk of lung cancer[192]. This algorithm was developed using data of primary care patients in QResearch, a computerised database of primary care records in England and Wales. Derivation of the algorithm was done using data from two-thirds of the practices while it was validated using the remaining one-third of practices. Predictor variables that were analysed to develop this risk algorithm were: current GP consulting for clinical symptoms of haemoptysis, loss of appetite and weight loss, recent GP consulting within the past 12 months for symptoms of cough, dyspnoea, tiredness, and hoarseness, body mass index, smoking status,

history of COPD, Townsend deprivation score, family history of lung cancer, previous cancer diagnosis, previous history of asthma, previous history of pneumonia, asbestos exposure and history of anaemia. The study outcome was the incident diagnosis of lung cancer during the subsequent 2 years. Variables which were found to be predictive of lung cancer were age, body mass index, Townsend score, smoking status, COPD, current GP consulting for haemoptysis, current loss of appetite, current weight loss and recent consultation for cough. A prior diagnosis of cancer was predictive of lung cancer only in females. Validation of the algorithm using the remaining one-third of QResearch practices, showed that it was well calibrated. In developing this algorithm, symptoms that were recorded in the period preceding lung cancer diagnosis when patients would likely be undergoing investigations for suspected cancer were not excluded. Therefore in these clinical situations where GPs may already be investigating patients' symptoms for possible lung cancer diagnosis, the association made by this algorithm between patients' current consulting for clinical symptoms and the incident diagnosis of lung cancer in the subsequent 2 years may only be stating the obvious. Based on the fact that a fundamental characteristic of good clinical prediction tools is not simply to predict clinical outcomes but to also provide opportunity for the outcomes to be prevented or delayed[156], it follows therefore that the clinical usefulness of this risk-prediction algorithm may be limited.

1.7 Summary of the evidence on lung cancer risk assessment scores

Most of the models that have been developed to predict the risk of lung cancer, use baseline risk factors to estimate an individual's risk of lung cancer. A case-control study which assessed the discriminatory performances of the Bach, Spitz and LLP models found these to be modest[195]. The need to identify other important risk factors (other than smoking) that have a different distribution in

lung cancer patients compared to those who will not develop lung cancer was suggested as a means of improving the lung cancer risk discriminatory power of these models[195].

Clinical symptoms in primary care have been shown to be independently associated with lung cancer up to 180 days before diagnosis[113]. However, to attempt to predict lung cancer risk using only alarm symptoms without the inclusion of other baseline risk factors would exclude more than 75% of cases[117]. The only lung cancer predictive model which has been developed using a combination of patients' baseline risk factors and symptoms in primary care did not exclude symptoms that were reported to the GP in the period leading up to lung cancer diagnosis in the model development. This model has the tendency to predict lung cancer in patients who are already being investigated for possible lung cancer by their GPs and its clinical usefulness is therefore limited.

1.8 Rationale of the thesis

Lung cancer survival is poor in the UK and delay in diagnosis has been recognised as an important factor contributing to this. There are currently no screening tests for lung cancer so earlier diagnosis is vital in order to improve treatment outcomes and overall survival. Several studies have suggested that the delay in lung cancer diagnosis may be partly due to late presentation of lung cancer symptoms by patients to general practice[112 139 196] while others suggest that delay in symptom recognition in general practice[113 139 197] may be to blame. While there is a well recognised need to address the issue of late presentation of symptoms of lung cancer to general practice, there is also a pressing need for research to understand the interactions between GPs and patients in primary care before the diagnosis of lung cancer is made[198].

To diagnose lung cancer earlier and improve survival, it is important that signs and symptoms are recognized promptly in primary care, especially among individuals who are at high risk of lung cancer[134]. The NICE referral guidelines developed to facilitate urgent referral of suspected lung cancer cases were based on a weak evidence base[142] and in some instances, may be misleading[134]. The low predictive power of the referral guidelines as a marker for lung cancer was demonstrated in a study conducted in a hospital trust in England which showed that only 42% of the patients urgently referred by their GPs for suspected lung cancer based on the criteria for urgent referral were diagnosed with lung cancer [142].

To further aid the identification and subsequent early investigation by GPs, of patients who are at high risk of lung cancer from those who present with non-specific symptoms associated with other illnesses, there is the need for a predictive tool that combines patients' baseline risk factors and early symptoms of lung cancer. The only risk prediction model which was developed using a combination of symptoms and other lung cancer baseline risk factors has been shown to be methodologically flawed[192] and necessitates further work to accurately determine the predictors of lung cancer in primary care.

Drawing on all the points above, this thesis aims to validate the use of lung cancer data from a large computerised database of UK general practice for research and then extensively explore the GP-patient interaction before lung cancer diagnosis over a 10 year period, with the aim of identifying factors which are associated with lung cancer. The database used for analysis in this thesis had been linked with Experian's Mosaic Public Sector ™ classification, a geo-demographic social marketing tool that classifies all households and postcodes within the UK into 61 types and 11 groups based on their typical demographics, consumer behaviour, lifestyle and attitudes[101]. This has enabled the identification of particular sectors of the UK where lung cancer incidence is highest with a view

to enabling focused and targeted public health interventions to improve lung cancer awareness and care. Following the identification of the predictors of lung cancer in general practice, the possibility of developing a predictive score will be explored. It is hoped that the results from this thesis will inform guidelines that will aid diagnosis and care of patients with lung cancer in primary care.

1.9 Thesis objectives

The aim of this thesis is to investigate the GP-patient interaction in the period before lung cancer diagnosis, with a view to determining the possibility of developing a predictive score for lung cancer that could be used to aid earlier diagnosis of future cases. The objectives that have been set in order to achieve this aim are to:

- Determine the validity of THIN database for studies on lung cancer in the UK and at the same time, identify the sectors of UK society where the incidence of lung cancer is highest.
- Explore the differences in the risk of lung cancer among different sub-groups of patients with similar recorded levels of cigarette smoke consumption, using a dataset of lung cancer cases and controls matched by age (year of birth), sex and general practice.
- Determine the independent predictors of lung cancer in a case-control dataset, matched only by practice and then develop a predictive score for lung cancer
- Investigate the validity of the lung cancer predictive score in an independent dataset of THIN patients

1.10 Outline of thesis sections

The following chapters of this thesis discuss the database that was analysed in this thesis, a description of how the analysed datasets were prepared and three chapters on studies that address the objectives of the thesis. The content of each chapter is detailed below:

Chapter 2: Description of The Health Improvement Network (THIN) database, key dates in THIN, the process of data preparation and an account of how the lung cancer population was derived, ethical approval for the studies in this thesis and statistical software.

Chapter 3: In the first study, the characteristics of patients with lung cancer in THIN is summarised and the validity of THIN for lung cancer research is assessed. The potential use of Experian's Mosaic Public Sector ™ classification to facilitate disease ascertainment by identifying particular sectors of the UK society where lung cancer incidence is highest, is also explored.

Chapter 4: A matched case-control dataset is developed for the primary purpose of piloting the methods for the development of a lung cancer score. This is followed by a description of studies using this dataset to explore differences in the risk of lung cancer, firstly among deprived compared to non-deprived smokers, then among depressed compared to non-depressed smokers. Lastly, a study which used the thesis dataset to assess the difference in smoking-associated risk of lung cancer among men and women is summarised.

Chapter 5: An unmatched dataset of lung cancer cases and controls is developed and this is used to identify the socio-demographic and early clinical factors predictive of lung cancer in general practice.

Chapter 6: A lung cancer risk-predictive model is derived using the lung cancer predictors identified in chapter 5, and this is followed by the validation of the model in an independent cohort of patients in THIN.

Chapter 7: A summary of the main findings in the thesis, what it adds to the existing knowledge of lung cancer and suggested future research.

Chapter 2. Description of the dataset and derivation of the lung cancer population

This chapter describes The Health Improvement Network (THIN) database - the computerised database of general practice records that was used for the analyses in this thesis. It describes the component files that make up the database including Experian's MOSAIC Public Sector™ variable which had been linked with the database and this is followed by a step-by step account of the process that was used to prepare the dataset for all the analytic work that was undertaken. The steps taken to derive the final population of lung cancer cases are also described and lastly, a brief summary of the ethical approval for this study, funding and the statistical software used for all the analyses in this thesis.

2.1 The Health Improvement Network database

The Health Improvement Network (THIN)[199] is a computerised longitudinal database of anonymised primary care records from the UK. In October 2009 when the data for this study were compiled, THIN contained data from 446 general practices across the UK with a total of 8.2 million patients. More than 3.2 million of these patients were actively registered and could be prospectively followed while the remaining patients who had historic data, had either left the practice or died.

In May 2002, THIN was set up through the collaboration between Epidemiology and Pharmacology Information Core (EPIC) - a research organisation that for many years facilitated access to the General Practice Research Database (GPRD) for medical research, and In Practice Systems (InPS)[200] who provide Vision software - the general practice interface software to about 2000 general practices in the UK[199 201]. On joining the THIN scheme, general practices

contribute data prospectively using the practice's Vision computer software without interruption to normal practice operation. All retrospective data are also uploaded into the patients' records, most of which were recorded using the Value Added Medical Products (VAMP) practice management system[202] that was used in the GPRD. Incremental data are downloaded monthly by EPIC, processed and added to existing data to create the THIN data that is made available to researchers[199].

2.1.1 Structure of THIN database

The database contains all records relating to patients such as information on signs and symptoms, diagnosis, prescriptions, routine health checks, preventative health information and referrals to secondary care[203]. These data are contained in four standard files - patient, medical, therapy and additional health data (AHD) files; as well as two linked files - postcode variable indicators (PVI) and dosage records (Table 2.1). All entries are organised by practice and each patient has a unique identifier to enable linking of patients' records across all files. Data are entered into THIN using Read codes which are a standard hierarchical classification system used by general practitioners in the UK to record patient medical information[203]. Table 2.2 shows the formats for the different files in THIN database.

Table 2.1. Structure of THIN database

	THIN data file	Information recorded
Standard files	Patient	Patient demographics, registration details such as date of registration with practice, date of transfer out of practice, date of death.
	Medical	Symptoms, diagnosis, interventions recorded in primary care as well as discharge summaries from hospital and letter from specialists
	Therapy	Prescriptions issued to patients (including formulation and strength of medications, dose and quantity)
	Additional health data	Lifestyle data, test results, details of death, immunizations and physical measurements
Linked files	Dosage	Dosage instructions
	Postcode variable indicators	Postcode-linked area based socioeconomic, ethnicity and environmental indices

Table 2.2 Example of file formats in THIN

The tables below show the formats of the different data files in THIN. This is followed by a description of what the different fields represent.

a. Patient data file

combid	prac	patid	patflag	yob	hh	regdate	regstat	xferdate	regrea	deathdate	amrdate	visdate	pracstart	lastdate
a6732 00??	a6732	00??	A	1961	48661	19881102	2	0			20010101	20070711	20010101	20090729
a6732 00?0	a6732	00?0	A	1970	37682	19960830	5	20050604	3		20010101	20070711	20010101	20090729
a6732 00?1	a6732	00?1	A	1938	22561	19410815	2	0			20010101	20070711	20010101	20090729
a6732 00?2	a6732	00?2	A	1952	21402	20020903	99	20061018	1	20061018	20010101	20070711	20010101	20090729
a6732 00?3	a6732	00?3	A	1912	10641	19581010	2	0			20010101	20070711	20010101	20090729

b. Medical file

combid	prac	patid	evdate	medcode	medflag	diagnosr	source	episode	NHSspec	locate	textid
a6831 01OF	a6831	01OF	19920518	8B41.00	R	00000D	0	0	000	I	1yYe
a6831 01OF	a6831	01OF	19921020	173..00	R	00000D	0	1	000	I	22Ie
a6831 01OF	a6831	01OF	19930527	ZZZZZ00	R	00000D	L	4	000	I	0rdr
a9928 02Y8	a9928	02Y8	20031007	8B28.00	R	00000c	0	1	000	I	0000001
a9928 02Y8	a9928	02Y8	20031007	8B63.12	R	00000c	R	0	000	I	0000001

c. Therapy file

combid	prac	patid	rxdate	drugcode	therflag	doscode	rxqty	rxdays	private	prscber	rxtype	opno	bnf	seqnoiss	maxnoiss	packinfo	dosgval
a6732 009Z	a6732	009Z	20020228	95617998	Y	0001826	60	0	N	000000B	1	0	2080200	1	0	0	-1
a6732 00?e	a6732	00?e	20061012	90841998	Y	0003563	30	0	N	0000004	1	0	1060400	0	0	0	1
a6732 00BF	a6732	00BF	20060106	86990998	Y	0000200	56	0	N	0000004	1	0	2060200	14	0	0	1
a6732 00Sr	a6732	00Sr	20041020	97085997	Y	0000424	2	0	N	000000A	1	0	3010101	11	0	0	-1
a6732 00dO	a6732	00dO	20060113	97217998	Y	0000200	56	0	N	0000003	1	0	2020100	29	0	0	1

d. Additional health data (AHD) file

combid	prac	patid	evdate	ahdcode	ahdflag	ahdval1	ahdval2	medcode	source	NHSspec	locate
a6732 009Z	a6732	009Z	20050413	1003040001	R	00000000	00000000	137S.00	0	000	I
a6732 00?e	a6732	00?e	19931001	1003050000	R	0	Y	1362.00	0	000	I
a6732 00BF	a6732	00BF	20031013	1003050000	R	0	Y	1363.00	0	000	I
a6732 00Sr	a6732	00Sr	20071003	1003050000	R	28	Y	136..00	0	000	I
a6732 00dO	a6732	00dO	20031007	1003050000	R	0	N	1361.00	0	000	I

Description of the fields in THIN dataset

combid	Patient identifier which is unique within the entire THIN dataset (combination of practice id and patient id)
prac	Practice identifier
patid	Patient identifier which is unique within practice
patflag	Flag which indicates the integrity of the data for that patient
yob	Year of birth
hh	Household identifier
regdate	Date of patients' registration with the practice
regstat	Registration status (for example, 2 = permanent, 5 = transferred out, 15 = walk-in centre, 99 = death)
xferdate	Date when patient was transferred out from the general practice
regrea	Additional registration information (for example, 1 = death, 3 = internal transfer, 23 = registration cancelled)
deathdate	Date of death
amrdate	Acceptable mortality recording date. This denotes the year when the practices' is deemed to be reporting all-cause mortality based on predicted numbers from national statistics. It is a measure of when data records from the practice became broadly reliable
visdate	Date when the practice started to use the vision software
pracstart	Earlier of amrdate and visdate
lastdate	Date of last data collection from the practice
evdate	Date of the event recorded
medcode	Read codes which are coded clinical language
medflag	Flag indicating integrity of the clinical record (for example, R = acceptable record, E = source invalid)
diagnosr	Identifier of person entering record
source	Variable indicating origin of record
episode	Episode type (for example, 1 = First ever episode, 2 = new event)
NHSspec	Secondary care specialty
locate	Location of consultation
textid	Link to free text comment
rxdate	Prescription date
drugcode	Multilex drug code
therflag	Flag indicating integrity of the record
rxqty	Quantity prescribed
rxdays	Duration of the prescription in days
prscber	System assigned identifier or prescriber
rxtype	Variable denoting if acute or repeat prescription
opno	Number of original packs ordered
bnf	BNF (British National Formulary) 1 code
seqnoiss	Issue sequence number for repeat prescriptions
maxnoiss	Maximum number of issues for repeat prescriptions
packinfo	pack size information
dosgval	The calculated daily dosage
ahdcode	AHD (Additional Health Data) code
ahdflag	Flag indicating integrity of record
ahdval1	AHD value 1
ahdval2	AHD value 2

2.1.2 Quality of data in THIN

To ensure high quality data, each general practice that contributes data to THIN receives expert advice and training on quality data recording, and audits are performed to ensure that practices are recording data to a sufficiently high standard[199]. The Health Improvement Network has been demonstrated to have high quality data[201] and several published validation studies have shown the database to be valid for pharmacoepidemiology research[204 205] with a high degree of completeness and accuracy for records of several disease diagnoses[206-208]. A recent study showed that the records of incidence of all cancers in THIN were consistent with that reported in cancer registries[209]. More specifically, the observed recording rates of pancreatic, colorectal and lung cancers in THIN which were lower than cancer registry rates between 2000-2002, had increased to approximately 80% of registry years in later years, after 2004[209].

In addition to the validation studies, there have been several publications to date from medical research conducted using THIN data and these include: a nested case-control study published in 2006, which assessed the risk of diabetes associated with prescribed glucocorticoids[210]. A study published in 2010 estimated the incidence of dementia and survival after a primary care diagnosis of dementia[211] using data from 353 UK general practices contributing to THIN. Another study published in 2011 assessed the trends in long-term oral glucocorticoid prescription in the UK[212]. More recent research conducted using THIN database includes a study which aimed to determine the prevalence of underlying disease in men with erectile dysfunction receiving phosphodiesterase type 5 inhibitors in the UK[213] and another study which evaluated the risk of myocardial infarction and death from coronary heart disease after discontinuation of low dose aspirin in individuals with a previous history of cardiovascular events[214].

2.1.3 Strengths and weaknesses of THIN

THIN dataset is well known to be a source of high quality data for epidemiological studies. As with any other data source, it inevitably has some limitations. This section highlights some of the key strengths of THIN and also considers some limitations of using data from general practice databases for research. The design of the study in this thesis was made following consideration of the limitations of the dataset.

2.1.3.1 Size

THIN is a large dataset containing primary care records of approximately 5.8% of the UK population in 2009. Due to its large size, an adequate number of patients with relatively rare outcomes can be identified. It is therefore a good source of data for studies investigating rare diseases such as lung cancer.

2.1.3.2 Scope of data recording

Data in THIN are collected during routine general practice consultation without interruption to normal practice operation and therefore reflects "real-life"[199]. A drawback of THIN and other computerised routine general practice databases however, is that the data recorded in GPs' medical record system are collected primarily for the purpose of patient and practice management and not for research[199]. In routine medical care, the recording of information tends to be selective rather than comprehensive[215] and there is a tendency therefore for GPs to record only the information that they require or which they consider relevant to the patients' condition at the time of consultation. Not only does this imply that certain information that is vital for research may be not be obtainable from the dataset, but there is a likelihood of ascertainment bias and misleading associations arising from differential surveillance of patients[215].

2.1.3.3 Representativeness

All individuals residing in the UK have a right to be registered with a GP and the care provided at general practices are free of charge. To a large extent, data from general practices represent all sections of the UK population. Evidence however suggest that there may be a slight over-representation of practices from more affluent areas in THIN[216 217]. Despite these assertions, validation studies have found data in THIN to be widely representative of the UK population. Also as mentioned above, records of cancer incidence in THIN have been found to closely resemble records of incidence in the cancer registries[209].

2.1.3.4 Temporality

The method of data entry into the THIN database enables the prospective follow-up of patients and an ability to identify the timing of data collection in relation to the outcome of interest. It is therefore possible to establish the cause-effect path and this overcomes any bias due to loss of temporality. Information in THIN is also continually updated and allows the investigation of any effects of new drugs or interventions on the outcome.

2.1.3.5 Diagnostic criteria

A limitation with routine general practice data is that the perception of morbidity may vary between different practices and even within GPs in the same practice[218]. Analyses done using these data are therefore based on the recorded diagnosis being the best diagnostic formulation. The accuracy and variation of lung cancer diagnoses in different general practices has not been assessed. However considering that lung cancer diagnosis is made following investigations in primary care or by the chest physician in secondary care, it is unlikely that

there will be significant variation in the GP diagnostic criteria for lung cancer in this study.

2.1.4 Measures of socioeconomic status in THIN

In addition to routine health information, patients' records in THIN have area-level information such as strategic health authority (SHA) regions. There are two area-based measures of socioeconomic status available in the THIN dataset. These are the Townsend quintile of deprivation and the Mosaic public sector ™ classification.

Townsend quintile of deprivation is a widely used and well validated measure of deprivation[219]. It measures the level of material deprivation for each output area (corresponding to approximately 125 households with similar characteristics[220]) using the following four indicators derived from census data[221]:

- Unemployment: the percentage of economically active residents aged 16-64 who are unemployed
- Car ownership: The percentage of private households who do not possess a car
- Home ownership: The percentage of private households not in owner occupied accommodation
- Overcrowding: The percentage of private households in overcrowded accommodation

Postcodes in the UK are matched to their output-area Townsend deprivation quintiles and during THIN data collection, the Vision software maps the anonymous id of patients in THIN to these quintiles using the patients' postcode.

Records of patients in THIN were also linked with another measure of socioeconomic status - The Mosaic public sector ™ variable. This is a lifestyle

segmentation tool originally designed by Experian to profile customers for the purpose of market research[222]. Mosaic Public Sector™ refines areas at a higher level than available deprivation markers by using data from 400 variables to classify all postcodes within the UK into 61 types, each type being a member of one of 11 groups. Of the 400 variables used to develop a Mosaic Public Sector™ profile, 54% are sourced from the 2001 Census while the other 46% are derived from sources such as the Experian Lifestyle Survey, consumer credit databases, the electoral roll, shareholder registers, Land registry data, Council Tax information, the Hospital Episode Statistics, the British Crime Survey, Expenditure and Food Survey and other sources[222]. Mosaic Public Sector™ classification is based on typical demographics, behaviour, consumer values, consumption patterns, lifestyle, education and social and health-related attitudes[101]. Table 2.3 shows the Mosaic public sector ™ classification into 11 groups and 61 types, as well as a concise description of the characteristics of individuals in these Mosaic types.

Table 2.3. Mosaic Public Sector™ groups and types

Code	Mosaic Public Sector™ group	Code	Mosaic Public Sector™ type
A (Symbols of success)	Career professionals living in sought after locations	A01	Financially secure people living in smart flats in cosmopolitan inner city locations
		A02	Highly educated senior professionals, many working in the media, politics and law
		A03	Successful managers living in very large houses in outer suburban locations
		A04	Financially secure couples, many close to retirement, living in sought after suburbs
		A05	Senior professionals and managers living in the suburbs of major regional centres
		A06	Successful, high earning couples with new jobs in areas of growing high tech employment
		A07	Well paid executives living in individually designed homes in rural environments
B (Happy families)	Younger families living in newer homes	B08	Families and singles living in developments built since 2001
		B09	Well qualified couples typically starting a family on a recently built private estate
		B10	Financially better off families living in relatively spacious modern private estates
		B11	Dual income families on intermediate incomes living on modern estates
		B12	Middle income families with children living in estates of modern private homes
		B13	First generation owner occupiers, many with large amounts of consumer debt
		B14	Military personnel living in purpose built accommodation
C (Suburban comfort)	Older families living in suburbia	C15	Senior white collar workers many on the verge of a financially secure retirement
		C16	Low density private estates, now with self reliant couples approaching retirement
		C17	Small business proprietors living in low density estates in smaller communities
		C18	Inter war suburbs many with less strong cohesion than they originally had
		C19	Attractive older suburbs, typically occupied by families but with increasing singles and childless couples
		C20	Suburbs sought after by the more successful members of the Asian community
D (Ties of community)	Close-knit, inner city and manufacturing town communities	D21	Mixed communities of urban residents living in well built early 20th century housing
		D22	Comfortably off manual workers living in spacious but inexpensive private houses
		D23	Owners of affordable terraces built to house 19th century heavy industrial workers
		D24	Low income families living in cramped Victorian terraced housing in inner city locations
		D25	Centres of small market towns and resorts containing many hostels and refuges
		D26	Communities of lowly paid factory workers, many of them of South Asian descent
		D27	Multi-cultural inner city terraces attracting second generation settlers from diverse communities
E (Urban intelligence)	Educated, young, single people living in areas of transient populations	E28	Neighbourhoods with transient singles living in multiply occupied large old houses
		E29	Economically successful singles, many living in privately rented inner city flats
		E30	Young professionals and their families who have gentrified terraces in pre 1914 suburbs
		E31	Well educated singles and childless couples colonising inner areas of provincial cities
		E32	Singles and childless couples in small units in newly built private estates
		E33	Older neighbourhoods increasingly taken over by short term student renters
		E34	Halls of residence and other buildings occupied mostly by students

F (Welfare borderline)	People living in social housing with uncertain employment in deprived areas	F35	Young people renting hard to let social housing often in disadvantaged inner city locations
		F36	High density social housing, mostly in inner London, with high levels of diversity
		F37	Young families living in upper floors of social housing
		F38	Singles, childless couples and older people living in high rise social housing
		F39	Older people living in crowded apartments in high density social housing
		F40	Older tenements of small private flats often occupied by highly disadvantaged individuals
G (Municipal dependency)	Low income families living in estate based social housing	G41	Families, many single parent, in deprived social housing on the edge of regional centres
		G42	Families with school age children, living in very large social housing estates on the outskirts of provincial cities
		G43	Older people, many in poor health from work in heavy industry, in low rise social housing
H (Blue collar enterprise)	Upwardly mobile families living in homes bought from social landlords	H44	Manual workers, many close to retirement, in low rise houses in ex-manufacturing towns
		H45	Older couples, mostly in small towns, who now own houses once rented from the council
		H46	Residents in 1930s and 1950s council estates, typically in London, now mostly owner occupiers
		H47	Social housing, typically in 'new towns', with good job opportunities for the poorly qualified
I (Twilight subsistence)	Older people living in social housing with high care needs	I48	Older people living in small council and housing association flats
		I49	Low income older couples renting low rise social housing in industrial regions
		I50	Older people receiving care in homes or sheltered accommodation
J (Grey perspectives)	Independent older people with relatively active lifestyles	J51	Very elderly people, many financially secure, living in privately owned retirement flats
		J52	Better off older people, singles and childless couples in developments of private flats
		J53	Financially secure and physically active older people, many retired to semi rural locations
		J54	Older couples, independent but on limited incomes, living in bungalows by the sea
		J55	Older people preferring to live in familiar surroundings in small market towns
		J56	Neighbourhoods with retired people and transient singles working in the holiday industry
K (Rural isolation)	People living in rural areas far from urbanisation	K57	Communities of retired people and second homers in areas of high environmental quality
		K58	Well off commuters and retired people living in attractive country villages
		K59	Country people living in still agriculturally active villages, mostly in lowland locations
		K60	Smallholders and self employed farmers, living beyond the reach of urban commuters
		K61	Low income farmers struggling on thin soils in isolated upland locations

2.2 Preparation of the dataset for this thesis

For the work in this thesis, data management and the initial cleaning of THIN database were performed by Chris JP Smith, the data manager in the Division of Epidemiology and Public Health, University of Nottingham. Following a rigorous systematic search of the Read code list, Barbara Iyen-Omofoman compiled the Read code lists for extraction of the THIN lung cancer population. This was used by Chris Smith to extract the entire THIN population of patients with a recorded code of lung cancer. Barbara Iyen-Omofoman performed subsequent data management of the lung cancer patient population and devised the eligibility criteria for the extraction of other populations that were used to develop the case-control populations analysed in this thesis. Chris Smith and Barbara Iyen-Omofoman worked together to extract the first case-control dataset (matched on age, sex and general practice) and then Barbara Iyen-Omofoman independently extracted the second case-control dataset (matched on practice alone).

Apart from the Read code lists for smoking status, Chronic Obstructive Pulmonary Disease (COPD) and depression which had previously been used and validated in other studies within the Division of Epidemiology and Public Health, Barbara Iyen-Omofoman compiled the Read code lists that were used to extract information on quantity of cigarettes smoked, lung cancer histology and the clinical symptoms and investigations in general practice. All stages of the data preparation for this thesis were supervised by Professor Richard Hubbard and Dr Laila Tata and they reviewed all Read code lists prior to data extraction.

A preliminary set of data analyses were conducted in the first nine months of this PhD using data from the version of THIN that was released in October 2008. Following a review of the methodology, the release of an updated version of THIN database in July 2009 and in-depth discussion with my PhD supervisors, it was decided that amendments should be made to the methodology initially used

and that I should repeat all the analyses using slightly different methods and the more recent version of the database. All the analyses described in this thesis on the lung cancer population as well as analyses using the two case-control datasets are the results of analyses done with the July 2009 version of THIN. The last set of analyses which were done to validate the lung cancer predictive score were done using data from the most recent version of THIN which was released in September 2010.

The following sections describes the steps that were taken to prepare the datasets for all the analytical work that was undertaken in this thesis.

2.2.1 Definition of incident lung cancer cases

To study the interaction between GPs and patients in the period before lung cancer is diagnosed, it was decided that the lung cancer cases included in the study should have their first date of lung cancer diagnosis within the study period. This would also enable a measure of the true survival of lung cancer and avoid any survival bias that may arise with prevalent cases. To ensure that only incident cases of lung cancer were included in the study, the patients had to have been actively registered in the general practice for at least one year prior to the first diagnosis of lung cancer.

An arbitrary study start date of January 1st 2000 was assigned for the study. The last date of data collection in the version of THIN that was used for this study was July 28th 2009. The study was therefore carried out on patients with a first diagnosis of lung cancer between the 1st of January 2000 and the last date of data collection - 28th of July 2009. Certain inclusion and exclusion criteria were applied in deriving the population that were included in the analyses but these will be described in a subsequent section after defining some key dates in

the dataset as well as amendments that were made to records that had incorrect dates.

2.2.2 Key dates in THIN and the derivation of study specific dates

Most of the work in this thesis entails the follow-up of patients from a start date to a defined end date. In THIN database, dates are provided by EPIC to indicate the date of registration of patients in their respective GP practices, the date that the various GP practices started contributing data to THIN, the dates when the general practices were deemed to have mortality records that were comparable with national records, the date of patients' death, the date of transfer of patients' from their practice if applicable as well as the date of last data collection from the practices by EPIC. Some of the dates provided by EPIC had to be combined in order to create new dates that map out the beginning and end of the periods when good quality follow-up data could be confidently ascertained from the patients. The dates provided by EPIC and the new dates created by combining the EPIC dates are detailed below.

2.2.2.1 EPIC dates

In THIN database, some of the key dates provided by EPIC are :

- Patient registration date (regdate)[199] - Date of patient registration with the general practice
- Vision date[199] - Date that the general practice joined the THIN scheme and started using the vision software to record consultations
- AMR date[199] - Date when the practice is deemed to be recording all-cause mortality based on predicted numbers from the national statistics given

the practice age/sex register. Data collected after this date are considered to be of high quality for research.

Other dates provided by EPIC in the THIN dataset are:

- The date of death of the patient
- The date of transfer of patients' from the practice (if applicable)
- Date of last data collection from the practice by EPIC

2.2.2.2 Dates derived from the combination of EPIC dates

To define clear inclusion and exclusion criteria for cases in the study, dates indicating the "start" and "finish" dates for each patient in the dataset had to be assigned. These were derived from a combination of some of the dates provided by EPIC. The following dates were created by combining original dates from EPIC:

- Practice start date[199] - This is used as a measure of when the practice started recording good quality data. It is the earlier of AMR (acceptable mortality reporting) date or vision date.
- Start date (S) - defined as the later of a patient's registration date at the practice or the practice start date
- Finish date (F) - This is the date of last data collection for a patient. It is the earlier of the "transfer-out" date, death date or date of last data collection for the practice

2.2.3 Amendments made to records with incorrect dates

Despite defining start and finish dates during which a patients' follow-up could be assessed, there remained some inconsistencies within the dataset and these had to be resolved before analyses for this study could be carried out. These inconsistencies include instances where the date of lung cancer diagnoses was later than the date of death or the patients' finish date or where the recorded date of death was later than the finish date.

In making amendments to these records with incorrect dates, consideration was taken of the fact that lung cancer has very poor survival and it is thus likely that some cases in the dataset may have been diagnosed post-mortem. Also, logistical issues with record keeping may result in some time lag before the entry of data into patients' electronic notes and this was also an issue that was considered. To avoid dropping data unnecessarily therefore, gaps that were deemed reasonable had to be determined based on an examination of the distribution of the incorrect time intervals. The determination of the reasonable gaps and the subsequent amendments to the data were done consecutively in the order shown in the following sections 2.2.3.1, 2.2.3.2 and 2.2.3.3.

2.2.3.1 Lung cancer cases with diagnoses date later than date of death

There were cases in the dataset whose first record of lung cancer diagnosis was after the recorded date of death. A distribution of the death-to-diagnosis interval was plotted for these cases and is shown in Figure 2.1.

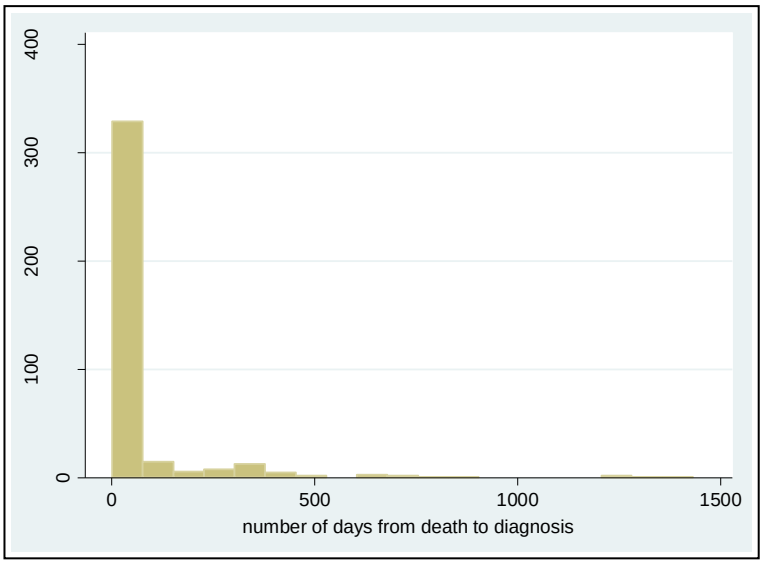


Figure 2.1:Histogram showing distribution of the interval between diagnosis and death in cases diagnosed after death (n=378)

Median interval =7 days after death (IQR 3 to 31 days)

The median interval between death and the subsequent record of lung cancer diagnosis in patients where diagnosis was recorded after death, was 7 days. The inter-quartile range was 3 to 31 days showing that 75% of them had their diagnosis made within the 31 days after death. Based on this, 31 days after death was considered a reasonable cut-off within which to accept records of lung cancer diagnosis. Patients whose lung cancer diagnoses were made within 31 days of death were considered to be most likely post-mortem diagnoses and they were retained in the study with their dates of death taken as the date of diagnoses (Figure 2.2). Records of lung cancer incidence made more than 31 days after death were considered to be a data entry error and these patients were excluded from further study analyses.

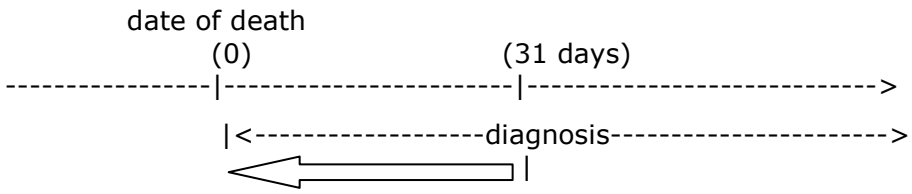


Figure 2.2: Re-coding of diagnosis date in cases diagnosed after death. These cases had their diagnosis date re-coded as the date of death.

2.2.3.2 Lung cancer cases with diagnosis date later than the finish date

By excluding or adjusting all cases with diagnosis date later than the recorded date of death in section 2.2.3.1 above, all dead cases now had diagnoses dates that were either on or before the date of death. There however remained a few patients in the dataset who had lung cancer diagnoses dates that were later than the recorded finish date. A distribution of the interval between the finish date and the diagnosis date were plotted in these cases and is shown below in Figure 2.3.

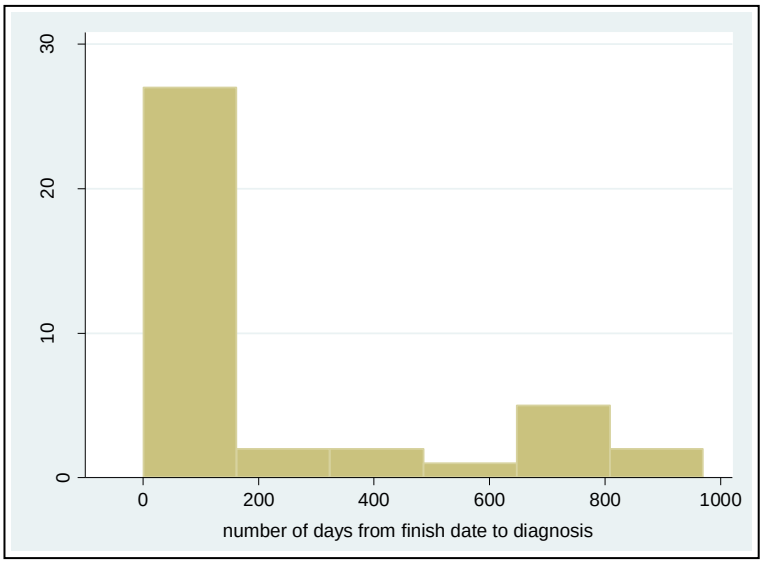


Figure 2.3: Histogram showing distribution of the interval between finish date and diagnosis in cases diagnosed after finish date (f) (n= 39)

Since all dead cases already had dates of diagnosis before or at death, this meant that the dead cases who had dates of diagnosis after their finish date, had finish dates that were earlier than their date of death. In this instance, the finish dates were re-coded to the date of death (Figure 2.4).

For the live cases whose had records of diagnoses after their recorded finish dates, a decision was made to use the same 31 day cut-off that was used in the exclusion of cases diagnosed after death. Cases who were diagnosed more than 31 days after their finish date were dropped from the dataset whereas those who

were diagnosed within 31 days of the study finish date had their finish dates re-coded to the diagnosis date (Figure 2.4).

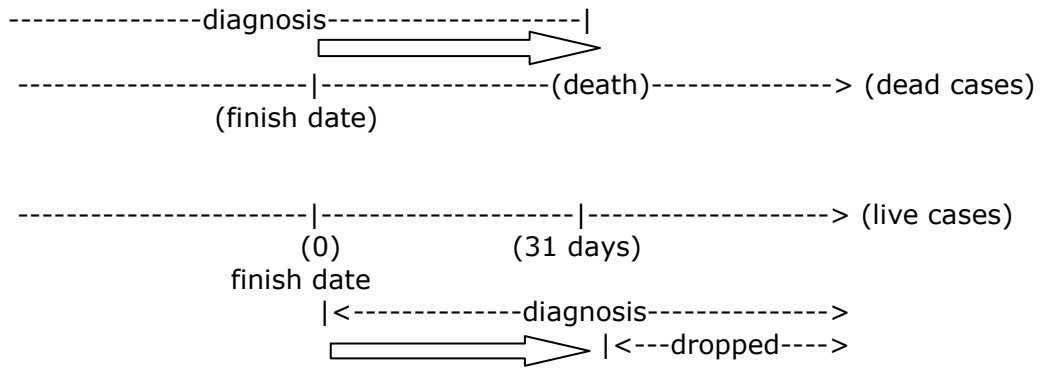


Figure 2.4: Re-coding of finish date in cases with diagnosis date after finish date

2.2.2.3 Lung cancer cases with deaths after the study finish date

Some cases had their recorded date of death later than the study finish date. Figure 2.5 shows the distribution of the interval between the study finish date and the date of death in these cases.

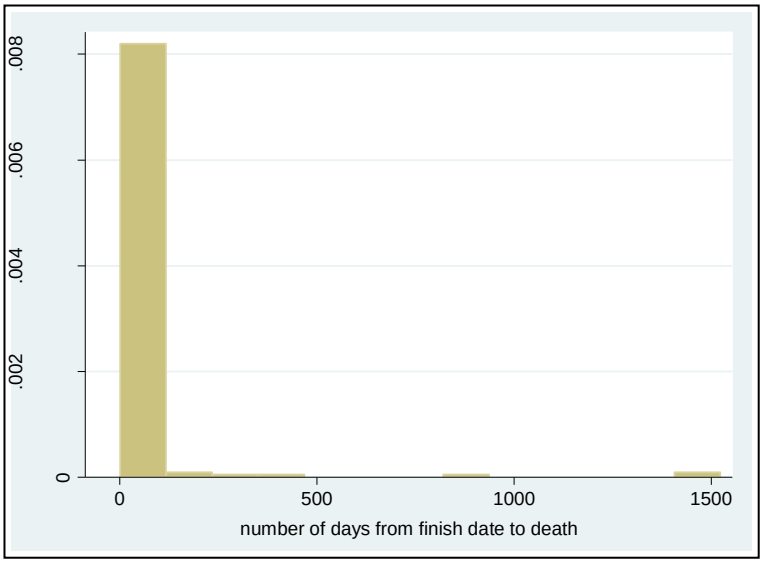


Figure 2.5: Histogram showing the distribution of the interval between f and death in cases where death was recorded after f (n=175)

Using the same 31-day cut-off previously used in excluding cases diagnosed after death and finish dates respectively, all remaining cases who had their deaths recorded more than 31 days after their finish dates were dropped from the dataset. Cases whose deaths were recorded within the 31 day period following the study finish date, had their finish dates re-coded as the date of death (Figure 2.6).

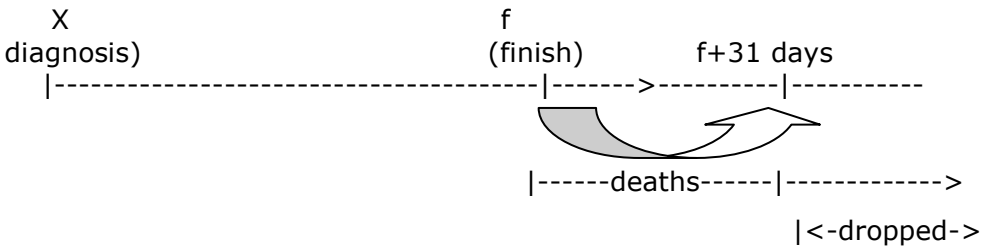


Figure 2.6: Re-coding of finish date in cases with date of death after finish date

2.3 Eligibility criteria for lung cancer cases in this study

After defining key dates in the dataset, the following inclusion and exclusion criteria were applied to derive the population of lung cancer cases that were studied in this thesis. The lung cancer Read code list that was used in extracting the population of lung cancer patients is in Appendix I.

2.3.1 Criteria for inclusion of patients in study

- First coded diagnosis of lung cancer between the 1st of January 2000 and the 28th of July 2009
- Actively registered in the GP practice for at least 1 year before diagnosis

2.3.2 Exclusion criteria

- Patients with missing month of diagnosis (If month was recorded but day was missing, day was re-coded as the first day of the given month)
- Cases with less than 1 year (365.25 days) between their start date and diagnosis date of lung cancer
- Cases with date of lung cancer diagnoses more than 31 days after death (cases diagnosed within 31 days after death had the date of diagnosis re-coded as the date of death: section 2.2.3.1).
- Cases with date of diagnoses more than 31 days after their finish date (cases diagnosed within 31 days after the finish date, had their finish date moved forward and re-coded as the date of diagnosis: section 2.2.3.2).

- Cases with date of death more than 31 days after the finish date (If death was within 31 days after finish date, the finish date was re-coded as the date of death: section 2.2.3.3).

Derivation of lung cancer cases for this study was based on the inclusion and exclusion criteria listed above. Figure 2.7 below shows the number of cases that were excluded from the study based on the criteria and shows how the final numbers in the study were obtained.

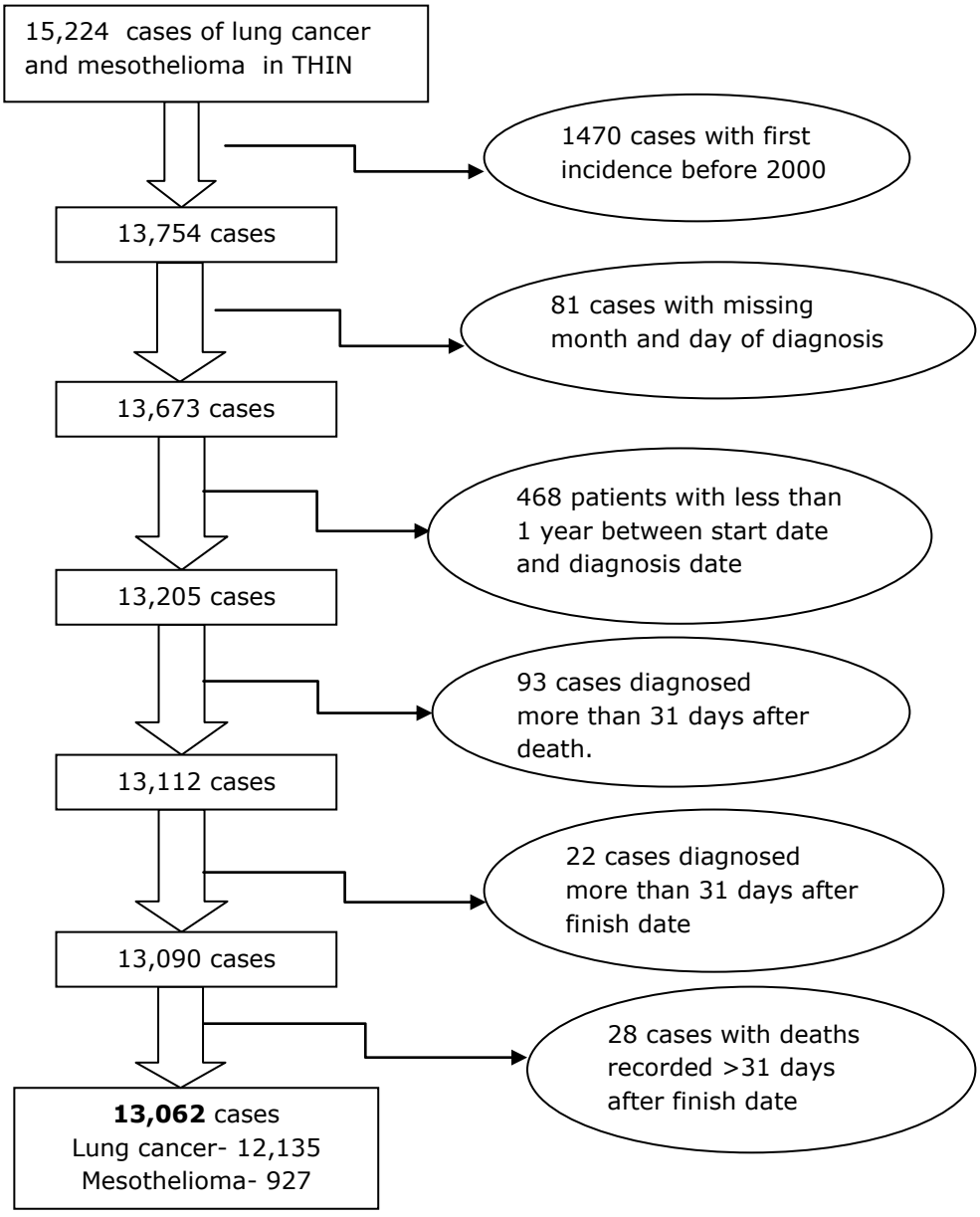


Figure 2.7: Flow chart showing how the population of lung cancer cases were derived from THIN dataset.

2.4 What proportion of lung cancer information in THIN is recorded as free text?

This section gives a general overview of free text data in THIN as well as a description of the free-text records in the medical dataset of the lung cancer cases that were included in this study. Free-text data were explored in these patients in order to determine how much of the information from variables such as histology, performance status and lung cancer staging were recorded by GPs as text and how much of these could be extracted for the analyses in this study.

2.4.1 Description of THIN free text

In THIN database, general practitioners are allowed to enter records as data comments or scanned information[223] and every entry in the medical dataset can have a data comment associated with it. These data comments are known as free text. Since this field may contain confidential information such as people's names, places, etc, not all of this information is made available for researchers. Fifty seven percent (168,037 comments) of free text in the medical records of all patients has been anonymised including the 10,000 most frequently used free text[223], and these anonymised comments have been linked to a 7 character unique identifier which can be looked up in an ancillary file called THINComments.

There are different types of free text data that can be obtained from patients' medical records in THIN and these are described below:

- The 7 character numeric identifier (anonymised text) which has been linked to a unique comment and can be looked up in the ancillary file called THINComments.

- 4-character alphanumeric textids which represent free text that are not one of the 168,037 anonymized comments (non-anonymised text). To ensure confidentiality of these texts, access can only be provided by EPIC on request and involves extrapolation of records by scrutiny of individual comment fields. Provision of access to these free text comments can therefore be a time-consuming and arduous task for the staff at EPIC and this access is quite expensive for researchers.
- A 7-character numeric identifier coded as "0000001" which is an empty text and does not code for anything (no text).

2.4.2 Free text in lung cancer patients' records

In the medical dataset of the 12,135 lung cancer cases in this study, a total of 1,896,389 free text records were identified. Of these, only 152,075 (8%) were anonymised texts which could be looked-up in the ancillary file and could potentially be retrieved if required. There were 1,184,309 (62.5%) non-anonymised texts with restricted and expensive access and the remaining 560,005 (29.5%) free texts were empty texts that did not code for anything. Figure 2.8 shows the proportion of the different types of free text comments in the medical dataset of patients with lung cancer in the thesis database.

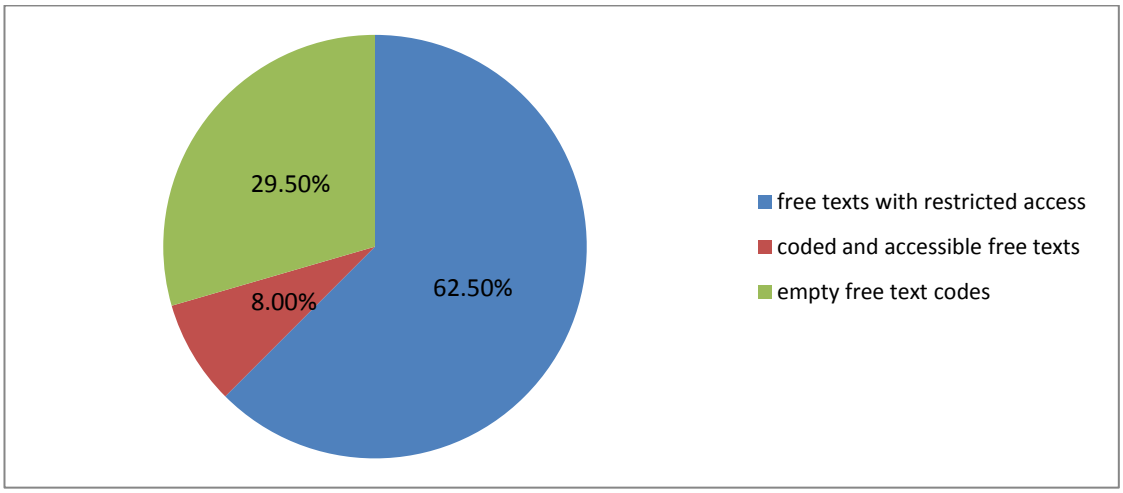


Figure 2.8: Types of free text in the medical dataset of patients with lung cancer

2.4.2.1 Free text data entries recorded with lung cancer Read codes

A total of 17,449 lung cancer Read code entries were identified in the medical dataset of cases in the thesis database. The free text records that were associated with these lung cancer entries were explored and the results showed that only 677 (3.9%) of these lung cancer-associated free text comments were coded texts that could be looked up in the ancillary file. There were 11,817 (67.7%) uncoded text comments that had not been anonymised and 4,955 (28.4%) were free text comments that were empty and did not code anything. Figure 2.9 shows the proportions of the different types of free text records that were associated with lung cancer entries.

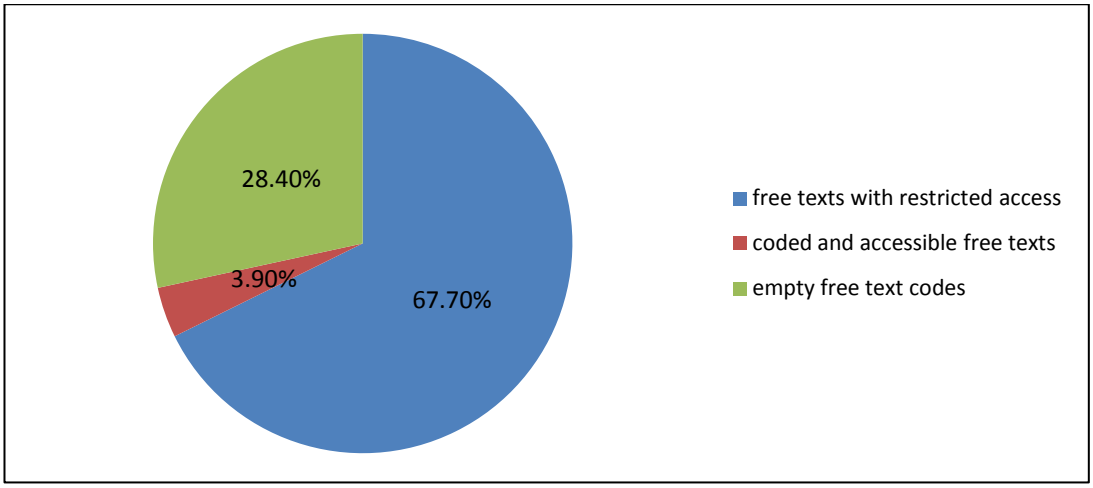


Figure 2.9: Free texts associated with lung cancer Read code entries

Among the 677 anonymised free text comments (3.9% of lung cancer Read code-associated free texts) that could be looked up in the ancillary file, the 20 most common free text entries were examined (Table 2.4) and apart from 37 entries of the histological sub-type "Adenocarcinoma", there was no other information from the free text entries that was considered relevant to this study.

Table 2.4. Most common free text comments associated with lung cancer Read code entries

Free text entry associated with lung cancer Read code entry	Total count
Right	56
Left	45
Inoperable	44
Cause of death	41
Metastatic	39
Adenocarcinoma	37
Lung	24
1A	23
Rt	17
1B	15
Primary	14
Probable	13
Left lower lobe	11
Right lower lobe	9
R	8
Confirmed	6
Radiotherapy	5
Right side	4
Lt	3
Recurrent	2

2.4.2.2 Non-anonymised free text in the lung cancer patients' dataset

As mentioned above in the introductory part of section 2.4.2, a total of 1,184,309 non-anonymised free text comments were identified in the medical dataset of the cases (62.5% of all free text records). These text comments were associated with medical entries entered using 18,432 Read codes from the July 2009 ('0907) EPIC Read code list. Although the free text comments were not accessed, an analysis was done to obtain the frequency of different Read codes associated with these comments. The median Read code frequency with non-anonymised free texts was 3 (IQR 1 to 14). Therefore, 75% of the Read code entries had 14 or less records associated with the free texts. There were 206 Read codes that were recorded in more than 1000 entries associated with non-anonymised free texts. The 38 most frequently recorded Read codes which had more than 4000 entries associated with non-anonymised texts are shown in Table 2.5.

Table 2.5. Most common Read codes associated with non-anonymised free texts in the case dataset

Read code associated with non-anonymised free text	Total count
Telephone encounter	59131
Letter from specialist	33723
Patient reviewed	29727
Had a chat to patient	20868
Home visit	16265
Cancer care review	11801
Medication requested	11041
Incoming mail NOS	10878
Administration NOS	9902
Seen in GP's surgery	9837
Dressing of wound	9237
Cough	8821
_Converted code	8087
Chest pain	7630
Seen in oncology clinic	7618
Administration	7504
Discharge summary	7451
Blood sample -> Lab NOS	7043
Communication from:	6912
Chest infection	6564
MED3 - doctor's statement	6542
C/O - cough	6524
Letter encounter from patient	6397
Incoming mail	5981
Seen in oncology clinic	5576
Discharged from hospital	5420
Third party encounter	5267
Medication review	5154
Chest infection NOS	5020
Discussion	4565
Patient's condition improved	4546
Comment note	4427
Letter encounter	4422
Patient given advice	4371
Lung cancer	4364
Seen in hospital casualty	4321
Hypertension monitoring	4243
Nursing care blood sample taken	4149

Investigation of the most common Read codes associated with the non-anonymised free text showed that the majority of these codes were Read codes for encounters with patients via telephone, chat or home visit. Further analysis to identify the most common Read code categories associated with non-anonymised free text was done (Table 2.6) and the results further confirmed that most of the non-anonymised free text entries were associated with Read codes of patient encounter by telephone, letter, chat or mail.

Table 2.6. Most common Read code categories associated with non-anonymised free texts in the medical dataset of cases

Read code category associated with non-anonymised free text	
9N	Patient encounter admin. Data
8H	Referral for further care
66	Chronic disease monitoring
R0	Symptoms
ZL	Administrative statuses
13	Social/personal history
F4	Disorders of eye and adnexa
7N	Subsidiary classification of laterality and operation sites
8B	Other therapy
9O	Prevention/screening admin.

2.4.3 Summary of findings from analyses of free text

Results from the preceding sections have shown that less than 10% (8%) of free text comments in the dataset of patients with lung cancer were easily accessible. Although 3.9% of the free text comments associated with entries of lung cancer Read codes were accessible, these did not provide much information of relevance to this study. Furthermore, examination of the Read code categories that were associated with the non-anonymised free text comments showed that the majority of these texts were associated with Read codes of encounters with patients by telephone, letter, chat, mail and may not provide very useful information in terms of signs and symptoms presented by the patients.

Following these findings, it was not considered a worthwhile exercise to request the manual extrapolation of non-anonymised free texts from EPIC for the work in this thesis.

2.5 Statistical software for data analyses

All the analyses undertaken in this study were performed using Stata release SE version 11 (StataCorp LP, Texas, USA). The statistical methods of analyses are described in more detail within each section.

2.6 Study ethics

Ethical approval for this study was granted by the Cegedim Strategic Data Medical Research scientific review committee in 2009. All records of patients in THIN are anonymised and do not contain any identifying information such as name, address, exact date of birth and NHS number[199].

2.7 Funding

This PhD research was funded by a grant from the Economic and Social Research Council (ESRC). I also wish to acknowledge funding by the British Lung Foundation (BLF) and the Nottingham Respiratory Biomedical Research Unit. The principal supervisor for this PhD research - Professor Richard Hubbard, is the BLF chair in Epidemiological Respiratory research.

Chapter 3. Validation of THIN and the distribution of lung cancer across sectors of society in the United Kingdom

This chapter describes a study which firstly, assessed the completeness and representativeness of THIN database to ensure that it was a valid source of data for lung cancer research. Then using Experian's Mosaic public sector ^{TM} variable which had been linked into THIN, the study identified detailed profiles of the UK sectors of society where lung cancer incidence is highest, as a means of exploring the potential of using this geo-demographic social marketing tool to facilitate lung cancer ascertainment. A brief justification for the study in this chapter is stated in the introduction and this is followed by the study methods, results, a discussion of the study findings with regards to what is already known and then a conclusion with a statement of what the study adds to current evidence.

3.1 Introduction

There exist socioeconomic variations in the incidence of lung cancer[90 96] and evidence from studies of other cancer screening services and treatments show unequal participation among different population sub-groups in screening services[224] as well as inequity in cancer treatment[225]. To increase earlier ascertainment of lung cancer and reduce lung cancer-related health inequalities, there is a public health need to enhance lung cancer awareness especially in sectors of society where lung cancer incidence is typically high, with a view to shortening the interval between symptoms and presentation to primary care. Computerised general practice records from THIN present a potentially useful

source of data to understand the current pathway of lung cancer diagnosis in general practice as well as identify the societal distribution of lung cancer.

There are two area-based measures of socioeconomic status in THIN - The Townsend quintile of deprivation and the Mosaic public sector™ classification, and these have been described in Chapter 2, "Description of the dataset and derivation of the lung cancer population". Compared with the well-known and commonly used Townsend Index[221] which measures the area-based level of material deprivation using four indicators: unemployment, car ownership, house ownership and overcrowding, Mosaic Public Sector™ classifications take account of more granular characteristics of the population living at different UK postcodes and therefore allows a clearer identification of the characteristics and differing needs of people[226]. To date, Mosaic classification has been used to a limited extent for the targeting of population public health services to those most in need[227] and studies have usefully applied it to demonstrate social disparities in health-related behaviours such as heavy episodic drinking[228] and smoking prevalence[102]. However no study yet has used Mosaic classifications to identify particular sectors of the UK society that may benefit from targeted public health efforts to improve lung cancer awareness and care.

Although THIN has been demonstrated to have a high degree of completeness and accuracy for records of several disease diagnoses[206-208] and cancers[209], it has not been fully exploited for lung cancer studies and its usefulness for this study and other lung cancer research will depend on its level of ascertainment and representativeness of lung cancer in the UK.

3.2 Methods

3.2.1 Derivation of variables analysed

3.2.1.1. Study population

All patients with a first recorded diagnosis of lung cancer between the 1st of January 2000 and the 28th of July 2009 were identified. The process used to derive the 12,135 incident cases of lung cancer used in this study, has been previously described in Chapter 2 (section 2.3).

3.2.1.2 Records of Lung cancer histology, Chronic Obstructive Pulmonary Disease (COPD) and smoking.

Records of lung cancer histology, Chronic Obstructive Pulmonary Disease (COPD) and smoking were obtained from patients in the study using Read code lists that were compiled after a thorough systematic search of the Read code dictionary (Read codes listed in Appendix I).

The Read codes for histology were developed based on the recommendations from the 2001 World Health Organisation classification of lung tumours[229]. The list of Read codes for smoking had been developed and used for other research in the Division of Epidemiology and Public Health. All records of smoking status before lung cancer diagnosis were extracted for each patient and based on their most recent smoking status before diagnosis, patients were classified as current smokers, ex smokers or non-smokers. Non-smokers who had previous records of being current or ex smokers, were re-classified as ex-smokers.

3.2.2 Characteristics of the lung cancer patients in THIN

To first address the need for validation, the completeness and representativeness of lung cancer data in THIN of the national UK population of patients with lung cancer were assessed. In doing this, the characteristics of patients with lung cancer in THIN as well as the incidence and survival rates of lung cancer in THIN between 2000 and 2009 were determined and these were compared with two reliable UK national lung cancer databases - The UK National Cancer Registry[230] and the National Lung Cancer Audit Database (LUCADA)[231].

Using basic descriptive statistics in STATA, the characteristics of the lung cancer patients were determined. Lung cancer patient characteristics such as histological types, Chronic Obstructive Pulmonary Disease (COPD) prevalence prior to lung cancer diagnosis and smoking status were also determined.

For the calculation of lung cancer incidence rate, the base population for analysis comprised of the entire population of patients registered in THIN general practices, who had contributed data after the 1st of January 2000 and who had records for at least one year in the dataset. Incidence rates with 95% confidence intervals (CI) were calculated as the total number of new lung cancer cases per 100,000 person-years at risk. Overall incidence rates in the population were calculated for the study period (2000-2009) and the results were stratified by calendar years (3-year periods), age (10-year age bands up to ≥90 years), sex, socioeconomic status and Strategic Health Authority (SHA) regions. Socioeconomic status was measured using the Townsend Index of multiple deprivation in quintiles[219] and the Mosaic Public Sector™ groups and types[222]. To assess the completeness of lung cancer ascertainment in THIN general practices and whether this varied by different UK SHA regions, the lung cancer incidence rates in THIN for each SHA were compared with the rates recorded by the National Cancer Registry[230]. Incidence rate ratios (IRR) between different population strata were obtained using multivariate Poisson regression. The

incidence rate ratios were further analysed using separate random effects Poisson regression models to adjust for any effects due to the variable reporting in different UK general practices[232].

Lung cancer survival rates were calculated from the period of first recorded lung cancer diagnosis to death or the date of last data collection from the general practice. Survival rates of lung cancer in THIN were compared with rates in the National Lung Cancer Audit database (LUCADA)[231], which is a good source of highly representative information on diagnosis and survival of lung cancer patients in NHS trusts throughout England, Wales and Scotland. Cox proportional hazards models were used to model survival data with age, sex and socioeconomic status to determine the relationship between these factors and lung cancer survival. The Cox proportional hazards assumption was assessed for each of the models by plotting the log minus log transformation of the Kaplan-Meier estimator of the survival function against time.

3.2.3 UK societal distribution of lung cancer

To identify the variation in lung cancer incidence across different UK sectors of society, the incidence rates of lung cancer in the different Mosaic Public Sector™ groups and types were determined. Because age and sex are used in part to derive the Mosaic Public Sector™ classification, models for the Mosaic analysis did not adjust for these covariates. Using the calculated lung cancer incidence rates in the different Mosaic types and the population make-up by Mosaic type in the different UK Primary care Trusts (PCT), the estimated number of lung cancer events in each PCT as well as the estimated incidence rates per 100,000 person years were estimated (this was jointly carried out with Experian UK).

3.3 Results

3.3.1 Characteristics of the lung cancer patients in THIN

3.3.1.1 General patient characteristics

Of the total number of 12,135 incident cases of lung cancer recorded in THIN between the 1st of January 2000 and the 28th of July 2009, there were 7,184 males (59.2%) and 4,951 females (40.8%). The median age at lung cancer diagnosis was 72.6 years (Inter-quartile range (IQR): 64.5-79.0). The median age at death was 73.8 years (IQR: 65.7-80.0).

3.3.1.1 Description of lung cancer types in THIN

The distribution of the different types of lung cancer description among patients in THIN is shown in Table 3.1. The most commonly recorded lung cancer description in patients' records were:

- Lung cancer
- Malignant neoplasm of the bronchus or lung
- Malignant neoplasm of trachea, bronchus and lung
- Malignant neoplasm of main bronchus
- Malignant neoplasm of upper lobe of lung

Table 3.1 Description of lung cancer types among patients in THIN database

Description	Frequency	Percent
Lung cancer	4,204	34.64
Malignant neoplasm of bronchus or lung	3,906	32.19
Malignant neoplasm of carina of bronchus	35	0.29
Malignant neoplasm of chest wall NOS	13	0.11
Malignant neoplasm of hilus of lung	53	0.44
Malignant neoplasm of lower lobe bronchus	60	0.49
Malignant neoplasm of lower lobe of lung	204	1.68
Malignant neoplasm of lower lobe, bronchus or lung	110	0.91
Malignant neoplasm of lower lobe, bronchus or lung NOS	30	0.25
Malignant neoplasm of main bronchus	452	3.72
Malignant neoplasm of main bronchus NOS	91	0.75
Malignant neoplasm of middle lobe bronchus	14	0.12
Malignant neoplasm of middle lobe of lung	50	0.41
Malignant neoplasm of middle lobe, bronchus or lung	35	0.29
Malignant neoplasm of middle lobe, bronchus or lung NOS	6	0.05
Malignant neoplasm of other sites of bronchus or lung	74	0.61
Malignant neoplasm of overlapping lesion of bronchus and lung	15	0.12
Malignant neoplasm of respiratory tract	28	0.23
Malignant neoplasm of thorax NOS	1	0.01
Malignant neoplasm of trachea	19	0.16
Malignant neoplasm of trachea NOS	4	0.03
Malignant neoplasm of trachea, bronchus and lung	1,763	14.53
Malignant neoplasm of upper lobe bronchus	116	0.96
Malignant neoplasm of upper lobe of lung	396	3.26
Malignant neoplasm of upper lobe, bronchus or lung	287	2.37
Malignant neoplasm of upper lobe, bronchus or lung NOS	43	0.35
Malignant neoplasm, overlap lesion of resp and intrathor organs	1	0.01
Pancoast's syndrome	31	0.26
[X]Malignant neoplasm of bronchus or lung unspecified	30	0.25
[X]Malignant neoplasm of respiratory and intrathoracic organs	62	0.51
[X]Malignant neoplasm/ill-defined sites	2	0.02
Total	12,135	100

3.3.1.3 Histological subtypes

Lung cancer histology records were available in the records of only 1,704 out of the 12,135 patients with lung cancer (14% of cases). This consisted of 1,659 records extrapolated from the patients' Medical and AHD datasets and 45 records retrieved from the medical free text comments. Small cell lung cancer was the histological type in 384 patients (22.5% of cases with histology), squamous cell carcinoma was the type in 689 patients (40.4%), adenocarcinoma was the histological type in 610 patients (35.8%) and 21 patients (1.2%) had large cell carcinoma.

3.3.1.4 Prevalence of Chronic Obstructive Pulmonary Disease (COPD)

Chronic Obstructive Pulmonary Disease (COPD) records were obtained from 3,082 patients with lung cancer (25.4% of cases with lung cancer). Analyses of the interval between the first diagnosis of COPD and the date of lung cancer diagnosis showed that the median time of COPD diagnosis was 3.9 years prior to lung cancer diagnosis (IQR 11 months to 8.4 years prior to lung cancer diagnosis).

3.3.1.5 Smoking status

Information on smoking was available in the records of 11,718 patients with lung cancer (96.6% of the population of lung cancer cases). Prior to the diagnosis of lung cancer, 5,537 patients (45.6%) were current smokers, 4,848 patients (40.0%) were ex-smokers, 1,230 patients (10.1%) had never smoked and 520 (4.3%) had no record of smoking status in their dataset. In total, 85.6% of patients with lung cancer had a history of ever-smoking before diagnosis (Figure 3.1).

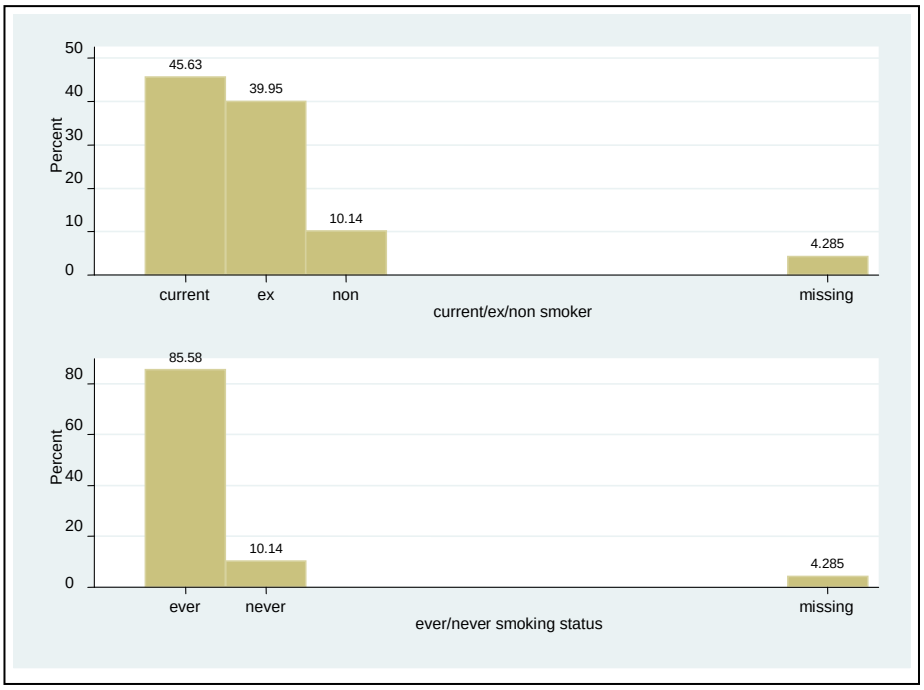


Figure 3.1: Last recorded smoking status of lung cancer patients prior to diagnosis
** "missing" includes lung cancer patients with no recorded smoking data as well as patients who either had their smoking records taken after disease diagnosis.

3.3.2 Lung cancer incidence in THIN

3.3.2.1 Overall incidence

The overall incidence of lung cancer in THIN for the whole study period from 2000 to 2009 was 41.4 per 100,000 person-years (95% CI 40.6-42.1). Lung cancer incidence increased by approximately 4% for every 3-year period (IRR 1.04, 95% CI 1.04-1.05) (Figure 3.2). The incidence rate in the 3-year period 2000-2002 was 33.1 per 100,000 person years (95% CI 31.9-34.3). The incidence rate in 2003-2005 was 42.8 per 100,000 person years (95% CI 41.5-44.2), incidence in 2006-2008 was 46.8 per 100,000 person years (95% CI 45.4-48.2) and the incidence rate in 2009 was 45.1 per 100,000 person years (95% CI 42.0-48.4).

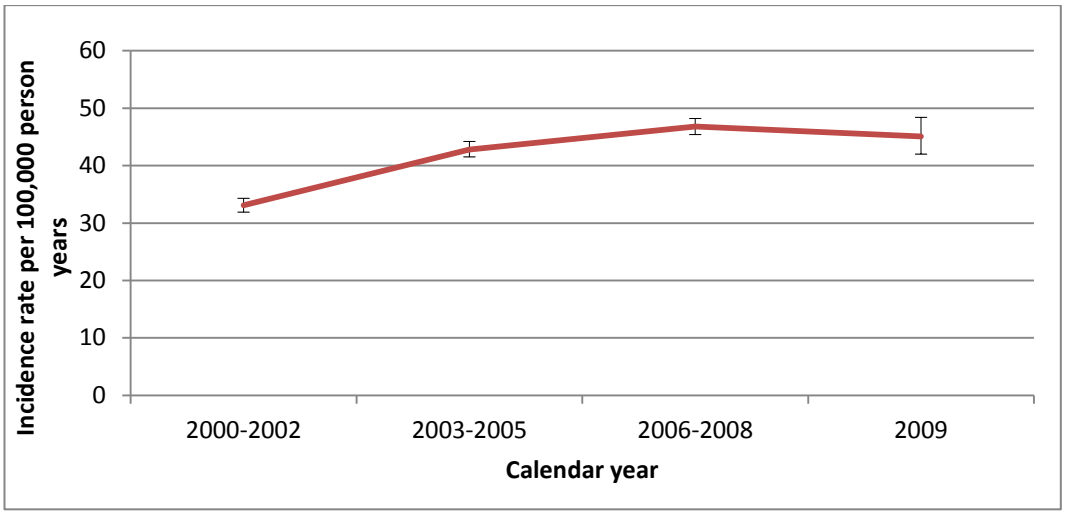


Figure 3.2: Trend in incidence of lung cancer, 2000-2009
Bars represent 95% confidence intervals

3.3.2.2 Lung cancer incidence by age-groups and sex

Incidence rates were 50% higher in males (49.4 per 100,000 person-years, 95% CI 48.2-50.5) compared with females (33.5 per 100,000 person-years, 95% CI 32.6-34.4) and increased with age, reaching a peak in the 80-90 year age-group in males and in the 70-80 year age-group in females (Table 3.2 & Figure 3.3).

Table 3.2. Overall incidence rates of lung cancer by age group and sex (2000-2009)

	Lung cancer events		**100,000 Person-yrs at risk**		**Rate/ 100,000 person-years (95% CI)**		
Age group (years)							
	Male	Female	Male	Female	All	Male	Female
0-40	30	29	75.6	72.4	0.4 (0.3-0.5)	0.4 (0.3-0.6)	0.4 (0.3-0.6)
40-50	168	147	22.1	21.3	7.3 (6.5-8.1)	7.6 (6.5-8.8)	6.9 (5.9-8.1)
50-60	793	574	19.3	18.9	35.7 (33.9-37.7)	41.0 (38.3-44.0)	30.3 (28.0-33.0)
60-70	1951	1285	14.5	14.9	110.0 (110-110)	134.4 (130-140)	86.3 (81.7-91.1)
70-80	2737	1781	9.6	11.7	212.0 (210-220)	286.2 (280-300)	151.6 (140-160)
80-90	1365	1029	3.9	7.0	219.4 (210-230)	348.1 (330-370)	147.2 (140-160)
> 90	139	105	0.5	1.5	120.2 (110-140)	283.7 (240-340)	68.2 (56.3-82.5)
All ages	7184	4951	145.5	147.8	41.4 (40.6-42.1)	49.4 (48.2-50.5)	33.5 (32.6-34.4)

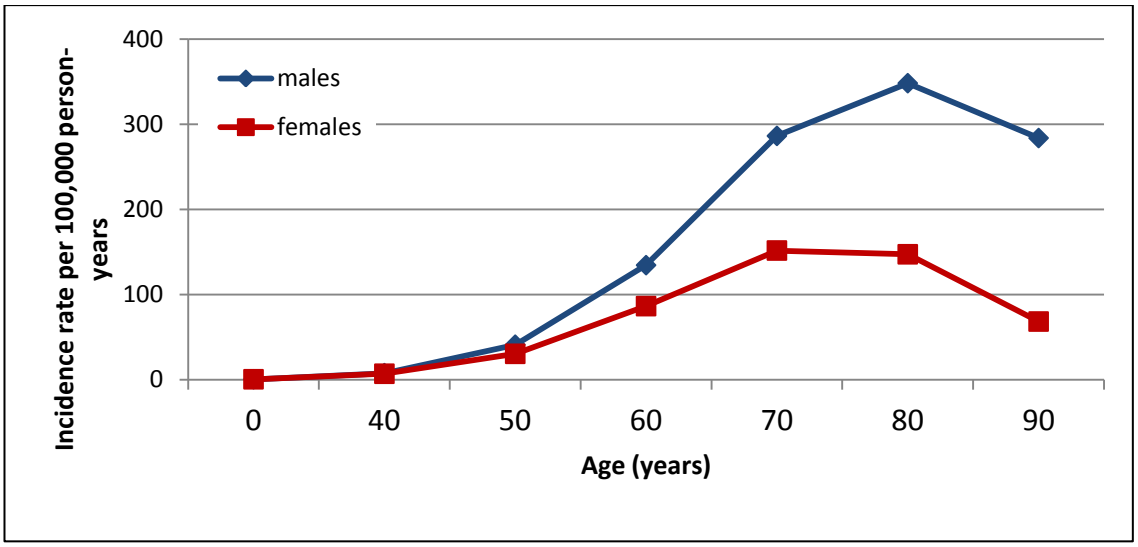


Figure 3.3: THIN lung cancer incidence rates by age and sex

3.3.2.3 Lung cancer incidence rates by Strategic Health Authority (SHA)

The SHAs with the highest lung cancer incidence rates in THIN were the North-West of England with 58.6 per 100,000 person-years (95% CI 55.9 - 61.5) followed by the North-East of England with 57.1 per 100,000 person-years (95% CI 52.6 - 61.9) and Scotland with an incidence rate of 54.4 per 100,000 person-years (95% CI 51.4 - 57.6) (Table 3.3). The lowest incidence rates for lung cancer were in London with 31.8 per 100,000 person-years (95% CI 29.8 - 33.8) followed by the South-East Coast of England with 32.3 per 100,000 person-years (95% CI 30.2 - 34.4) and the East Midlands with an incidence rate of 35.0 per 100,000 person-years (95% CI 31.9 - 38.3). Comparing lung cancer incidence rates in THIN in the SHA regions over the 3 year period from 2006-2008 (when lung cancer incidence in THIN had increased from the initial stages of the study and reached a plateau) with the 2003-2007 lung cancer incidence rates recorded by the National Cancer Registry[230], the rates in THIN and registry were comparable in 9 of the 13 SHAs (Table 3.3). THIN incidence rates were higher than registry rates in the South-West of England but the rates were lower than registry rates in London, Northern Ireland and the West Midlands. The overall lung cancer incidence rate in THIN for all the SHAs between 2006-2008 was 46.8

per 100,000 person-years and this accounts for 93.2% of the national cancer registry incidence rate of 50.2 per 100,000 person-years.

Table 3.3. Distribution and incidence rates of THIN lung cancer cases by UK Health authority

Strategic health authority (SHA)	Overall lung cancer incidence rate in THIN / 100,000 person-years (95% CI)	Number of new cases of lung cancer in THIN 2006-2008	100,000 person years at risk	THIN 2006-2008 lung cancer incidence rates/ 100,000 person years (95% CI)	UK national cancer registry age-standardised incidence rates of lung cancer (2003-2007)/ 100,000 person yrs[230]	Crude lung cancer incidence rate ratio (THIN compared to Registry rates)
East Midlands	35.0 (31.9 - 38.3)	172	4.1	41.7 (35.9-48.4)	47.1 (46.3-47.9)	0.89
East of England	36.7 (34.3 - 39.1)	331	7.6	43.5 (39.1-48.5)	40.6 (39.9-41.2)	1.07
London*	31.8 (29.8 - 33.8)	358	9.9	36.1 (32.5-40.0)	48.7 (48.0-49.4)	0.74
North East	57.1 (52.6 - 61.9)	211	3.3	63.6 (55.5-72.7)	68.2 (66.9-69.5)	0.93
North West	58.6 (55.9 - 61.5)	605	9.3	65.1 (60.1-70.5)	59.3 (58.6-60.1)	1.10
Northern Ireland*	35.3 (32.1 - 38.9)	146	3.8	38.8 (33.0-45.6)	49.2 (47.8-50.6)	0.79
Scotland	54.4 (51.4 - 57.6)	479	7.4	64.9 (59.4-71.0)	69.2 (68.3-70.1)	0.94
South Central	36.5 (34.5 - 38.6)	459	11.3	40.5 (36.9-44.4)	39.4 (38.6-40.2)	1.03
South East Coast	32.3 (30.2 - 34.4)	337	9.3	36.1 (32.5-40.2)	39.7 (39.0-40.5)	0.91
South West**	42.4 (40.2 - 44.8)	476	10.2	46.5 (42.5-50.8)	38.9 (38.3-39.6)	1.20
Wales	44.8 (41.7 - 48.0)	319	6.1	52.6 (47.2-58.7)	52.8 (51.8-53.9)	1.00
West Midlands*	36.5 (34.3 - 38.7)	372	9.3	39.8 (36.0-44.1)	46.5 (45.8-47.2)	0.86
Yorkshire & Humber	47.4 (43.9 - 51.2)	233	4.4	52.6 (46.2-59.8)	56.9 (56.0-57.7)	0.92
Overall	41.4 (40.6 - 42.1)	4498	96.2	46.8 (45.4-48.2)	50.2 (49.9-50.5)	0.93

* SHAs with lower incidence of lung cancer recorded in THIN compared to national cancer registry
** SHAs with higher incidence of lung cancer recorded in THIN compared to national cancer registry
(There is an overlap of the 95% confidence intervals in the incidence rates in the other 9 SHAs)

3.3.2.4 Lung cancer incidence rates by deprivation

3.3.2.4.1 Lung cancer incidence rates by Townsend deprivation quintiles

There was a strong relationship between socioeconomic deprivation and lung cancer incidence (Table 3.4). Using the Townsend Index as a measure of area level deprivation, the highest lung cancer incidence rate of 61.5 per 100,000 person-years (95% CI 59.1-64.1) in the most deprived Townsend quintile was over twice the incidence rate of 28.7 per 100,000 person-years (95% CI 27.5-30.0) in the least deprived quintile. After adjusting for the effects of age, sex and general practice (Table 3.4), there was an 11% increase in lung cancer incidence for every category increase in Townsend quintile (IRR 1.11, 95% CI 1.10-1.12) and the rate of lung cancer for people in the most deprived Townsend quintile was 2.2 times higher than the rate for people in the least deprived quintile (IRR 2.2, 95% CI 2.0-2.3).

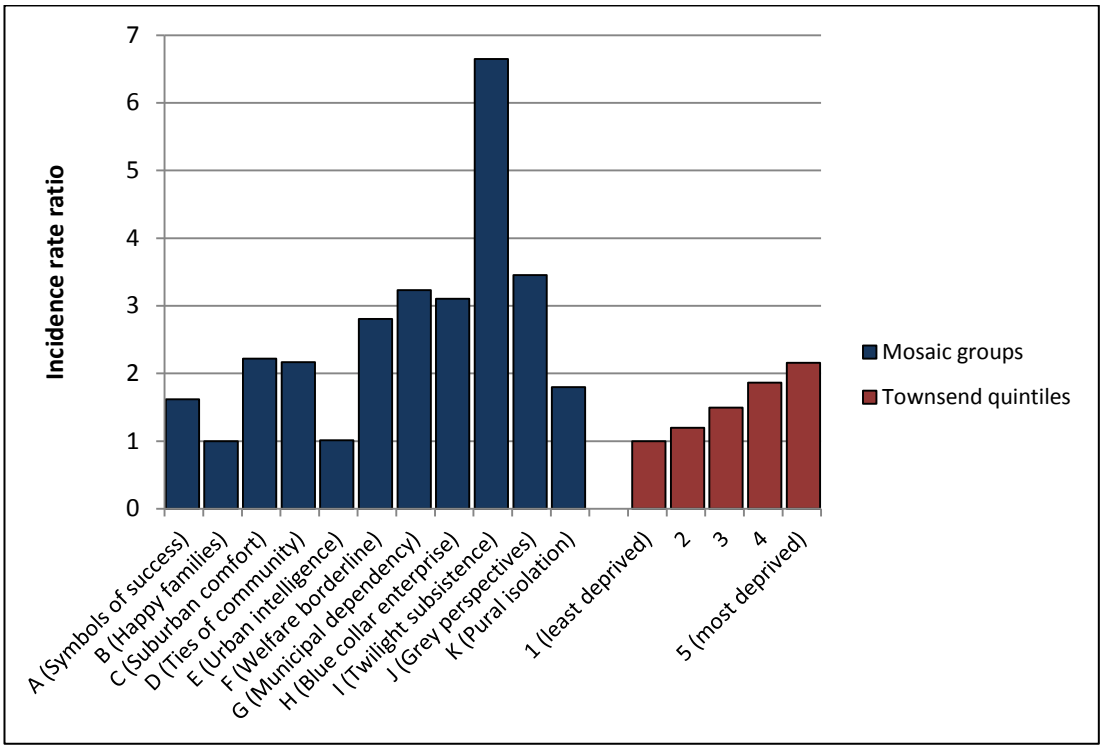
Table 3.4. Overall incidence of lung cancer by Townsend Index quintiles and Mosaic Public Sector™ groups

	Lung ca events	Person-yrs at risk	Rate per 100,000 p/y (95% CI)	Incidence rate ratios (95% CI) §
Townsend index of deprivation				
1 (least deprived)	2069	72.0	28.7 (27.5 - 30.0)	1.00
2	2243	61.4	36.5 (35.1 - 38.1)	1.20 (1.12-1.27)
3	2439	58.1	42.0 (40.4 - 43.7)	1.49 (1.41-1.59)
4	2653	51.4	51.7 (49.8 - 53.7)	1.86 (1.75-1.98)
5 (most deprived)	2245	36.4	61.5 (59.1 - 64.1)	2.16 (2.02-2.31)
missing	484	14.0	34.5 (31.5 - 37.7)	1.29 (1.14-1.46)
Mosaic Public Sector™ group				
A (Symbols of success)	690	27.9	24.7 (23.0 - 26.6)	1.62 (1.45-1.81)
B (Happy families)	613	34.9	17.6 (16.2 - 19.0)	1.00
C (Suburban comfort)	1700	46.7	36.4 (34.7 - 38.1)	2.22 (2.02-2.44)
D (Ties of community)	1608	40.0	40.2 (38.2 - 42.2)	2.17 (1.97-2.38)
E (Urban intelligence)	233	11.4	20.5 (18.0 - 23.3)	1.01 (0.86-1.19)
F (Welfare borderline)	566	9.0	62.6 (57.6 - 68.0)	2.81 (2.49-3.17)
G (Municipal dependency)	1008	15.4	65.5 (61.6 - 70.0)	3.23 (2.91-3.59)
H (Blue collar enterprise)	1791	33.4	53.7 (51.2 - 56.2)	3.10 (2.83-3.41)
I (Twilight subsistence)	866	6.7	129.3 (121.0 - 138.2)	6.65 (5.98-7.39)
J (Grey perspectives)	1239	20.4	60.7 (57.4 - 64.2)	3.45 (3.12-3.83)
K (Rural isolation)	425	14.1	30.0 (27.3 - 33.0)	1.80 (1.58-2.05)
99 (Missing)	1394	33.3	41.9 (39.8 - 44.2)	1.95 (1.73-2.20)

§ Townsend Index incidence rate ratios adjusted for age, sex and general practice
Mosaic Public Sector group incidence rate ratios adjusted for general practice

3.3.2.4.2 Lung cancer incidence rates by Mosaic Public Sector ™ groups and types

Compared with Townsend Index quintiles, there were wider variations in the incidence of lung cancer across Mosaic Public Sector™ groups (Table 3.4, Figure 3.4). The highest lung cancer incidence rate of 129.3 per 100,000 person-years (95% CI 121.0-138.2) was found in Mosaic Public Sector™ group I (Twilight subsistence). Mosaic Public Sector™ groups F, G and J also had high rates of lung cancer incidence. After adjusting for any effects due to the variable reporting of general practices, the lung cancer incidence rate in Mosaic group I where incidence was highest, was 6.6 times higher when compared with the rate in Mosaic group B where the incidence of lung cancer was lowest (IRR 6.65, 95% CI 6.0-7.4).



Reference groups (Mosaic group B ; Townsend quintile 1)

Figure 3.4: Lung cancer incidence rate ratios by Mosaic Public Sector ™ groups and by Townsend quintiles (adjusted for age, sex and practice)

Analyses of the 61 Mosaic Public Sector™ types (Table 3.5 & Figure 3.5) showed the highest lung cancer incidence rate of 191.7 per 100,000 person-years (95% CI 173.8-211.5) in Mosaic Public Sector™ type I50 (Cared for pensioners). The next highest incidence rate of 174.2 per 100,000 person-years (95% CI 151.1-200.7) was found in Mosaic Public Sector™ type I48 (Old people in flats). Lung cancer incidence was lowest for people in Mosaic Public Sector™ type B10 (Upscale new owners) with a rate of 6.2 per 100,000 person-years (95% CI 4.4-8.7). The incidence rate of lung cancer in Mosaic type I50 was 31.2 times higher (IRR 31.2, 95% CI 21.9-44.5) when compared to the rate in Mosaic type B10.

Table 3.6 summarizes the typical characteristics of the Mosaic Public Sector™ groups and types where lung cancer incidences were highest in the UK.

Table 3.5. Incidence rates (per 100,000 person years) by mosaic types

Mosaic type	Lung ca events	Person-yrs at risk	Rate per 100,000 p/y (95%CI)
A01 Global connections	10	0.38	26.0 (14.0 - 48.4)
A02 Cultural leadership	41	2.19	17.8 (13.8 - 25.5)
A03 Corporate chieftains	59	3.62	16.3 (12.6 - 21.0)
A04 Golden empty nesters	114	3.50	32.6 (27.1 - 39.2)
A05 Provincial privilege	165	4.41	37.4 (32.1 - 43.6)
A06 High technologists	134	7.53	17.8 (15.0 - 21.1)
A07 Semi-rural seclusion	167	6.28	26.6 (22.9 - 31.0)
B08 Just moving in	5	0.64	7.8 (3.2 - 18.7)
B09 Fledgling nurseries	38	4.56	8.3 (6.1 - 11.5)
B10 Upscale new owners	33	5.34	6.2 (4.4 - 8.7)
B11 Families making good	139	7.46	18.3 (15.8 - 22.0)
B12 Middle rung families	273	10.74	25.4 (22.6 - 28.6)
B13 Burdened optimists	120	5.86	20.5 (17.1 - 24.5)
B14 In military quarters	5	0.32	15.6 (6.5 - 37.5)
C15 Close to retirement	298	10.29	29.0 (25.9 - 32.4)
C16 Conservative values	395	7.46	53.0 (48.0 - 58.5)
C17 Small time business	305	8.75	34.9 (31.2 - 39.0)
C18 Sprawling subtopia	392	9.13	42.9 (38.9 - 47.4)
C19 Original suburbs	223	7.69	29.0 (25.4 - 33.1)
C20 Asian enterprise	87	3.42	25.4 (20.6 - 31.3)
D21 Respectable rows	228	6.82	33.4 (29.4 - 38.1)
D22 Affluent blue collar	426	9.63	44.3 (40.2 - 48.7)
D23 Industrial grit	445	10.65	41.8 (38.1 - 45.8)
D24 Coronation street	262	6.03	43.5 (38.5 - 49.1)
D25 Town centre refuge	140	2.84	49.4 (41.8 - 58.3)
D26 South Asian industry	14	0.78	17.9 (10.6 - 30.3)
D27 Settled minorities	93	3.30	28.2 (23.0 - 34.6)

E28 Counter cultural mix	40	1.43	27.9 (20.5 - 38.1)
E29 City adventurers	22	1.30	17.0 (11.2 - 25.8)
E30 New urban colonists	47	2.46	19.1 (14.4 - 25.4)
E31 Caring professionals	51	2.10	24.3 (18.5 - 32.0)
E32 Dinky developments	34	2.00	17.0 (12.1 - 23.8)
E33 Town gown transition	31	1.36	22.8 (16.0 - 32.4)
E34 University challenge	8	0.71	11.2 (5.6 - 22.5)
F35 Bedsit beneficiaries	26	0.82	31.8 (21.7 - 46.7)
F36 Metro multiculture	61	1.80	33.9 (26.4 - 43.6)
F37 Upper floor families	135	3.24	41.7 (35.2 - 49.3)
F38 Tower block living	30	0.41	72.8 (50.9 - 104.2)
F39 Dignified dependency	243	2.08	117.1 (103.3 - 132.8)
F40 Sharing a staircase	71	0.70	100.9 (80.0 - 127.3)
G41 Families on benefits	76	3.21	23.7 (18.9 - 29.6)
G42 Low horizons	401	6.43	62.4 (56.6 - 68.8)
G43 Ex-industrial legacy	531	5.75	92.4 (84.9 - 100.6)
H44 Rustbelt resilience	529	8.61	61.4 (56.4 - 66.9)
H45 Older right to buy	487	6.23	78.2 (71.6 - 85.5)
H46 White van culture	486	10.44	46.6 (42.6 - 50.9)
H47 New town materialism	289	8.10	35.7 (31.8 - 40.1)
I48 Old people in flats	191	1.10	174.2 (151.1 - 200.7)
I49 Low income elderly	277	3.52	78.6 (69.9 - 88.4)
I50 Cared for pensioners	398	2.08	191.7 (173.8 - 211.5)
J51 Sepia memories	124	1.32	93.8 (78.6 - 111.8)
J52 Childfree serenity	119	2.74	43.4 (36.3 - 51.9)
J53 High spending elders	258	4.49	57.5 (50.9 - 64.9)
J54 Bungalow retirement	268	3.18	84.3 (74.8 - 95.0)
J55 Small town seniors	405	7.41	54.7 (49.6 - 60.3)
J56 Tourist attendants	65	1.28	50.8 (39.9 - 64.8)
K57 Summer playgrounds	43	0.76	56.4 (41.8 - 76.0)
K58 Greenbelt guardians	137	5.33	25.7 (21.7 - 30.4)
K59 Parochial villagers	134	4.09	32.8 (27.7 - 38.9)
K60 Pastoral symphony	91	3.19	28.5 (23.2 - 35.0)
K61 Upland hill farmers	20	0.78	25.8 (16.6 - 39.9)
0 (no data)	261	6.36	41.0 (36.3 - 46.3)
99 (unclassified)	52	1.34	38.8 (29.6 - 50.9)

Table 3.6. Mosaic groups and types with the highest incidence of lung cancer

Mosaic groups	
I Twilight subsistence	Older people living in social housing with high care needs
G Municipal dependency	Low income families living in estate based social housing
F Welfare borderline	People living in social housing with uncertain unemployment in deprived areas
Mosaic types	
I50 Cared-for pensioners	Older people receiving care in homes or sheltered accommodation
I48 Old people in flats	Older people living in small council and housing association flats
F39 Dignified dependency	Low income couples and pensioners living in crowded apartments in high density social housing

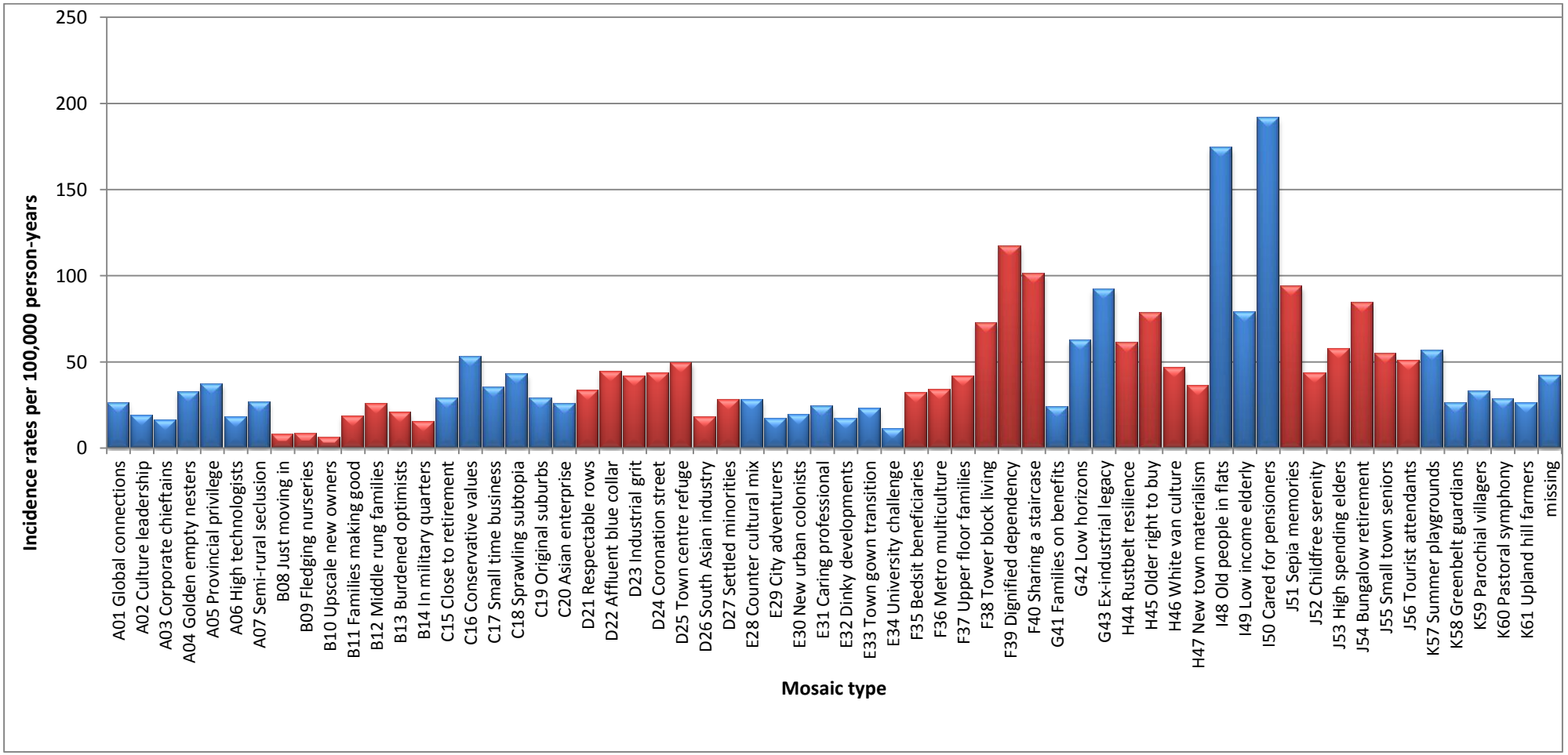


Figure 3.5: Lung cancer incidence by Mosaic Public Sector™ type

The estimated lung cancer incidence rates in each UK Primary Care Trust was derived using the THIN incidence rates of lung cancer for the different Mosaic types and the population of each Mosaic type in the different Primary Care Trusts (PCTs) in the UK. Figure 3.6 shows the estimated lung cancer incidence rates in the different regions in the UK.

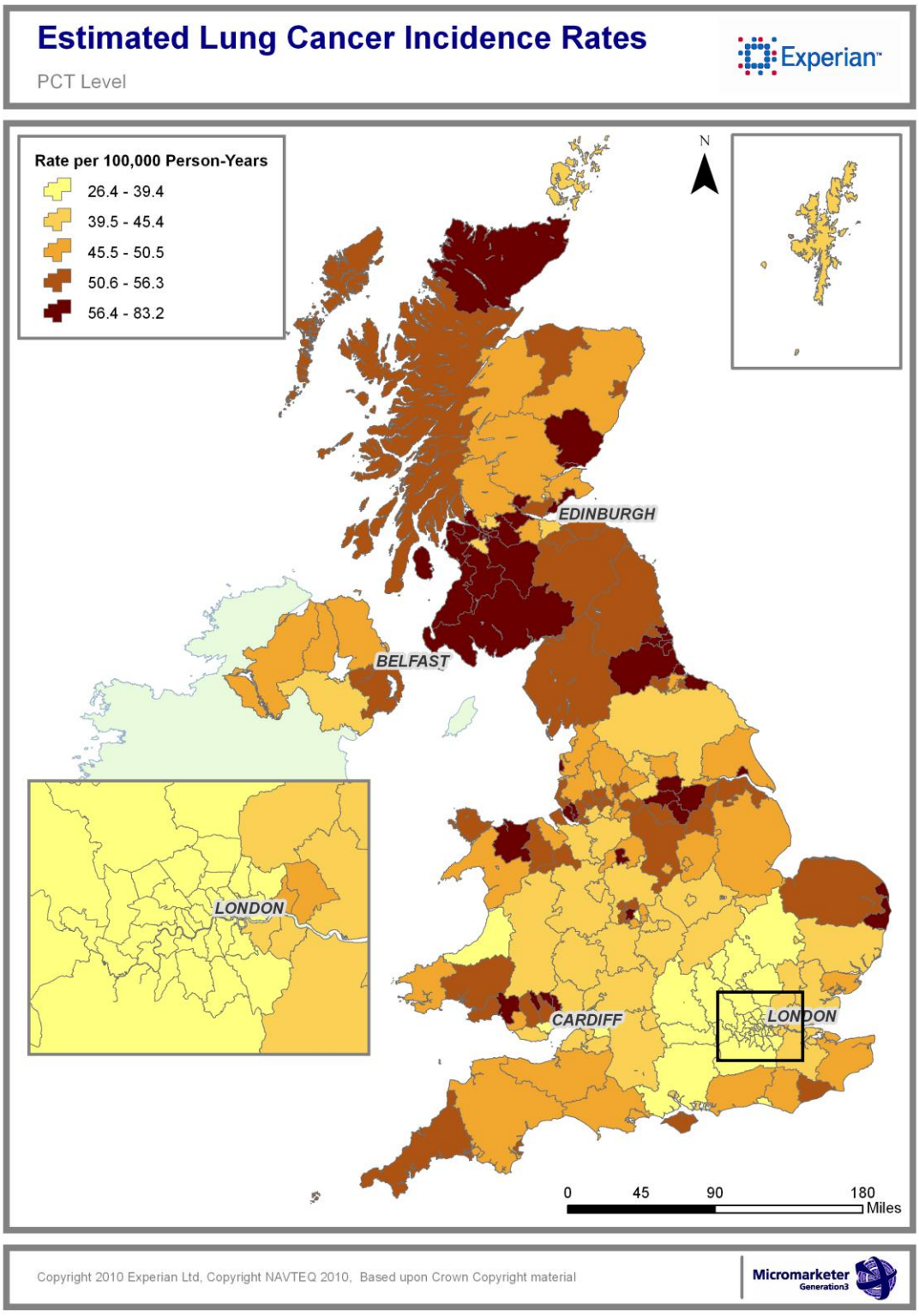


Figure 3.6: Estimated number of people in each primary care trust (PCT) in the UK likely to have lung cancer.
This was calculated using the population of each Mosaic type in the PCTs and the THIN lung cancer incidence rate by Mosaic type.
(Mapping by Experian UK)

3.3.3 Lung cancer survival in THIN

3.3.3.1 Overall survival

Among the 12,135 lung cancer cases studied, 8,885 (73.2%) died during the study period. Six months after diagnosis, 57% of the cases were still alive; one year after, 37% of the cases were alive and five years after, only 11% of the cases were alive. The median survival for the cases was 232 days (IQR: 76-630 days). This was only slightly better than survival in the National Lung Cancer Audit database (LUCADA)[231] where the median survival was 203 days with a one year survival of 32%.

3.3.3.2 Lung cancer survival by age and sex

Lung cancer survival worsened with increasing age at diagnosis (Table 3.7). For patients diagnosed at 40 years of age or less, the 1-year and 5-year survival were 52% and 31% respectively. One year and five year survival after lung cancer diagnosis at ages between 80 to 90 years were 29% and 6% respectively.

Table 3.7. Survival of lung cancer patients by age at diagnosis

Age at diagnosis	Median survival in days (IQR)	6 months survival	1-year survival	5-year survival	Unadjusted hazards ratio	95% CI	p-value
<40	457 (248- .)	85%	52%	31%	1.00	-	-
40-50	341 (148-1150)	70%	48%	17%	1.35	0.92-1.97	0.126
50-60	287 (116-830)	65%	42%	15%	1.54	1.08-2.20	0.018
60-70	274 (85-736)	61%	42%	13%	1.68	1.18-2.40	0.004
70-80	218 (72-604)	55%	36%	9%	1.94	1.36-2.77	<0.001
80-90	164 (54-443)	47%	29%	6%	2.41	1.69-3.45	<0.001
>90	147 (46-403)	40%	26%	-	2.72	1.85-4.01	<0.001

Male lung cancer patients died earlier than female patients with a median survival for males of 221 days (IQR: 72-580 days) compared with 251 days (IQR: 83-709 days) for females (Figure 3.7).

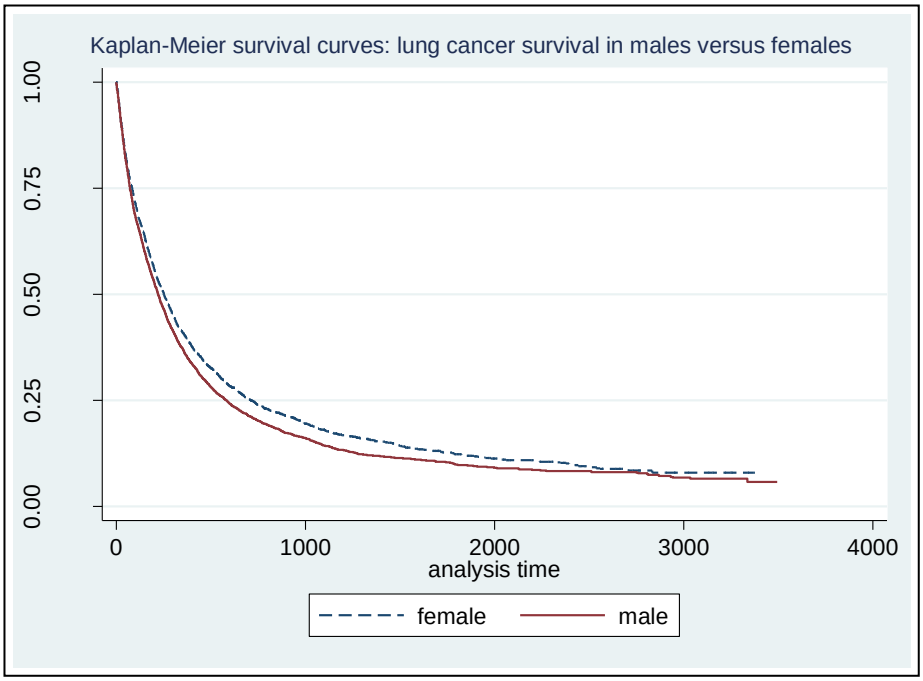


Figure 3.7: Kaplan-Meier survival plots showing lung cancer survival by sex

The percentages of males alive at 6 months, 1 year and 5 years after diagnosis were 55%, 36% and 10% respectively. Survival for females on the other hand at 6 months, 1 year and 5 years were 59%, 40% and 12% respectively (Table 3.8). Survival for patients in THIN was better than survival in the cancer registry[16], where the one year lung cancer survival was 27% for men and 30% for women. After adjusting for the effect of age at diagnosis, male lung cancer patients in THIN had 11% worse survival than female lung cancer patients (Hazards ratio for death - 1.11, 95% CI 1.06 to 1.16).

Table 3.8. Survival of lung cancer patients by sex

Lung cancer follow-up period	Males	females
6 months	0.55 (55%)	0.59 (59%)
1 year	0.36 (36%)	0.40 (40%)
5 years	0.10 (10%)	0.12 (12%)

3.3.3.2 Lung cancer survival by deprivation

Using the Townsend index deprivation quintile as a measure of socioeconomic status, survival did not differ across socioeconomic groups (Figure 3.8 & Table 3.9)

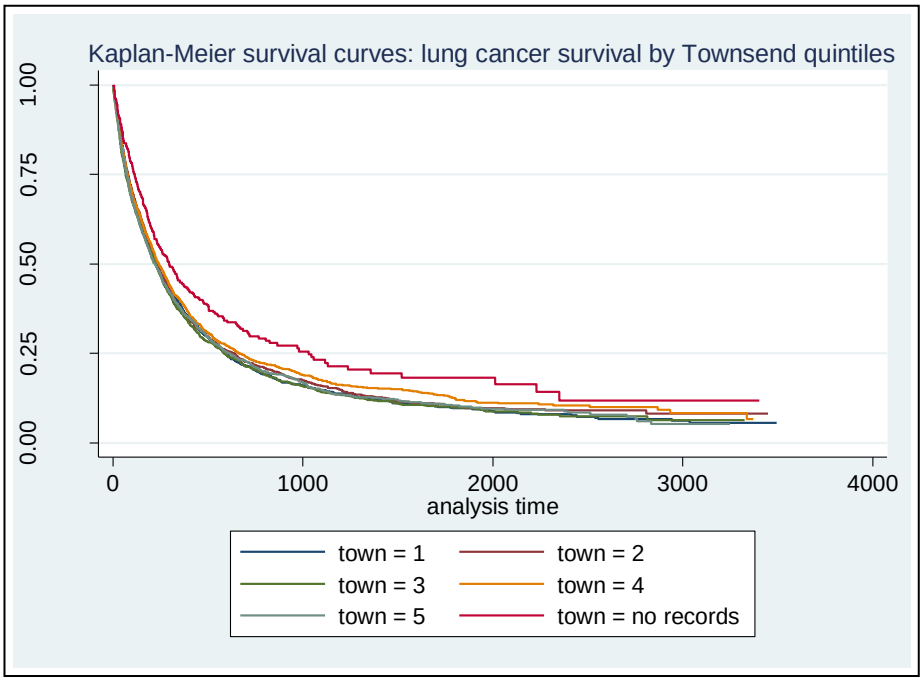


Figure 3.8: Kaplan-Meier survival plots showing lung cancer survival by Townsend deprivation quintiles

Table 3.9. Survival of lung cancer patients by Townsend deprivation quintiles

Townsend quintile	Median survival in days (IQR)	6 months survival	1-year survival	5-year survival	Unadjusted hazards ratio	95% CI	p-value
1	223 (78-593)	56%	37%	9.7%	1.00	-	-
2	232 (79-640)	57%	36%	10%	0.98	0.91-1.05	0.53
3	224 (67-587)	56%	36%	9.9%	1.03	0.96-1.10	0.46
4	242 (76-666)	58%	39%	12%	0.94	0.88-1.01	0.10
5	221 (72-608)	55%	37%	10%	1.01	0.94-1.09	0.82
missing	296 (116-1032)	64%	44%	18%	0.78	0.68-0.88	<0.001

3.4 Discussion

The overall incidence of lung cancer recorded in THIN general practices was 41.4 per 100,000 person-years between 2000 and 2009, however incidence from 2000-2002 was lower than in the latter periods of the study. This compares favourably with findings from a previous study which showed that the observed recording rates of pancreatic, colorectal and lung cancers in THIN prior to 2004 were lower than expected based on the national cancer registry data but increased and were more comparable to registry rates after 2004[209]. It has been suggested that a large increase in the recruitment of general practices to THIN in 2003 associated with receipt of training in data entry, experience in using the Vision software, and the institution of cancer quality improvement measures by the national Health Service in 2003 may have all contributed to the increase in recording of these cancers[209]. The introduction of the Quality and Outcomes Framework (QOF)[233] in 2004 which encourages general practitioners to record all new cases of cancer may also partly explain the increase in cancer recording in THIN. After comparing the lung cancer incidence rate in THIN with incidence rate recorded by the national cancer registry[230], this study confirms that THIN captures a higher proportion of lung cancer incidence in more recent years.

3.4.1 Lung cancer Incidence

There are two reliable national lung cancer databases in the UK against which THIN data were compared to assess its completeness and representativeness. These are the National Lung Cancer Audit database (LUCADA)[231] which has been shown to be highly representative of people with lung cancer in England[84]; and the national cancer registry data reported by the Office for National Statistics (ONS)[234] which is a good source of information on lung cancer incidence. Data reported by the ONS are systematically collected from all regional cancer registries in England, Wales, Scotland and Northern Ireland.

Reassuringly, the sex distribution of lung cancer cases in THIN, the median age at diagnosis and at death, and the increasing incidence with greater socioeconomic deprivation were all comparable to findings from LUCADA[84]. Comparison of the lung cancer incidence rate in THIN with the incidence rate reported by the national cancer registry[230], showed the incidence rate in THIN to be over 93% of the cancer registry incidence rate. Geographical variations in lung cancer incidence in THIN were also mostly similar to registry data. The highest incidence rates were in the North-West of England, North-East of England and Scotland while the South East Coast and London had the lowest incidence. Cancer registry data however, shows incidence in London to be exceptionally high compared to other SHA regions in southern England. This is in contrast to THIN where the lowest incidence of lung cancer was in London, which may be due to THIN's over recruitment of practices covering slightly more affluent areas[216 217]. The population of THIN also has an over-representation of practices from the South-East of England where incidence rates are among the lowest so it is therefore unsurprising that the crude overall lung cancer incidence in THIN is marginally lower than the incidence rates based on registry data. The difference between THIN and registry incidence rates may also be partly

attributed to the fact that about 6.8% of cases included in the UK cancer registries are from death certificates only[18].

3.4.2 Societal distribution of lung cancer

The association that was found between lung cancer incidence and greater socioeconomic deprivation was independent of age, sex and general practice and is consistent with findings from other studies[90 96]. Variations in lung cancer incidence were however, more marked in the Mosaic groups and types than in Townsend deprivation quintiles. Mosaic Public Sector™ segmentation classifies UK households and postcodes into several lifestyle groups and types based on finer characteristics which has enabled the identification of much higher incidence rates of lung cancer in specific sectors of society. Mosaic Public Sector™ types I50 (Cared for pensioners), I48 (Old people in flats) and F39 (Dignified dependency) had the highest lung cancer incidence rates and this was unsurprising considering the fact that these Mosaic Public Sector™ types are characterised mostly by older people who have poor levels of education, are mostly reliant on state benefits and live relatively less healthy lifestyles including above average smoking rates.

Mosaic classification is done at the household as well as the postcode level and although about half (54%) of the data used for Mosaic profiling are sourced from the 2001 Census, the other 46% are derived from sources such as the Experian Lifestyle Survey, consumer credit databases, the electoral roll, shareholder registers, Land registry data, Council Tax information, the Hospital Episode Statistics, the British Crime Survey, Expenditure and Food Survey and other sources[222]. Mosaic profiling is therefore based on an exchange of information which enhances a deeper understanding of the characteristics of people in the various groups and types[226] unlike the Townsend Index which uses a less

complex classification of postcodes based on measures of socioeconomic deprivation from Census data[221]. To accurately target public health resources and develop tailored public health campaigns and interventions, the differing needs of deprived populations have to be identified and understood and in this regard, Mosaic classification is particularly valuable.

3.4.3 Lung cancer survival

Median survival for people with lung cancer in THIN was only slightly better than survival in LUCADA[84]. The survival estimates in THIN and LUCADA were marginally higher when compared with survival in the cancer registry[16] and most likely reflect the different methods of case ascertainment[141]; in particular, the registry ascertains cases with a diagnosis of lung cancer only on a death certificate whilst these cases, having no supporting clinical data prior to death, may not have been recorded in THIN nor LUCADA.

Socioeconomic deprivation did not affect survival of people with lung cancer in THIN and this is consistent with the findings from LUCADA[84]. This lack of association may reflect the dismal prognosis of lung cancer in general and the lack of effective treatments for most people with lung cancer.

3.4.4 Strengths and limitations of this study

Some of the limitations of using general practice data such as THIN for this study include the limited scope of data recording and variation in the diagnostic criteria for medical conditions that were previously discussed in Chapter 2 (Section 2.1.3 - strengths and weaknesses of THIN). Although it was considered necessary to explore the survival of lung cancer patients in THIN in relation to the cancer histology and patients' performance status, there were insufficient data on these

variables to enable these analysis. Performance status records were available for only 14 patients with lung cancer (1.15% of lung cancer patients in THIN) and despite retrieving some histology records from the medical free text comments, histology records were available for only 1704 patients overall (14% of patients in the dataset). Due to the lack of power that may result from analysis of these few numbers, the effect of performance status and histology on lung cancer survival were therefore not explored.

Detailed information about how Mosaic groups and types are derived are not disclosed by Experian and this limits the ability to assess the validity of their methods. Some health information have also been used in deriving the Mosaic classifications and this may confound the identification of groups with the highest lung cancer incidence. By using data from 400 variables to profile all postcodes in the UK into 61 Mosaic types, it is not likely that any postcode or household will conform with all of the values characteristic of its type and in fact, a few postcodes may not fall into any category. However, it is worth noting that Mosaic types identify groups of individuals and households that are as similar as possible to each other and as different as possible to other groups[101].

A major strength of this study is that this is the first lung cancer study to incorporate the Experian's Mosaic Public Sector™ classification tool and this tool provides a finer and more detailed classification of the UK population than any other socio-demographic classification markers such as Townsend deprivation index[221] and therefore allows programs and interventions to be tailored to the specific needs of the population.

3.5 Conclusion

The analyses in this study have shown that general practice data from THIN are representative of lung cancer in the UK and capture the vast majority of cases from cancer registries. UK general practice data are thus a potentially valuable tool for lung cancer research as they are the only source of detailed prospectively collected health information available at a population level both before and after lung cancer diagnosis. Linkage of patients' records to Experian's Mosaic Public Sector™ classification has also provided a more refined knowledge of the sectors of society where lung cancer incidence is highest in the UK. As such, Mosaic could be used outside general practice as an important tool to reduce lung cancer-related health inequalities by enabling tailored public health campaigns and interventions to be more precisely and thus effectively targeted geographically to specific lifestyle groups in society.

Chapter 4. The use of a matched case-control dataset to explore differences in the smoking-associated risk of lung cancer

The previous chapter assessed the validity of lung cancer records in THIN database concluding that it was representative of lung cancer in the UK and therefore a valid source of data for lung cancer research. In this chapter, a dataset of lung cancer cases and controls, matched on age (year of birth), sex and general practice is developed with the primary aim of piloting the methods for the development of a lung cancer risk-prediction score, including the assessment of the timing of symptoms and other clinical features that are likely to be predictive of lung cancer. The chapter goes on to describe several studies that were conducted using the matched case-control dataset to investigate the association between cigarette smoking and lung cancer in different subgroups of patients in general practice. In particular, socioeconomically deprived individuals and those with depression were studied because these are subgroups of people with particularly high smoking prevalence and high levels of cigarette smoke addiction. Since age and sex are associated with lung cancer incidence (shown in results in Chapter 3, section 3.3.2.2 - Lung cancer incidence by age groups and sex), performing these analyses in a population of cases and controls matched by age, sex and general practice allows the confounding effects of age, sex and the variable recording in general practices, to be dealt with during the design stage of the study.

4.1 Derivation of the matched case-control dataset

4.1.1 Criteria for selection of cases

The cases included in this matched case-control dataset were the incident cases of lung cancer derived in Chapter 2 (section 2.3). As mentioned earlier, only incident cases of lung cancer first diagnosed between the 1st of January 2000 and the 28th of July 2009 were included in the study. The eligibility criteria for case selection are as summarised below:

- First coded diagnosis of lung cancer between the 1st of January 2000 and the 28th of July 2009
- Actively registered in the GP practice for at least 1 year before diagnosis
- Exclusion of cases without a month of diagnosis, cases with a date of diagnosis more than 31 days after death, diagnosis more than 31 days after the finish date and cases with a recorded date of death more than 31 days after the finish date.

The total number of eligible cases that were identified in THIN database were 12,135. Lung cancer is rare in individuals less than 40 years and analysis of the THIN dataset in this thesis has shown lung cancer to be rare in individuals less than 40 years (59 patients less than 40 years; 0.49% of the case population). Based on this, subsequent analysis in this thesis excluded patients less than 40 years. In total, 12,076 eligible cases were available to develop the matched case-control dataset for the analyses in the following studies.

4.1.2 Criteria for selection of controls

Each case in the dataset was matched with up to four controls randomly selected from the patient population in THIN. Controls were matched to cases using the following criteria.

- Same sex as their matched case
- Same age (year of birth) as the matched case
- Registered at the same general practice as the case
- Have general practice records for at least 1 year prior to the date of lung cancer diagnosis in the matched case (also known as the index date)
- No record of lung cancer or mesothelioma in their record
- Alive and contributing to THIN at the time of lung cancer diagnosis in the matched case

4.1.3 Overall matched case-control population

A total of 5,256 cases were matched with 4 controls each, 4,008 cases matched with 3 eligible controls each, 1,933 controls matched with 2 controls each and 691 cases each had only 1 eligible control. In total therefore, there were 49,493 patients in the case-control population comprising of 11,888 cases and 37,605 controls. There were 188 cases who did not have any eligible controls to match with and these cases were excluded from further analyses in this study. All cases and controls were derived from 445 UK general practices.

There were 7,025 male and 4,863 female lung cancer cases in the dataset, making up 59.1% and 40.9% of the lung cancer population respectively. The median age of lung cancer diagnosis was 72.5 years (IQR 64.5 to 78.8 years). The median follow-up time prior to lung cancer diagnoses was similar in the cases and controls at 9.5 years (IQR 5.5 years to 13.5 years) and 9.4 years (IQR 5.4 years to 13.2 years) respectively.

4.2 Factors to be investigated in this chapter

The following sections in this chapter describe studies that firstly explore the features of general practice patients before lung cancer diagnosis and then investigate several hypotheses on the variation in lung cancer risk among different sub-groups of smokers.

Section 4.4 uses the matched case-control dataset to identify factors that are predictive of lung cancer in general practice.

Section 4.5 investigates whether the association between cigarette smoking and lung cancer differs between individuals of different socioeconomic groups.

Section 4.5 investigates whether there is variation in the risk of lung cancer among smokers with a history of depression compared to those who have no history of depression in general practice.

Section 4.6 summarises the result from another research project which was done using the dataset created in this thesis, to investigate whether the risk of lung cancer differs between men and women with the same recorded quantity of cigarettes smoked.

4.3 Definition of variables analysed in this chapter

This section describes the variables that were analysed in the studies in this chapter. While some variables were exclusive to one study, others were common to more than one study. Detailed analyses for the different studies are discussed in the relevant sections.

4.3.1 Age and sex

Demographic information such as date of birth and sex are available for all patients in THIN database. Children up to the age of 15 years of age have their month and year of birth recorded in THIN, however on reaching the age of 15, only the year of birth is recorded. For the purpose of analyses, the date of birth of individuals over the age of 15 years in THIN was assumed as the 1st of July of the recorded year of birth. The following studies in this thesis have included only patients aged 40 years of age or older and age was defined as age on the index date of lung cancer. Since the cases and controls in this chapter were matched on age and sex, these variables were identical for all patients in a matched set.

4.3.2 Deprivation

Townsend quintile of deprivation was previously described in Chapter 2 (Section 2.1.5). All patients in THIN are assigned to a Townsend quintile corresponding to their level of deprivation, and these quintiles are made available with the demographic records of patients in the database. The Townsend quintiles range from 1 to 5, with quintile 1 representing the least deprived quintile and quintile 5 representing the most deprived quintile.

4.3.3 Smoking

All records of smoking status were retrieved from patient's records using the smoking Read codes listed in Appendix I. Patients were categorised according to their smoking status prior to lung cancer, as current, ex or non smokers. Records of daily cigarette consumption prior to the diagnosis of lung cancer were

also retrieved from patients who were "current- " or "ex-" smokers. Two types of records of daily cigarette consumption were extracted from the patients' notes:

- The last record of daily consumption prior to the lung cancer index date
- The highest ever recorded daily consumption prior to the index date

In obtaining the quantity of cigarettes smoked, all smoking records made within the six months before lung cancer diagnosis were excluded to account for a possible change in the cases' cigarette consumption in the months preceding lung cancer diagnosis. Based on their cigarette consumption, patients were classified as: non-smokers, trivial/light smokers (1 to 9 cigarettes smoked daily), moderate smokers (10 to 19 cigarettes smoked daily) and heavy/very heavy smokers (20+ cigarettes per day). Current smokers who had no record of their daily cigarette consumption were recorded as such - (smoker with no recorded quantity) and patients who had no recorded smoking information and who were not known to be non-smokers were included in a separate category (missing smoking records).

4.3.4 Clinical features

The symptoms and diagnoses that were analysed in cases and controls were defined using two sources. Firstly, the symptoms recommended by the NICE guidelines[147] for referral of suspected cases of lung cancer and indications for chest x-ray; These were cough, haemoptysis, chest/shoulder pain, voice hoarseness, dyspnoea and weight loss. In addition, the six most common symptoms and diagnoses in the records of patients with lung cancer other than the symptoms in the NICE guidelines (complete list of most common symptoms and diagnoses in the medical records of patients with lung cancer is shown in Appendix II) were assessed; These were upper respiratory tract infections (URTI), lower respiratory tract infections (LRTI), non-specific chest infections,

constipation, depressive disorders and Chronic Obstructive Pulmonary Disease (COPD). Records of these symptoms and diagnoses prior to lung cancer diagnosis, were extracted from patients' datasets using lists of Read codes for the different conditions (Read codes listed in Appendix I).

Records of chest x-rays, blood tests and general practice consultations for symptoms other than those already assessed, were also retrieved from the patients' records.

4.4 The use of a matched case-control dataset to identify the factors predictive of lung cancer

As stated in chapter 1, a major objective of this thesis is to develop a lung cancer risk-prediction score using patient features in primary care that are predictive of lung cancer before diagnosis. In order to ensure that the timing of clinical features for the development of the score were accurately determined, it was considered a worthwhile exercise to pilot the methods for identifying the lung cancer predictors using the matched case-control dataset developed in this chapter prior to the score development with a different dataset. Since the cases and controls in the dataset in this chapter have been matched on age and sex, the effect of these variables in predicting lung cancer cannot be assessed. However, this pilot study enabled the identification of other predictors in general practice as well as allow the timing of symptoms and other clinical features to be determined.

4.4.1 Methods

Conditional logistic regression was used to estimate the relative odds and 95% CI for lung cancer, by smoking status, daily cigarette consumption and deprivation. Before conducting analyses on patients' clinical features, the median period of general practice follow-up for the cases and controls were assessed to ensure that they were comparable. The pattern and frequency of symptom presentation in cases and controls prior to lung cancer diagnosis were then assessed by way of frequency plots for the different symptom records. This allowed an estimation of the time periods when symptom consultation patterns differed in the cases and when they could be used to predict a future diagnosis of lung cancer. To identify the precise time periods when clinical factors were independently associated with lung cancer, conditional logistic regression analyses were done to estimate the odds ratio and 95% CI for lung cancer with the different clinical factors firstly in the 0-6 months and the 6-24 month periods, and then over shorter 6-monthly time periods: 0-6 months, 6-12 months, 12-18 months and the 18-24 months before diagnosis. To determine the independent predictors of lung cancer, multivariate analyses were done using the smoking, deprivation and clinical variables that were associated with lung cancer in univariate analyses at the 6-24 month period before diagnosis using a statistical significance cut-off level of $p<0.05$. Variables that were not significant in multivariate analysis were removed from the model and those variables that were previously not associated with lung cancer in univariate analysis were again checked for significance in the final model.

4.4.2 Results

4.4.2.1 Socioeconomic deprivation and smoking characteristics of cases and controls

Using the Townsend deprivation quintiles as a measure of socioeconomic deprivation, increasing deprivation was associated with a greater likelihood of lung cancer (Table 4.1).

Smoking status of patients prior to lung cancer diagnosis were available for 34,313 controls (91.2% of controls) and 11,383 cases (95.8% of cases). In total therefore, there were smoking records for 45,696 out of the 49,493 patients in the dataset (92.3% of patients). Results in Table 4.1 show that a higher proportion of controls were non smokers compared to cases (39.2% and 10.6% respectively). Compared to controls, the likelihood of a case being a current smoker was 11.43 (95% CI 10.59-12.34) and the likelihood of a case being an ex-smoker was 5.33 (95% CI 4.95-5.75). Patients with lung cancer smoked more cigarettes per day compared to controls and controls were more likely to be trivial smokers of less than 1 cigarette per day.

Analysis of the highest and the latest recorded quantity of cigarettes smoked daily by cases and controls up to 6 months before diagnosis shows that in the period before lung cancer diagnosis, there was a reduction in the proportion of cases who were heavy and very heavy smokers as well as an increase in the proportion of cases who were moderate, light and trivial smokers. Although the controls showed a similar decrease in heavy cigarette consumption over time, these were not as marked as in the cases. Based on this finding and taking into account the fact that individuals' smoking consumption can change over time, the highest ever recorded daily cigarette consumption was used as a proxy marker of patients' cigarette exposure in all subsequent analyses.

Analysis using a combination of patients' smoking status and daily cigarette consumption showed an increase in the odds ratio for lung cancer with an increase in the daily cigarette consumption and the odds were greater in current smokers compared to ex smokers. The odds ratio for lung cancer among current smokers of 40+ cigarettes per day was 21.97 (95% CI 18.65-25.88) whereas the odds ratio among ex smokers of 40+ cigarettes per day was 8.56 (95% CI 7.08-10.34).

Table 4.1 Socioeconomic deprivation and smoking status of cases and controls

	Control n(%) n=37,605	Case n(%) n=11,888	Unadjusted odds ratio for lung cancer (95% CI)
Townsend deprivation quintile			
5 (most deprived)	5,064 (13.47)	2,196 (18.47)	2.26 (2.08-2.44)
4	6,742 (17.93)	2,609 (21.95)	1.88 (1.75-2.03)
3	7,420 (19.73)	2,380 (20.02)	1.48 (1.38-1.59)
2	8,187 (21.77)	2,200 (18.51)	1.19 (1.11-1.28)
1 (least deprived)	8,735 (23.23)	2,037 (17.13)	1.00
Missing Townsend records	1,457 (3.87)	466 (3.92)	1.65 (1.41-1.94
Smoking status prior to lung cancer diagnosis			
Current smoker	7,369 (19.60)	5,458 (45.91)	11.43 (10.59-12.34)
Ex smoker	12,403 (32.98)	4,748 (39.94)	5.33 (4.95-5.75)
Non smoker	14,541 (38.67)	1,177 (9.90)	1.00
Missing smoking records	3,292 (8.75)	505 (4.25)	1.89 (1.67-2.13)
Daily cigarette consumption up to 6 months before diagnosis			
Highest record of cig/day			
Very heavy (40+/day)	730 (1.94)	685 (5.76)	15.06 (13.21-17.16)
Heavy (20-39/day)	3,949 (10.50)	3,607 (30.34)	14.02 (12.91-15.22)
Moderate (10-19/day)	3,720 (9.89)	2,410 (20.27)	9.04 (8.31-9.84)
Light (1-9/day)	2,234 (5.94)	983 (8.27)	5.86 (5.30-6.49)
Trivial (<1/day)	143 (0.38)	20 (0.17)	1.95 (1.20-3.20)
Smoker, no quantity recorded	8,460 (22.50)	2,169 (18.25)	3.23 (2.98-3.51)
Non smoker	14,729 (39.17)	1,260 (10.60)	1.00
Missing smoking records	3,640 (9.68)	754 (6.34)	2.38 (2.14-2.65)
Latest record of cig/day			
Very heavy (40+/day)	459 (1.22)	376 (3.16)	13.02 (11.11-15.26)
Heavy (20-39/day)	2,935 (7.80)	2,482 (20.88)	12.80 (11.73-13.97)
Moderate (10-19/day)	3,940 (10.48)	2,852 (23.99)	10.20 (9.40-11.08)
Light (1-9/day)	3,257 (8.66)	1,957 (16.46)	8.24 (7.55-8.99)
Trivial (<1/day)	185 (0.49)	38 (0.32)	2.86 (1.97-4.16)
Smoker, no quantity recorded	8,460 (22.50)	2,169 (18.25)	3.22 (2.97-3.49)
Non smoker	14,729 (39.17)	1,260 (10.60)	1.00
Missing smoking records	3,640 (9.68)	754 (6.34)	2.41 (2.16-2.68)
Smoking status and highest daily cigarette consumption			
Current V heavy (40+/d)	340 (0.90)	465 (3.91)	21.97 (18.65-25.88)
Current Heavy (20-39/d)	2,389 (6.35)	2,578 (21.69)	16.90 (15.44-18.49)
Current Mod (10-19/d)	2,028 (5.39)	1,645 (13.84)	11.41 (10.37-12.54)
Current Light (1-9/d)	1,138 (3.03)	597 (5.02)	7.19 (6.36-8.13)
Current Trivial (<1/d)	46 (0.12)	7 (0.06)	2.30 (1.00-5.27)
Ex V heavy (40+/d)	390 (1.04)	220 (1.85)	8.56 (7.08-10.34)
Ex Heavy (20-39/d)	1,560 (4.15)	1,029 (8.66)	9.70 (8.71-10.80)
Ex Mod (10-19/d)	1,692 (4.50)	765 (6.44)	6.02 (5.38-6.74)
Ex Light (1-9/d)	1,096 (2.91)	386 (3.25)	4.37 (3.80-5.02)
Ex Trivial (<1/d)	97 (0.26)	13 (0.11)	1.73 (0.94-3.17)
Non-smoker	14,729 (39.17)	1,260 (10.60)	1.00
Smoker, no quantity recorded	8,460 (22.50)	2,169 (18.25)	3.20 (2.95-3.47)
Missing smoking records	3,640 (9.68)	754 (6.34)	2.41 (2.17-2.69)

4.4.2.2 Clinical features prior to lung cancer diagnosis

4.4.2.2.1 Duration of registration in general practice

To ensure that the clinical records of cases and controls were comparable, the average period of their registration in general practice prior to the index date of lung cancer were determined. Cases had a median general practice registration duration of 9.5 years (IQR 5.5 years to 13.5 years) while controls had a registration duration of 9.4 years (IQR 5.4 years to 13.2 years) before lung cancer diagnosis in their matched case.

4.4.2.2.2 Overall pattern of consultations by cases and controls

The median number of consultations per case in the 5 years before lung cancer diagnosis was 287 (IQR 142 to 510) and the median number of consultations per control within the same period was 198 (IQR 79 to 393). Within the 2 years before diagnosis, the median number of consultations per case was 168 (IQR 89 to 278) and the median number per control was 107 (IQR 42 to 204). Plots of the frequency of consultations among patients in the dataset within the 5 year and 2 year periods before diagnosis (Figure 4.1 & Figure 4.2), show a similar pattern of consultation in the cases and controls up to the year before lung cancer diagnosis when there is a considerable increase in the consultation frequency for cases.

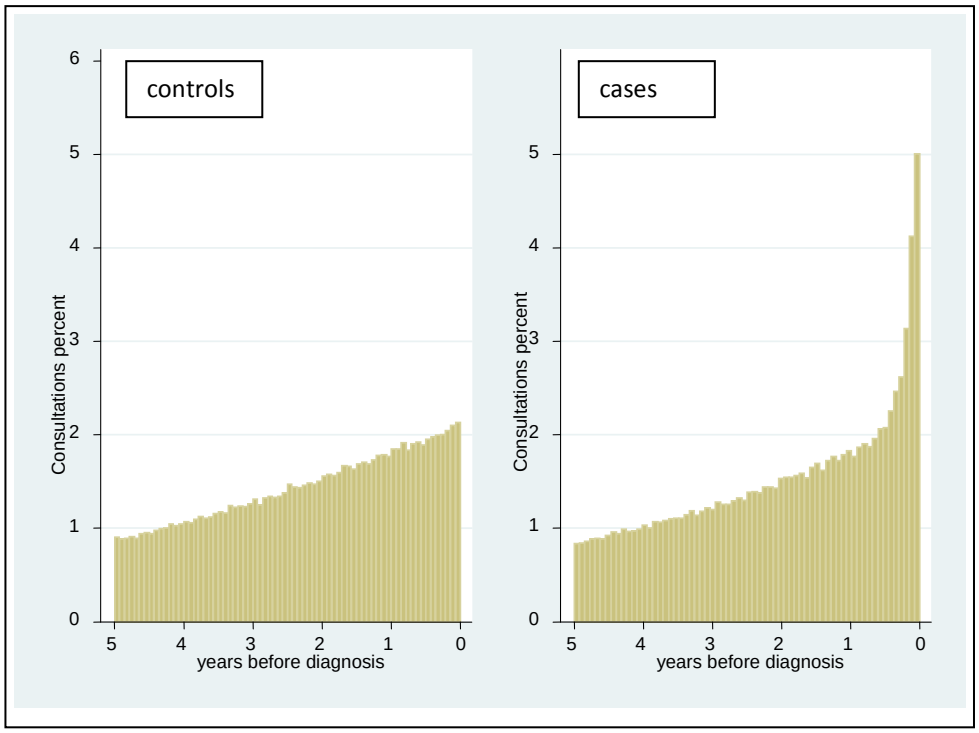


Figure 4.1 General consultations by cases and controls, 5 years before lung cancer diagnosis
(the height of the bars are scaled so that the sum of their height equals 100)

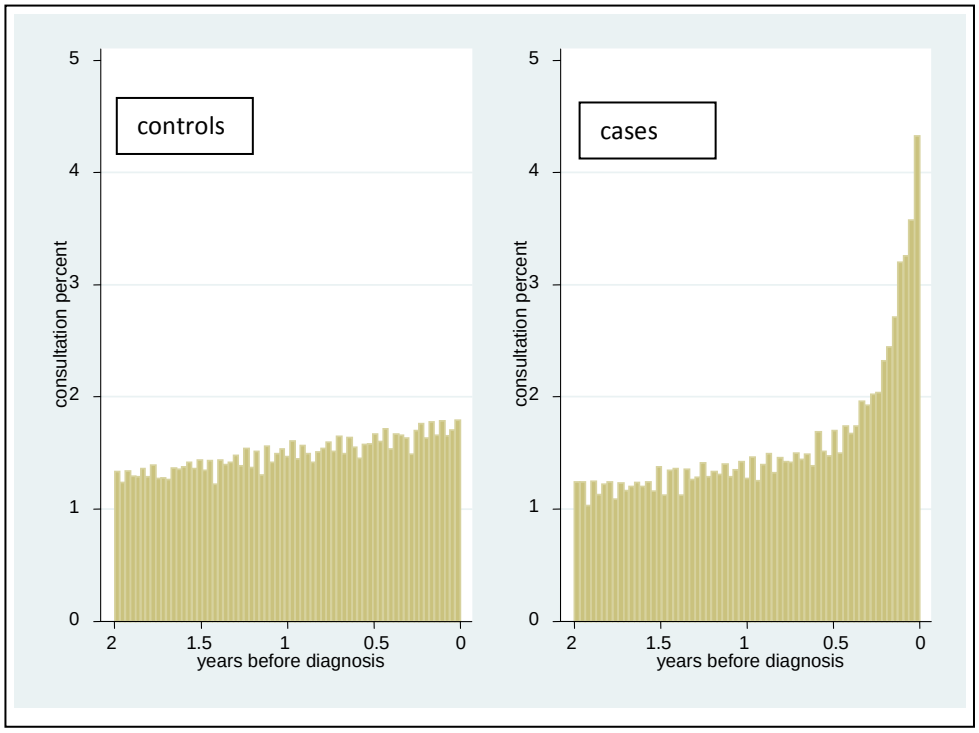


Figure 4.2 General consultation by cases and controls, 2 years before lung cancer diagnosis
(the height of the bars are scaled so that the sum of their height equals 100)

4.4.2.2.2 Pattern of symptom consultations prior to lung cancer diagnosis

As shown in Figure 4.3a, there was a considerable increase in the frequency of symptom presentation by cases, for all the lung cancer symptoms detailed in the NICE guidelines. This increase in symptom presentation in general practice are shown to have occurred within the year before lung cancer diagnosis. Plots of the most commonly recorded symptoms and diagnosis in the dataset of lung cancer cases - constipation, depression, URTI, LRTI, chest infections and COPD, also show an increase in the general practice presentation of these symptoms before lung cancer diagnosis (Figure 4.3b). However, among the most commonly recorded symptoms in the case dataset, the increase in the pattern of general practice presentation were more marked for LRTI, chest infections, depression and COPD.

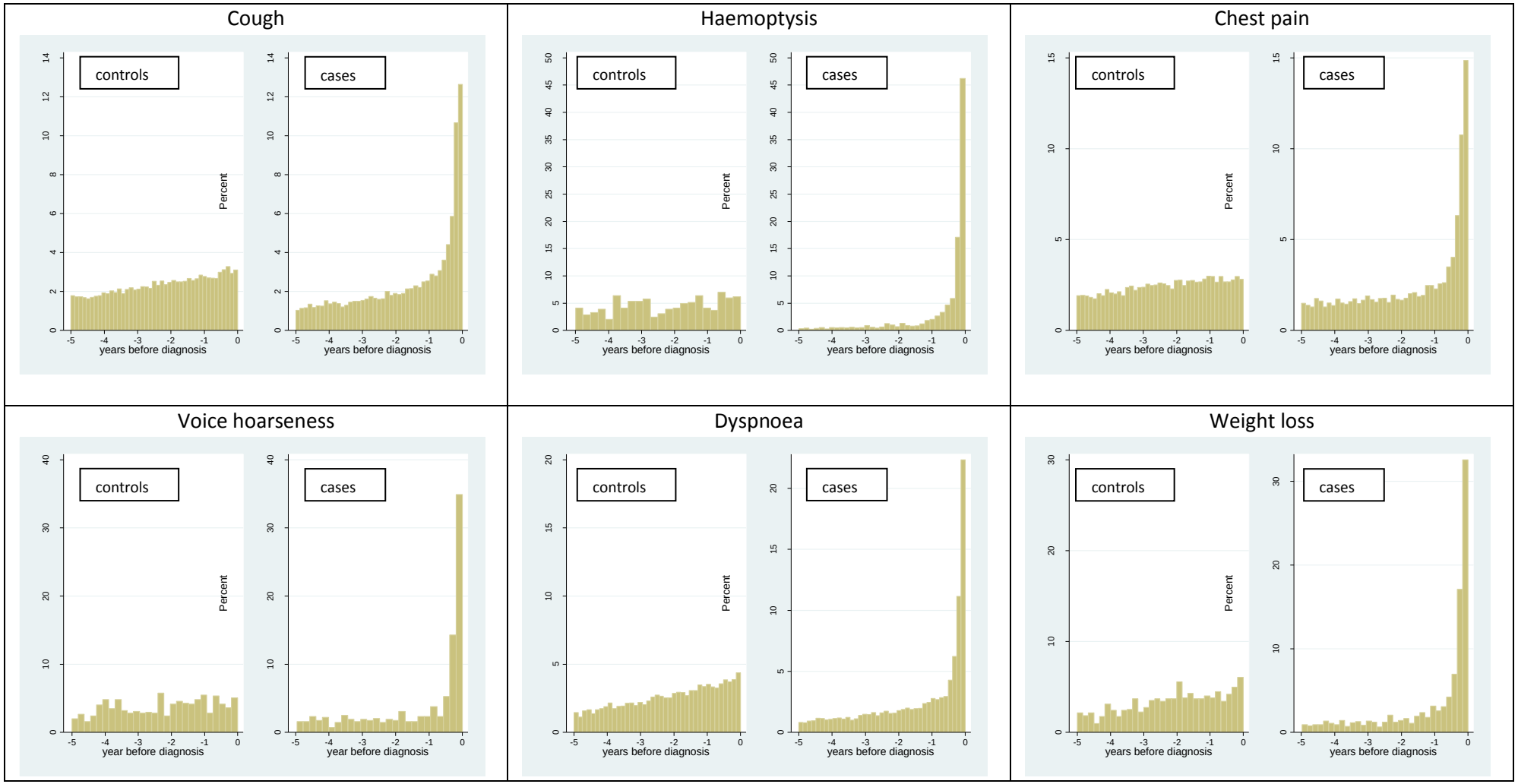


Figure 4.3a Plots showing the frequency of symptom records* in cases and controls, 5 years before lung cancer diagnosis

*Symptoms recommended by the NICE guidelines for indications for chest x-ray or referral of suspected cases of lung cancer

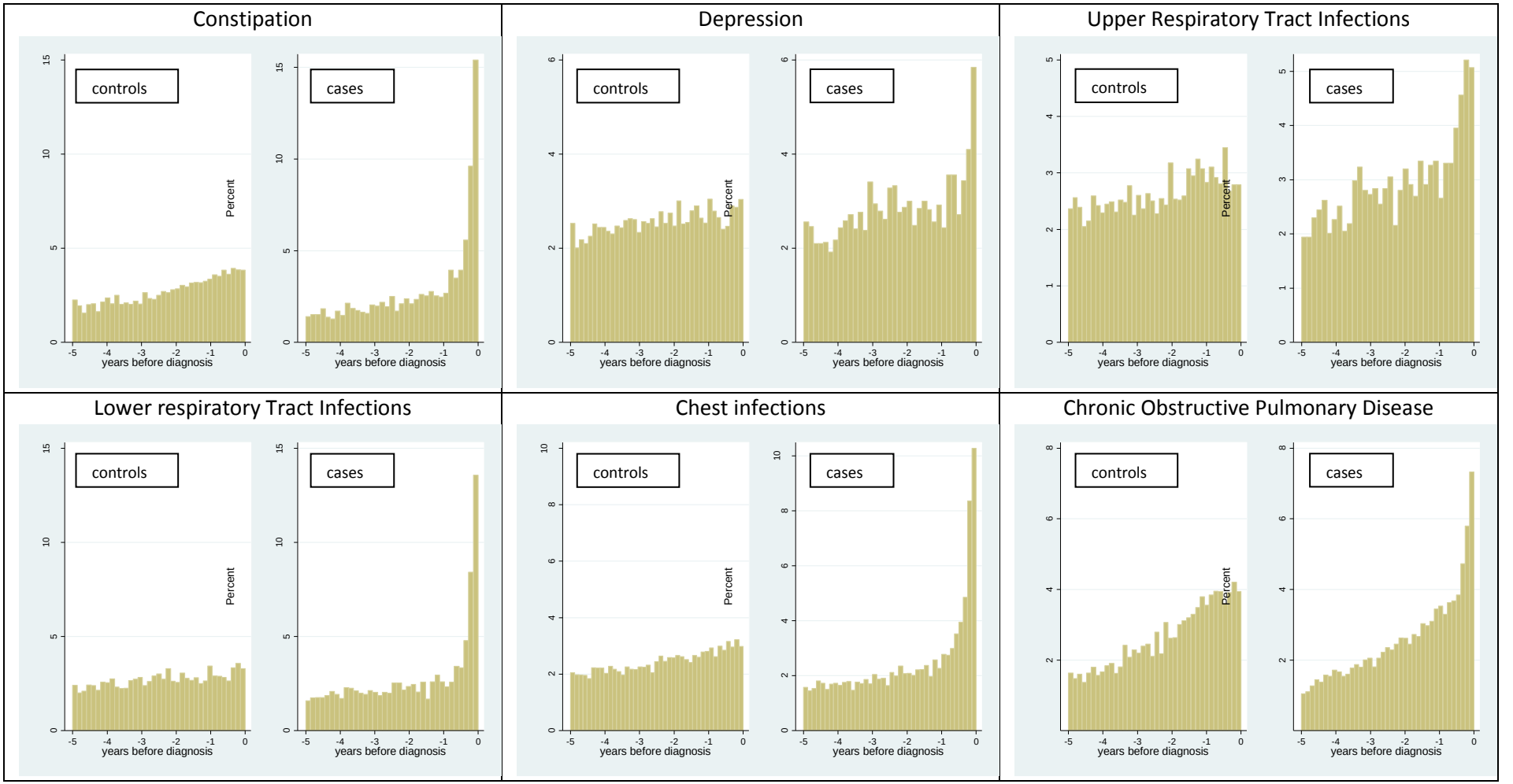


Figure 4.3b Plots showing the frequency of symptom records in cases and controls, 5 years before lung cancer diagnosis**

**six of the most commonly recorded symptoms and diagnosis in the medical records of cases before lung cancer diagnosis
(the height of the bars are scaled so that the sum of their height equals 100)

4.4.2.2.3 Pattern of clinical investigations before lung cancer diagnosis

Figure 4.4 shows an increase in the frequency of chest x-rays and blood investigations among cases compared to controls, before the diagnosis of lung cancer was made.

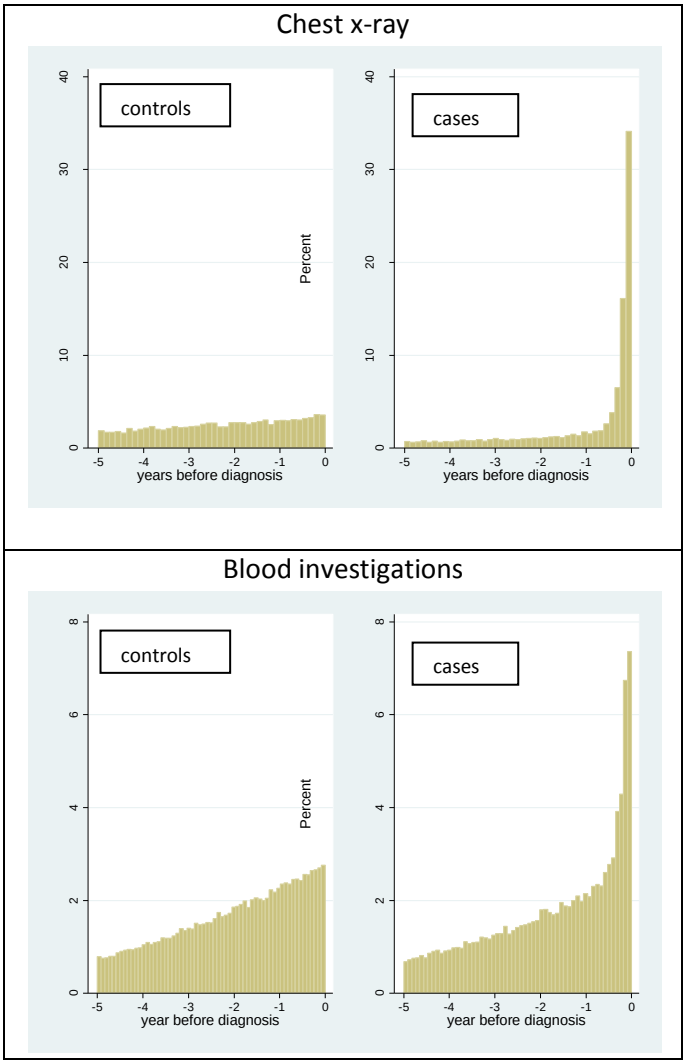


Figure 4.4 The frequency of chest x-ray and blood investigations, 5 years before lung cancer diagnosis
(the height of the bars are scaled so that the sum of their height equals 100)

4.4.2.2.4 Symptoms and investigations associated with lung cancer in general practice

Table 4.2 shows the univariate association between lung cancer and patients' clinical features in the 0-6 and 6-24 month periods before diagnosis. Results from the 6-monthly sub-analysis of the 6-24 month clinical records are also shown. The largest proportion of symptoms and investigations by cases were made in the 0-6 month period before diagnosis and the symptoms with the largest odds ratio for lung cancer were haemoptysis and weight loss while the investigation that was most strongly associated with lung cancer was chest investigations. The majority of records made 6-24 months before diagnosis were made in the 6-12 month period before diagnosis.

Table 4.2 Univariate association between lung cancer and general practice symptoms and investigations up to 24 months before diagnosis

Symptom before lung cancer	Control n(%) N=37,605	Case n(%) N=11,888	Unadjusted OR for lung cancer (95% C)
Cough			
0-6 months	2,259 (6.01)	3,232 (27.19)	6.15 (5.77-6.55)
6-24 months	4,722 (12.56)	2,589 (21.78)	1.95 (1.85-2.07)
6-12 months	2,009 (5.34)	1,386 (11.66)	2.33 (2.17-2.51)
12-18 months	1,949 (5.18)	1,072 (9.02)	1.76 (1.63-1.91)
18-24 months	1,848 (4.91)	937 (7.88)	1.64 (1.51-1.79)
Haemoptysis			
0-6 months	54 (0.14)	1,108 (9.32)	75.52 (56.21-101.48)
6-24 months	128 (0.34)	272 (2.29)	6.82 (5.50-8.44)
6-12 months	46 (0.12)	161 (1.35)	11.12 (7.99-15.47)
12-18 months	38 (0.10)	82 (0.69)	6.76 (4.56-10.01)
18-24 months	48 (0.13)	56 (0.47)	3.66 (2.48-5.41)
Chest/shoulder pain			
0-6 months	1,330 (3.54)	1,953 (16.43)	5.69 (5.26-6.15)
6-24 months	3,330 (8.86)	1,463 (12.31)	1.46 (1.36-1.56)
6-12 months	1,315 (3.50)	697 (5.86)	1.74 (1.58-1.92)
12-18 months	1,301 (3.46)	548 (4.61)	1.34 (1.21-1.49)
18-24 months	1,253 (3.33)	476 (4.00)	1.21 (1.09-1.36)
Voice hoarseness			
0-6 months	65 (0.17)	227 (1.91)	10.93 (8.26-14.46)
6-24 months	189 (0.50)	95 (0.80)	1.56 (1.22-2.00)
6-12 months	69 (0.18)	42 (0.35)	1.90 (1.29-2.81)
12-18 months	68 (0.18)	29 (0.24)	1.30 (0.84-2.02)
18-24 months	63 (0.17)	31 (0.26)	1.50 (0.97-2.32)
Dyspnoea			
0-6 months	1,020 (2.71)	2,465 (20.74)	10.01 (9.19-10.90)
6-24 months	2,119 (5.63)	1,439 (12.10)	2.26 (2.10-2.43)
6-12 months	930 (2.47)	720 (6.06)	2.44 (2.20-2.70)
12-18 months	889 (2.36)	587 (4.94)	2.06 (1.85-2.30)
18-24 months	760 (2.02)	511 (4.30)	2.08 (1.85-2.34)
Weight loss			
0-6 months	125 (0.33)	629 (5.29)	17.17 (14.03-21.02)
6-24 months	297 (0.79)	239 (2.01)	2.45 (2.05-2.92)
6-12 months	105 (0.28)	118 (0.99)	3.54 (2.71-4.62)
12-18 months	92 (0.24)	89 (0.75)	2.85 (2.11-3.84)
18-24 months	111 (0.30)	52 (0.44)	1.36 (0.98-1.91)

Constipation			
0-6 months	623 (1.66)	762 (6.41)	4.05 (3.62-4.53)
6-24 months	1,384 (3.68)	626 (5.27)	1.38 (1.25-1.53)
6-12 months	588 (1.56)	285 (2.40)	1.47 (1.27-1.71)
12-18 months	539 (1.43)	239 (2.01)	1.33 (1.14-1.56)
18-24 months	486 (1.29)	208 (1.75)	1.27 (1.07-1.50)
Depression			
0-6 months	678 (1.80)	431 (3.63)	2.06 (1.82-2.34)
6-24 months	1,526 (4.06)	640 (5.38)	1.36 (1.24-1.50)
6-12 months	640 (1.70)	285 (2.40)	1.46 (1.26-1.68)
12-18 months	651 (1.73)	255 (2.15)	1.25 (1.08-1.45)
18-24 months	629 (1.67)	251 (2.11)	1.30 (1.12-1.51)
URTI			
0-6 months	731 (1.94)	417 (3.51)	1.86 (1.64-2.10)
6-24 months	1,975 (5.25)	735 (6.18)	1.19 (1.08-1.30)
6-12 months	708 (1.88)	284 (2.39)	1.27 (1.10-1.47)
12-18 months	771 (2.05)	270 (2.27)	1.09 (0.95-1.26)
18-24 months	665 (1.77)	258 (2.17)	1.24 (1.07-1.43)
LRTI			
0-6 months	529 (1.41)	926 (7.79)	6.40 (5.70-7.17)
6-24 months	1,267 (3.37)	835 (7.02)	2.22 (2.02-2.44)
6-12 months	493 (1.31)	348 (2.93)	2.34 (2.03-2.71)
12-18 months	465 (1.24)	306 (2.57)	2.13 (1.83-2.47)
18-24 months	446 (1.19)	280 (2.36)	1.99 (1.71-2.33)
Chest infections			
0-6 months	1,457 (3.87)	2,145 (18.04)	5.91 (5.48-6.38)
6-24 months	3,291 (8.75)	1,994 (16.77)	2.16 (2.03-2.31)
6-12 months	1,354 (3.60)	1,022 (8.60)	2.54 (2.33-2.78)
12-18 months	1,352 (3.60)	804 (6.76)	1.94 (1.77-2.13)
18-24 months	1,261 (3.35)	747 (6.28)	1.94 (1.76-2.14)
COPD			
0-6 months	660 (1.76)	1,183 (9.95)	6.31 (5.70-6.99)
6-24 months	1,234 (3.28)	1.403 (11.80)	4.01 (3.69-4.36)
6-12 months	659 (1.75)	748 (6.29)	3.82 (3.42-4.27)
12-18 months	595 (1.58)	670 (5.64)	3.73 (3.33-4.19)
18-24 months	504 (1.34)	576 (4.85)	3.76 (3.32-4.26)
Chest x-rays			
0-6 months	1153 (3.07)	5,990 (50.39)	39.48 (35.83-43.51)
6-24 months	2,752 (7.32)	1,870 (15.73)	2.41 (2.25-2.57)
6-12 months	1,042 (2.77)	893 (7.51)	2.87 (2.61-3.16)
12-18 months	984 (2.62)	682 (5.74)	2.24 (2.02-2.48)
18-24 months	940 (2.50)	564 (4.74)	1.93 (1.73-2.15)
Blood investigations			
0-6 months	12,923 (34.37)	6,967 (58.61)	2.88 (2.75-3.01)
6-24 months	20,047 (53.31)	7,071 (59.48)	1.24 (1.18-1.30)
6-12 months	12,042 (32.02)	4,338 (36.49)	1.17 (1.12-1.22)
12-18 months	11,221 (29.84)	3,985 (33.52)	1.13 (1.08-1.18)
18-24 months	10,488 (27.89)	3,723 (31.32)	1.12 (1.06-1.17)

In mutivariate analysis of all the socio-demographic and clinical records of patients associated with lung cancer in the 6-24 months before diagnosis (Table 4.3), voice hoarseness, constipation, depression and upper respiratory tract infections were found not to be associated with lung cancer in the 6-24 month period and were excluded from the final model. In conducting this analysis, the highest daily cigarette consumption ever recorded (after exclusion of records

made 6 months prior to diagnosis) was used as a proxy marker for patients' cigarette exposure.

Table 4.3 Multivariate modelling of the clinical features associated with lung cancer 6-24 months before diagnosis

	Univariate OR	p-value	Adjusted OR (95% CI)	p-value §
Smoked qty(highest)				
Current V heavy (40+/d)	21.97 (18.65-25.88)		18.02 (15.24-21.31)	
Current Heavy (20-39/d)	16.90 (15.44-18.49)		14.40 (13.13-15.79)	
Current Mod (10-19/d)	11.41 (10.37-12.54)	<0.001	10.32 (9.36-11.38)	
Current Light (1-9/d)	7.19 (6.36-8.13)		6.75 (5.95-7.65)	<0.001
Current Trivial (<1/d)	2.30 (1.00-5.27)		2.16 (0.93-4.99)	
Ex V heavy (40+/d)	8.56 (7.08-10.34)		6.67 (5.48-8.11)	
Ex Heavy (20-39/d)	9.70 (8.71-10.80)		7.70 (6.89-8.60)	
Ex Mod (10-19/d)	6.02 (5.38-6.74)		4.98 (4.43-5.59)	
Ex Light (1-9/d)	4.37 (3.80-5.02)		3.73 (3.23-4.30)	
Ex Trivial (<1/d)	1.73 (0.94-3.17)		1.61 (0.86-3.00)	
Non-smoker	1.00		1.00	
Smoker, no qty recorded	3.20 (2.95-3.47)		2.97 (2.74-3.23)	
Missing smoking records	2.41 (2.17-2.69)		2.52 (2.26-2.82)	
Townsend score				
5 (most deprived)	2.26 (2.08-2.44)		1.47 (1.34-1.61)	
4	1.88 (1.75-2.03)	<0.001	1.36 (1.25-1.48)	<0.001
3	1.48 (1.38-1.59)		1.25 (1.16-1.36)	
2	1.19 (1.11-1.28)		1.15 (1.06-1.24)	
1 (least deprived)	1.00		1.00	
9(no record)	1.65 (1.41-1.94		1.15 (0.96-1.39)	
Cough	1.95 (1.85-2.07)	<0.001	1.36 (1.26-1.45)	<0.001
Haemoptysis	6.82 (5.50-8.44)	<0.001	3.72 (2.90-4.77)	<0.001
Chest pain	1.46 (1.36-1.56)	<0.001	1.12 (1.04-1.22)	0.004
Dyspnoea	2.26 (2.10-2.43)	<0.001	1.20 (1.10-1.32)	<0.001
Weight loss	2.45 (2.05-2.92)	<0.001	1.60 (1.30-1.97)	<0.001
LRTI	2.22 (2.02-2.44)	<0.001	1.28 (1.15-1.43)	<0.001
Chest infections	2.16 (2.03-2.31)	<0.001	1.30 (1.20-1.40)	<0.001
COPD	4.01 (3.69-4.36)	<0.001	1.69 (1.53-1.87)	<0.001
Chest x-rays	2.41 (2.25-2.57)	<0.001	1.27 (1.27-1.50)	<0.001
Blood tests	1.24 (1.18-1.30)	<0.001	1.07 (1.01-1.13)	<0.001

§ P-values for binary variables were obtained using the Wald's test of significance. In variables with more than 2 categories, p-values were obtained from the likelihood ratio test

4.4.3 Discussion and conclusion

This study has identified the socio-demographic and clinical predictors of lung cancer in general practice up to two years before diagnosis and also identified the timing before diagnosis when patients' features can be used to predict a future diagnosis of lung cancer.

There was an increase in the frequency of general consultations, consultations for clinical symptoms of lung cancer and clinical investigations in cases up to two years before lung cancer diagnosis and this was most marked within the year before diagnosis. After excluding records made in the 6 months before lung cancer diagnosis, patients' socio-demographic and clinical features were found to be independently associated with lung cancer 6-24 months before diagnosis. The socio-demographic features associated with lung cancer were patients' smoking status, daily cigarette consumption and deprivation (measured using Townsend deprivation quintiles). Since the cases and controls in the dataset were matched on age and sex, the effect of age and sex could not be accounted for. Clinical features that were independently associated with lung cancer were cough, haemoptysis, chest/shoulder pain, dyspnoea, weight loss, lower respiratory tract infections, chest infections, Chronic Obstructive Pulmonary Disease (COPD), chest x-rays and blood investigations. Despite being predictive of lung cancer, the majority of symptom records were relatively uncommon in the records of cases.

As previously stated, this study was done to pilot the methods for the development of a lung cancer predictive score and results from the study will be used to inform decisions on the relevant time periods when clinical symptoms can be used to reliably develop a predictive score. In the next chapter, a similar but more detailed study will be conducted using a case-control study that has not been matched on age and sex which will enable the identification of lung

cancer predictors including age and sex. Results of that study will then be applied in developing a predictive score for lung cancer.

4.5 Is there variation in the smoking associated risk of lung cancer by deprivation?

4.5.1 Introduction

Previous studies as well as results from this thesis have shown an increase in lung cancer risk among individuals of lower socioeconomic status[92]. Smoking is strongly associated with lung cancer incidence and it is highly prevalent among individuals of lower socioeconomic status[235]. Evidence from a meta-analysis however shows that the socioeconomic differences in lung cancer incidence remains even after adjusting for the level of cigarette smoke consumption[90]. Studies have also shown that self-reported smoking only accounts for 15% to 50% of the socioeconomic variation in lung cancer risk[93 236 237]. Although the differential exposure to factors such as diet and occupational exposure are known to account for some of the socioeconomic differences in lung cancer incidence, a substantial part of the inequalities remain even after these have been adjusted for and they do not fully account for the difference in lung cancer risk[97]. Fidler et al.[99] demonstrated that at similar levels of reported daily cigarette consumption, the saliva cotinine levels among individuals with higher levels of deprivation were higher than the cotinine levels in less deprived individuals. Results of the study may be explained by possible misclassification of smoking status or a difference in smoking behaviour between individuals of different socioeconomic groups. It however suggests the possibility that individuals of different socioeconomic status may be exposed to different smoking-associated risks per cigarette smoked.

This study uses the matched case-control dataset developed in this chapter to test the hypothesis that for each stratum of smoking, the dose-related risk of lung cancer is higher in individuals of lower socioeconomic status compared to individuals of higher socioeconomic status.

4.5.2 Methods

Conditional logistic regression analyses were performed to estimate the odds ratios and 95% CI for lung cancer associated with socioeconomic status and smoking. The odds ratio for lung cancer by smoking was stratified by Townsend quintiles to assess whether the overall risk of lung cancer differed among smokers from different socioeconomic groups ; and this was also assessed in males and females separately. Interaction terms were used to assess for any interaction between smoking and socio-economic status. Statistical significance was assumed at $p<0.05$ using the Wald's test of significance.

4.5.3 Results

In all the Townsend quintiles, smoking prevalence was higher among lung cancer cases than controls. Also, daily cigarette consumption increased with increasing levels of deprivation. Table 4.4 shows the distribution of cases and controls in the different Townsend quintiles and by category of smoking.

Table 4.5 shows an increase in the odds ratio for lung cancer with higher daily cigarette consumption. Compared to individuals who had never smoked, the odds for lung cancer in individuals who smoked 10 to 19 cigarettes daily was 9.04 (95% CI 8.30-9.83) and this increased to 14.17 (95% CI 13.07-15.35) in individuals who smoked 20 or more cigarettes daily. After stratifying this analysis by Townsend quintiles, there remained an increase in the odds of lung

cancer with a higher number of cigarettes smoked daily across all Townsend quintiles. The odds ratio for lung cancer among smokers in the different Townsend quintiles did not show a significant trend of increasing lung cancer risk with increasing deprivation and the findings were similar in males and females (Table 4.6), however the overall risk of lung cancer among smokers was greater in females than males.

Further investigation of the odds ratios for lung cancer among never smokers in the different Townsend quintiles showed that among never smokers, the risk of lung cancer increased with increasing deprivation such that individuals from Townsend quintile 5 had a 60% increase in lung cancer risk compared to individuals in Townsend quintile 1 (odds ratio 1.60; 95% CI 1.16-2.20). The lung cancer odds ratio for trend with increasing deprivation among non-smokers was 1.08 (95% CI 1.01-1.14).

Table 4.4. The distribution of cases and controls in the Townsend quintiles and by smoking category

	Townsend 1 (n=10,772)		Townsend 2 (n=10,387)		Townsend 3 (n=9,800)		Townsend 4 (n=9,351)		Townsend 5 (n=7,260)	
Daily cigarettes smoked	Controls (%) n=8,735	Cases (%) n=2,037	Controls (%) n=8,187	Cases (%) n=2,200	Controls (%) n=7,420	Cases (%) n=2,380	Controls (%) n=6,742	Cases (%) n=2,609	Controls (%) n=5,064	Case (%) n=2,196
Heavy/very heavy	773 (8.9)	617 (30.3)	798 (9.8)	707 (32.1)	908 (12.2)	839 (35.3)	1,052 (15.6)	1,019 (39.1)	941 (18.6)	936 (42.6)
Moderate smoker	691 (7.9)	363 (17.8)	737 (9.0)	395 (18.0)	719 (9.7)	492 (20.7)	761 (11.3)	585 (22.4)	652 (12.9)	471 (21.5)
Trivial/light smoker	514 (5.9)	197 (9.7)	489 (6.0)	197 (8.9)	486 (6.6)	191 (8.0)	438 (6.5)	213 (8.2)	352 (7.0)	166 (7.6)
Non-smoker	3,916 (44.8)	307 (15.1)	3,573 (43.6)	280 (12.7)	2,903 (39.1)	232 (9.8)	2,303 (34.2)	222 (8.5)	1,522 (30.1)	170 (7.7)
Smoker (no quantity)	2,037 (23.3)	440 (21.6)	1,886 (23.0)	466 (21.2)	1,651 (22.3)	464 (19.5)	1,521 (22.6)	412 (15.8)	1,052 (20.8)	315 (14.3)
No smoking records	804 (9.2)	113 (5.5)	704 (8.6)	155 (7.1)	753 (10.2)	162 (6.8)	667 (9.9)	158 (6.1)	545 (10.8)	138 (6.3)

Table 4.5. The odds ratio for lung cancer by smoking category and stratified by Townsend quintiles

Daily cigarettes smoked	Overall OR (n=49,493)*	Townsend 1 (n=10,772)	Townsend 2 (n=10,387)	Townsend 3 (n=9,800)	Townsend 4 (n=9,351)	Townsend 5 (n=7,260)
Heavy/very heavy	14.17 (13.07-15.35)	13.49 (10.31-17.66)	16.78 (12.39-22.74)	20.65 (14.83-28.77)	13.94 (10.40-18.67)	10.86 (7.97-14.80)
Moderate smoker	9.04 (8.30-9.83)	6.99 (5.34-9.15)	9.81 (7.14-13.49)	12.75 (9.12-17.82)	8.81 (6.49-11.95)	7.19 (5.21-9.94)
Trivial/light smoker	5.63 (5.09-6.22)	4.40 (3.25-5.96)	6.25 (4.37-8.94)	5.89 (4.04-8.58)	5.28 (3.66-7.64)	4.54 (3.08-6.68)
Non-smoker	1.00	1.00	1.00	1.00	1.00	1.00
Smoker (no quantity)	3.23 (2.98-3.50)	2.61 (2.07-3.29)	3.29 (2.50-4.32)	4.48 (3.29-6.10)	3.03 (2.27-4.04)	2.78 (2.00-3.87)
No smoking records	2.38 (2.14-2.65)	1.55 (1.11-2.18)	2.54 (1.78-3.61)	2.66 (1.79-3.95)	2.17 (1.49-3.15)	1.85 (1.23-2.79)

* includes those with missing records on Townsend quintile

Table 4.6. The odds ratio for lung cancer by smoking category in males and females, stratified by Townsend quintiles

			Males			
Daily cigarettes smoked	**Overall OR** (n=28,991)*	**Townsend 1** (n=6,591)	**Townsend 2** (n=6,169)	**Townsend 3** (n=5,714)	**Townsend 4** (n=5,337)	**Townsend 5** (n=4,095)
Heavy/very heavy	11.24 (10.11-12.51)	12.11 (8.58-17.09)	12.06 (8.27-17.57)	15.86 (10.33-24.37)	9.30 (6.39-13.51)	8.04 (5.27-12.28)
Moderate smoker	7.62 (6.78-8.55)	6.38 (4.47-9.10)	8.41 (5.60-12.64)	9.62 (6.19-14.94)	6.57 (4.35-9.93)	4.40 (2.80-6.92)
Trivial/light smoker	5.00 (4.38-5.70)	5.81 (3.95-8.56)	5.36 (3.41-8.43)	4.90 (2.98-8.05)	3.91 (2.43-6.27)	3.45 (2.01-5.94)
Non-smoker	1.00	1.00	1.00	1.00	1.00	1.00
Smoker (no quantity)	3.09 (2.78-3.44)	3.15 (2.35-4.22)	2.78 (1.99-3.88)	4.65 (3.09-6.99)	2.28 (1.57-3.31)	2.48 (1.58-3.89)
No smoking records	2.20 (1.91-2.53)	1.64 (1.07-2.50)	2.32 (1.48-3.64)	2.52 (1.54-4.15)	1.64 (1.00-2.67)	1.20 (0.70-2.06)
			Females			
	n=20,502*	n=4,181	n=4,218	n=4,086	n=4,014	n=3,165
Heavy/very heavy	19.60 (5.41-7.42)	19.03 (11.92-30.81)	30.32 (17.63-52.14)	32.08 (18.47-55.71)	23.56 (14.59-39.06)	14.49 (9.12-23.03)
Moderate smoker	10.91 (9.63-12.37)	8.99 (5.78-14.00)	11.42 (6.78-19.22)	19.48 (11.35-33.43)	12.11 (7.60-19.30)	11.58 (7.20-18.60)
Trivial/light smoker	6.34 (5.41-7.42)	2.61 (1.58-4.31)	7.93 (4.35-14.46)	7.78 (4.28-14.11)	7.21 (3.99-13.00)	5.70 (3.25-10.00)
Non-smoker	1.00	1.00	1.00	1.00	1.00	1.00
Smoker (no quantity)	3.07 (2.70-3.50)	1.69 (1.14-2.51)	4.22 (2.57-6.94)	3.75 (2.31-6.09)	4.14 (2.60-6.57)	2.75 (1.67-4.52)
No smoking records	2.47 (2.08-2.94)	1.47 (0.83-2.62)	2.72 (1.53-4.83)	2.55 (1.31-4.98)	2.88 (1.59-5.19)	2.99 (1.56-5.72)

* includes those with missing records on Townsend quintile

4.5.4 Discussion and conclusion

In this study, there was no evidence to support the hypothesis that the risk of lung cancer associated with smoking increases with increasing deprivation. At increasing levels of cigarette consumption, there was an increase in the risk of lung cancer and this risk was similar across the Townsend quintiles. Cigarette consumption was however higher among individuals from more deprived Townsend quintiles and a higher proportion of cases who smoked were from more deprived quintiles. The finding of the lack of a difference in the risk of lung cancer among smokers from different socioeconomic groups was consistent in both males and females. However, the increased risk of lung cancer associated with smoking was higher in females than males.

A major strength of this study is the large size of the THIN dataset which provides sufficient power to the study. By using patients' highest ever smoking record up to 6 months before lung cancer diagnosis, any effect due to a change in cigarette consumption in the months leading up to lung cancer diagnosis has been minimised in this study.

The reliance on patients' reported smoking consumption may introduce bias due to a possible underestimation of smoking status by certain patients. Also, the misclassification of smoking status by GPs may introduce residual confounding into the study, although any effect due to misclassification would affect the cases and controls similarly and should therefore not make a difference to the study results. Information on risk factors such as occupational exposure, diet and alcohol consumption which may be higher among and therefore increase the risk of lung cancer in individuals of lower socioeconomic groups, were not available in THIN database and could not be adjusted for in the study. Nonetheless, the finding of an increase in the baseline risk of lung cancer among non-smokers who were deprived compared to non-deprived non-smokers

suggests that these factors may marginally increase the risk of lung cancer among deprived individuals and supports findings from other studies that factors such as diet, occupational, environmental exposures and other lifestyle factors contribute to the association between socioeconomic status and lung cancer risk. The increase in lung cancer risk among female smokers compared to non-smokers is an interesting finding which warrants further exploration and this will be explored in another study in this chapter.

In conclusion, the findings of this study fail to provide support for the hypothesis that the risk of lung cancer is higher in more deprived smokers compared to less deprived smokers with similar levels of reported daily cigarette consumption; and suggests contrary to previous studies[236 237], that most of the socioeconomic difference in lung cancer risk are due to smoking. The socioeconomic gradient in lung cancer incidence is therefore driven by the greater smoking prevalence among people of lower socioeconomic status and to tackle these inequalities, smoking cessation programs targeted to socioeconomically deprived communities need to be intensified.

4.6 Is there an increase in smoking-associated risk of lung cancer in depressed compared to non-depressed smokers?

4.6.1 Introduction

Despite the fact that cigarette smoking is the most important risk factor for lung cancer[25 38], certain host factors increase the susceptibility of people to start and continue smoking, to smoke more heavily and to develop lung cancer[238]. Smoking prevalence is higher among individuals with depression[239-241], perhaps in part because nicotine from cigarettes has been reported to provide temporary relief from the symptoms of depression[242 243]. Compared to smokers without reported depression, smokers with depression also have a higher risk of being nicotine dependent[244], they are less likely to quit smoking[245] and they have a greater likelihood of smoking relapse[246]. Depression may also cause an alteration in the body's immune system and consequently increases the risk of immune-related conditions such as cancer[104 247].

It has been suggested that depression increases the risk of several cancers including lung cancer, yet evidence from the few studies that have examined this association is not consistent. In a Finnish cohort study, depression was found to modify the effect of smoking on lung cancer risk in men such that the relative risk of lung cancer among smokers compared with non-smokers was considerably higher for those with elevated depressiveness scores (19.67; 95% CI 2.57-150.7) than for men at normal depressiveness scores (3.38; 95% CI 1.09-10.52)[107]. In another prospective study in the United States, depression was positively associated with smoking-related cancers in individuals who smoked at least 15 cigarettes daily[108]. A prospective cohort study of persons aged 71 years and older in Massachusetts, USA, found an increase in the risk of several cancers including lung cancer among chronically depressed individuals regardless of their smoking status [105].

Using the thesis matched case-control dataset of patients in THIN, this study examined the association between depression and subsequent lung cancer risk in UK general practice patients. In doing this, the association between smoking and lung cancer was stratified by depression to determine whether people with depression are more at risk from the adverse effects of smoking.

4.6.2 Methods

Records of depression up to one year before the lung cancer index date were obtained from the cases and controls. Depression records made in the year preceding diagnosis were excluded to ensure that records related to patients' imminent diagnosis of lung cancer were not included in the analyses. Also, the highest recorded daily cigarette consumption for patients were obtained (detailed in section 4.3). Conditional logistic regression analyses were used to estimate the odds ratios and 95% confidence intervals (CIs) for lung cancer associated with depression and smoking. The odds ratio for lung cancer with depression was also obtained after adjusting for the effects of smoking. To estimate the increase in lung cancer risk among smokers with depression and those without depression, the analysis of the association between lung cancer and smoking were stratified by depression. Interaction terms were used to assess for any interaction between smoking and depression. Statistical significance was assumed at 0.05 using the Wald's test of significance.

4.6.3 Results

Records of depression were present in the general practice notes of 20.9% of cases and 17.1% of controls prior to one year before the cases' lung cancer index date. Univariate analysis of depression and lung cancer showed that depression was associated with a 30% increased odds of lung cancer (OR 1.30;

95% CI 1.24 - 1.38)(Table 4.7). Smoking was also associated with an increase in the odds ratio for lung cancer and the odds increased with an increase in daily cigarette consumption. On adjusting the association between depression and lung cancer by smoking, the odds ratio for lung cancer among people with depression decreased to 1.06 (95% CI 0.99-1.12).

Table 4.8 shows the association between cigarette smoking and lung cancer stratified by depression. Compared to individuals with no record of depression, individuals with a history of depression were more likely to smoke and to smoke more heavily, with a higher proportion of them being moderate and heavy/very heavy smokers. The increase in the odds ratio for lung cancer with higher daily cigarette consumption was similar in both depressed and non-depressed groups of patients. There was no effect modification by a diagnosis of depression on the association between smoking habit and lung cancer risk.

To ensure that results from the stratified matched analyses using conditional logistic regression were not distorted due to the large number of missing depression values which would have resulted in some dropped cases or controls, these analyses were repeated by breaking the matching and using unconditional logistic regression adjusted for age and sex. The results from these analyses were very similar to those of the matched analyses.

Table 4.7. Frequency of depression and smoking prevalence among cases and controls

Variable	Controls n(%) n=37,605	Cases n(%) n=11,888	Total n(%)	Unadjusted odds ratio for lung cancer	95% confidence intervals
Depression					
History of depression	6,436 (17.1)	2,487 (20.9)	8,923 (18.0)	1.30	1.24 - 1.38
No history of depression	31,169 (82.9)	9,401 (79.1)	40,570 (82.0)	1.00	
Smoking					
Heavy/very heavy smoker	4,679 (12.4)	4,292 (36.1)	8,971 (18.1)	14.17	13.07 - 15.35
Moderate smoker	3,720 (9.9)	2,410 (20.3)	6,130 (12.4)	9.04	8.30 - 9.83
Trivial/light smoker	2,377 (6.3)	1,003 (8.4)	3,380 (6.8)	5.63	5.09 - 6.22
Non smoker	14,729 (37.2)	1,260 (10.6)	15,989 (32.3)	1.00	-
Smoker, no record of quantity	8,460 (22.5)	2,169 (18.3)	10,629 (21.5)	3.23	2.98 - 3.50
Missing smoking records	3,640 (9.7)	754 (6.3)	4,394 (8.9)	2.38	2.14 - 2.65

Odds ratio for lung cancer with depression after adjusting for smoking was 1.06 (95% CI 0.99-1.12)

Table 4.8. Association between smoking and lung cancer, stratified by depression

	No history of depression			**History of depression**		
Smoking status	controls n(%) n=31,169 (100)	cases n(%) n=9,401 (100)	Odds ratio for lung cancer (95% CI)	controls (%) n=6,436 (100)	cases (%) n=2,487 (100)	Odds ratios for lung cancer (95% CI)
Heavy/very heavy smoker	3,524 (11.3)	3,179 (33.8)	13.5 (12.3-14.8)	1,155 (18.0)	1,113 (44.8)	14.8 (11.0-20.0)
Moderate smoker	2,939 (9.4)	1,848 (19.7)	8.8 (8.0-9.7)	781 (12.1)	562 (22.6)	9.2 (6.8-12.6)
Trivial/light smoker	1,942 (6.2)	816 (8.7)	5.7 (5.1-6.4)	435 (6.8)	187 (7.5)	4.6 (3.1-6.9)
Non smoker	12,393 (39.8)	1,051 (11.2)	1.00	2,336 (36.3)	209 (8.4)	1.00
Smoker, no record of qty	7,043 (22.6)	1,820 (19.4)	3.3 (3.0-3.7)	1,417 (22.0)	349 (14.0)	2.8 (2.0-3.8)
Missing smoking records	3,328 (10.7)	687 (7.3)	2.3 (2.1-2.6)	312 (4.9)	67 (2.7)	1.7 (0.9-3.0)

4.6.4 Discussion and conclusion

In this study, patients with a history of depression were found to have a 30% increased risk of lung cancer compared with patients with no history of depression and this increased lung cancer risk was explained by cigarette smoking. Cigarette smoking was higher among patients with a recorded history of depression compared to those with no history of depression, and they were more likely to smoke more heavily. On stratified analysis, a history of depression did not appear to make people more vulnerable to the carcinogenic effects of smoking.

As previously mentioned in chapter 2, a strength of THIN database - the data source for this study, is its large size, providing data on a vast number of patients and enabling the study of associations between different exposures and rare outcomes such as lung cancer. Records of depression and smoking in the database were collected during routine consultation in general practice and the results are therefore applicable to UK general practices. By matching cases and controls in our study by age, sex and general practice, any confounding due to these variables were controlled for during the design stage of the study.

The study is limited by the fact that the diagnosis of depression was not based on standardised psychiatric criteria but on the assessment of GPs. Patients were noted to have a history of depression when they had records of a previous diagnosis of depression in their general practice notes up to one year prior to the lung cancer index date. Although this is not an ideal way to assess clinical depression, findings from these analyses consequently reflect the association between these assessments of depression in general practice and subsequent lung cancer incidence. Patients with lung cancer are known to commonly have psychological distress and depressive symptoms which is related with their functional limitations and symptoms[248] and this can lead to the possibility of reverse causation in the association between depression and lung cancer. To

minimise any effect due to reverse causality in this study, records of depression that were made within the year before lung cancer diagnosis were excluded. Since evidence from previous studies show that the majority of patients with lung cancer have symptoms for a median of 12 months before diagnosis[112], it is unlikely that the diagnosis of depression among the cases in this study were related to their impending lung cancer diagnosis.

Previous studies have shown an increase in lung cancer risk among depressed compared to non-depressed smokers[107 108] and it has been suggested that there may be differences in the smoking behaviour such as much deeper inhalation or smoking more of the cigarette in depressed compared to non-depressed smokers[249]. It has also been argued that depression modifies the effect of smoking on lung cancer[107]. In this study however, there was no difference in smoking-associated risk of lung cancer between depressed and non-depressed individuals who smoked similar quantities of cigarettes daily, suggesting that the increased risk of lung cancer observed among depressed individuals is mostly explained by their higher prevalence of cigarette smoking and more heavy smoking.

The small non-significant 6% excess risk of lung cancer with depression which remained after adjusting for smoking may partly be due to residual confounding due to possible misclassification of smoking in the general practice records, or passive smoking. It has been proposed that depression alters the body's immune functions and suppresses cellular immunity through activation of the Hypothalamic-Pituitary-Adrenal axis and the ensuing abnormal secretion of adrenal steroids[104 250]. This impaired cellular immunity has however been noted to promote the development and progression of certain cancers associated with viruses but there is no known evidence as yet, to show that it increases the risk of lung cancer. Certain behavioural factors such as low levels of physical activity, poor dietary habits and high alcohol consumption are also known to increase the

risk of lung cancer[72 79 80], and the less healthy behaviours of depressed people may be another mechanism by which lung cancer risk is slightly increased in these individuals.

In conclusion therefore, this study found an increase in lung cancer risk among general practice patients with a history of depression and this was largely explained by smoking. Smoking increased the risk of lung cancer in depressed and non-depressed individuals and there was no evidence to support a significantly higher smoking-associated risk of lung cancer in depressed compared to non-depressed individuals. The possibility that an interplay of genetic factors and other behavioural risk factors such as high alcohol consumption and poor dietary intake may marginally increase the risk of lung cancer among individuals with depression cannot be excluded. However given the fact that depression and other mental health conditions are associated with a higher prevalence of cigarette smoking and as a consequence, lung cancer, it is important that smoking-cessation interventions are incorporated into the NICE guidelines for the management of patients with depression[251] in order to prevent lung cancer and other chronic conditions in the long term.

4.7 The association between smoking quantity and lung cancer in men and women

This section summarizes the results of a study[252] which investigated whether the risk of lung cancer differs between men and women with the same recorded quantity of cigarettes smoked. The study was conducted as part of a PhD project by Dr Helen Powell, a clinical fellow in the Division of Epidemiology and Public Health. The study used the matched case-control dataset developed by Barbara Iyen-Omofoman and which has been described in this thesis chapter. The initial

data management and extraction of the variables for the study were done by Barbara Iyen-Omofoman while Helen Powell carried out the data organisation and performed the statistical analyses.

4.7.1 Study summary

Previous evidence had shown that women who smoke have a 25% greater risk of coronary heart disease than male smokers[253] and even though an examination of this relationship in lung cancer had shown conflicting results[87 89 254 255], no study had assessed the effect in a UK population. The study also tested the hypothesis that if women are at higher risk of smoking-related lung cancer, it may be because they have smaller lung volumes than men.

Using conditional logistic regression, odds ratios for lung cancer were calculated according to the highest recorded daily cigarette consumption in men and women separately. Results showed that in women, there was a 19-fold increase in the risk of lung cancer (odds ratio 19.10; 95% CI 16.98-21.49) in heavy smokers compared to never-smokers. This was more than for men smoking the same quantity (odds ratio 12.81; 95% CI 11.52-14.24). A test for interaction showed strong evidence of a difference in effect of cigarette quantity smoked on lung cancer between men and women (interaction $p<0.001$) and this effect remained even after adjusting for height (a proxy marker for lung volume).

Based on the results of this study, it was concluded that moderate and heavy smoking carry a higher risk of lung cancer in women than men and this difference is not explained by a difference in their lung volumes. Extrapolating risk estimates for lung cancer in men to women will therefore underestimate the adverse impact of smoking in women.

Chapter 5. The use of an unmatched case-control dataset to identify the socio-demographic and early clinical features predictive of lung cancer in general practice

In the previous chapter, the study used a case-control dataset matched on age, sex and general practice to pilot the methods to identify the independent predictors of lung cancer and determine the timing of clinical features before lung cancer diagnosis. Also using the same dataset, studies were conducted to investigate variations in the smoking-associated risk of lung cancer in certain groups of general practice patients while controlling for the confounding effects of age, sex and variability in general practice recording. The next phase of this study entails the identification of lung cancer predictors that can be used to develop a predictive score for lung cancer. To develop a predictive score that is robust and widely applicable, the effects of age and sex have to be accounted for. This chapter therefore uses a case-control dataset that has not been matched on age or sex, to identify the socio-demographic and early clinical features predictive of lung cancer and that can be used to develop a predictive score for lung cancer.

5.1 introduction

Most lung cancer patients experience symptoms before diagnosis[114] and in a study of recently diagnosed patients, symptoms were recalled starting between 4 months and 2 years before diagnosis[112]. In the UK where the GP is the gatekeeper to specialised health care, most patients present with symptoms to their GP before the diagnosis of lung cancer is made[112 113]. A case-control study of patients from 21 general practices in Exeter, UK[113] showed that GP records of haemoptysis, dyspnoea, abnormal spirometry and smoking were independently

associated with lung cancer up to 180 days before diagnosis. Although the UK NICE referral guidelines[147] were developed to facilitate urgent referral of suspected lung cancer cases, the evidence base for this has been questioned[112] [134] [142]. Because many lung cancer symptoms are non-specific, GPs face a difficult challenge in deciding which patients merit investigation. A key step is to estimate the risk of lung cancer by taking into account a combination of socio-demographic features and clinical symptoms.

This study aims to use the thesis dataset of cases and unmatched controls to identify the pattern and frequency of early pre-diagnostic symptoms, clinical investigations and patients' socio-demographic factors that are independently associated with lung cancer.

5.2 Methods

5.2.1 Cases and controls

The cases included in this matched case-control dataset were the incident cases of lung cancer derived in Chapter 2 (section 2.3). The eligibility criteria for case selection have also been detailed in section 4.1.1. Similar to the matched case-control study in chapter 4, the 59 lung cancer cases less than 40 years of age (0.49% of cases) were excluded from the analyses in this chapter. In total, there were 12,076 eligible cases for this study.

For each case, 10 randomly selected controls were selected. The controls were selected and assigned to cases using the following criteria

- Controls had to be registered in the same general practice as the case
- At least 40 years or older on the date of diagnosis of the case
- Alive and contributing to THIN on the lung cancer index date of the case

- Have at least 1 year of active data prior to the case index date

Ten eligible controls were randomly assigned to each case and when there were less than 10 eligible controls for a case, all eligible controls were assigned.

5.2.2 Socio-demographic and clinical features

The socio-demographic information analysed were: age, sex, Townsend deprivation quintiles (measure of socio-economic status) and smoking history. Definition of the variables - Age, sex, Townsend deprivation quintiles and smoking, as well as the data extraction process for these variables were as described in the previous chapter for the matched case-control dataset (section 4.3). In the dataset analysed in the previous chapter, age and sex were used as matching variables. In the case-control dataset analysed in this chapter, patients' age was categorised into 5-year age bands and in addition to sex, was included in the analyses.

Based on the highest ever recorded number of cigarettes smoked daily, the smoking records of current or ex-smokers were further categorised as trivial (less than 1 cigarette daily), light (1-9 cigarettes daily), moderate (10-19 cigarettes daily), heavy (20-39 cigarettes daily) or very heavy (40+ cigarettes daily). As detailed in the previous chapter, current or ex-smokers who had no records of daily cigarette consumption were recorded as such (smoker with no recorded quantity) and patients who had no recorded smoking information and who were not known to be non-smokers were included in a separate category (missing smoking records).

All consultations made by the cases and controls in their registered general practices were retrieved from the database. Details of the symptoms analysed, chest x-rays and blood investigations have been described in the previous

chapter (section 4.3 4). In extracting and categorising records of blood tests for this study, blood investigations were classified based on the outcome of the tests into: normal result, abnormal result, test done with no recorded result and no record of blood tests. In addition, the frequency of general practice consultations for symptoms and diagnoses other than those already assessed in the study were retrieved from patients' records.

The Read codes used to extract patients' records of symptoms, chest x-rays and blood tests are listed in Appendix I.

5.2.3 Timing of clinical records

Similar to the method described in the study in the previous chapter, all symptoms, diagnoses and investigations over the 2-year period before lung cancer diagnosis (or the diagnosis date of the matched case, for controls) were retrieved from patients' records. Since a chest x-ray is the initial investigation for suspected lung cancer[147], the timing of chest x-rays prior to lung cancer diagnosis in cases and prior to the pseudo-date in controls were determined and compared, as a means of determining the time period before diagnosis when GPs started preliminary investigations for suspected lung cancer in cases. There was a steep increase in the chest x-ray frequency in cases (but not controls) within the 4 months prior to diagnosis, so all symptoms, blood tests and other consultations recorded within this period were excluded.

5.2.4 Statistical analysis

Univariate logistic regression models were used to calculate the relative odds of lung cancer by socio-demographic factors (age, sex and Townsend deprivation quintiles), smoking, symptoms, diagnoses, blood tests and number of consultations in the 2 years before diagnosis. These analyses were done separately for records made in the 4-12 and the 13-24 month periods prior to diagnosis. Multivariate modelling was done using only variables that were associated with lung cancer in univariate analysis, using a significance cut-off level of $p<0.05$. Variables that were not statistically significant in the multivariate analysis were removed from the model and those that previously showed no association with lung cancer in the univariate model were re-checked for significance in the final model.

5.3 Results

5.3.1 Population socio-demographic characteristics

Of the 12,076 cases eligible for this study, 12,073 cases were assigned 10 controls each, 2 cases did not have any eligible controls and were excluded, and the remaining case had only 1 eligible control, giving a total of 132,805 patients in the study population which comprised of 12,074 cases and 120,731 controls. Compared to controls, people with lung cancer were more likely to be male, be older, live in households located in more deprived areas and they were more likely to have a current or ex smoking history (Table 5.1)

Table 5.1. Socio-demographic characteristics and smoking status of lung cancer cases and controls

	Case n(%) n =12,074	Control n(%) n = 120,731	Unadjusted odds ratio for lung cancer (95% CI)
Age at diagnosis (years)			
>80	2,639 (21.86)	10,797 (8.94)	48.80 (39.72-59.97)
75-80	2,305 (19.09)	8,191 (6.78)	56.19 (45.69-69.10)
70-75	2,212 (18.32)	9,940 (8.23)	44.43 (36.13-54.64)
65-70	1,750 (14.49)	11,201 (9.28)	31.20 (25.34-38.40)
60-65	1,488 (12.32)	13,475 (11.16)	22.05 (17.90-27.16)
55-60	896 (7.42)	15,439 (12.79)	11.59 (9.37-14.33)
50-55	469 (3.88)	15,963 (13.22)	5.87 (4.70-7.32)
45-50	220 (1.82)	16,756 (13.88)	2.62 (2.06-3.34)
40-45	95 (0.79)	18,969 (15.71)	1.00
Sex			
Male	7,154 (59.25)	58,034 (48.07)	1.57 (1.51-1.63)
Female	4,920 (40.75)	62,697 (51.93)	1.00
Townsend deprivation quintile			
5 (most deprived)	2,234 (18.50)	15,997 (13.25)	1.94 (1.82-2.07)
4	2,640 (21.87)	21,071 (17.45)	1.74 (1.64-1.85)
3	2,421 (20.05)	23,791 (19.71)	1.41 (1.33-1.50)
2	2,236 (18.52)	26,540 (21.98)	1.17 (1.10-1.25)
1 (least deprived)	2,064 (17.09)	28,681 (23.76)	1.00
Missing Townsend records	479 (3.97)	4,651 (3.85)	1.43 (1.29-1.59)
Smoking status and qty			
Current V heavy (40+/d)	471 (3.90)	1,466 (1.21)	12.52 (11.14-14.09)
Current Heavy (20-39/d)	2,589 (21.44)	10,928 (9.05)	9.24 (8.61-9.90)
Current Mod (10-19/d)	1,665 (13.79)	8,247 (6.83)	7.87 (7.29-8.49)
Current Light (1-9/d)	607 (5.03)	3,765 (3.12)	6.28 (5.68-6.96)
Current Trivial (<1/d)	7 (0.06)	144 (0.12)	1.89 (0.89-4.05)
Current, no qty recorded	439 (3.64)	4,495 (3.72)	3.81 (3.40-4.26
Ex V Heavy (40+/d)	221 (1.83)	841 (0.70)	10.24 (8.75-12.00)
Ex Heavy (20-39/d)	1,043 (8.64)	4,258 (3.53)	9.55 (8.75-10.42)
Ex Mod (10-19/d)	777 (6.44)	4,394 (3.64)	6.89 (6.27-7.57)
Ex Light (1-9/d)	399 (3.30)	2,837 (2.35)	5.48 (4.87-6.17)
Ex Trivial (<1/d)	13 (0.11)	289 (0.24)	1.75 (1.00-3.06)
Ex, no qty recorded	1,780 (14.74)	16,027 (13.27)	4.33 (4.02-4.66)
Non smoker	1,300 (10.77)	50,676 (41.97)	1.00
Missing smoking records	763 (6.32)	12,364 (10.24)	2.41 (2.20-2.64)

5.3.2 Duration of registration in the general practices

The average period of follow-up in the general practices prior to the lung cancer index date (defined in the cases as the date of lung cancer diagnosis and defined in controls as the date of lung cancer diagnosis in the matching case) was similar in cases and controls. The median follow-up for the cases was 9.5 years (inter-quartile range 5.5 years to 13.5 years) and the median follow-up for controls was 9.1 years (inter-quartile range 5.2 years to 13.2 years).

5.3.3 Overall consultation by cases and controls

For the entire duration of being registered in the general practices, the median number of consultations per case was 421 and the median number of consultations per control was 192. In the two years before lung cancer diagnosis in cases, the median number of consultations by the cases and controls were 170 and 64 consultations respectively. Plots of the consultation pattern in cases and controls over the 5-year and 2-year periods prior to lung cancer diagnosis (Figure 5.1 and Figure 5.2) showed a similar consultation pattern in cases and controls up to the year before diagnosis, when the consultation frequency in cases increased considerably.

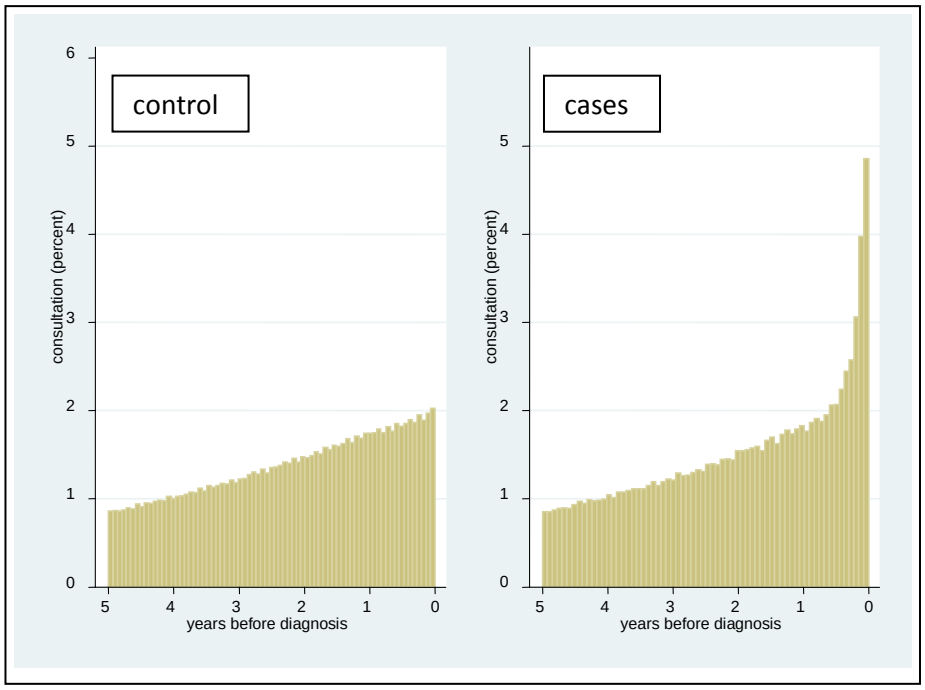


Figure 5.1: Plot of general consultation by controls and lung cancer cases, 5 years before lung cancer diagnosis
(the height of the bars are scaled so that the sum of their height equals 100)

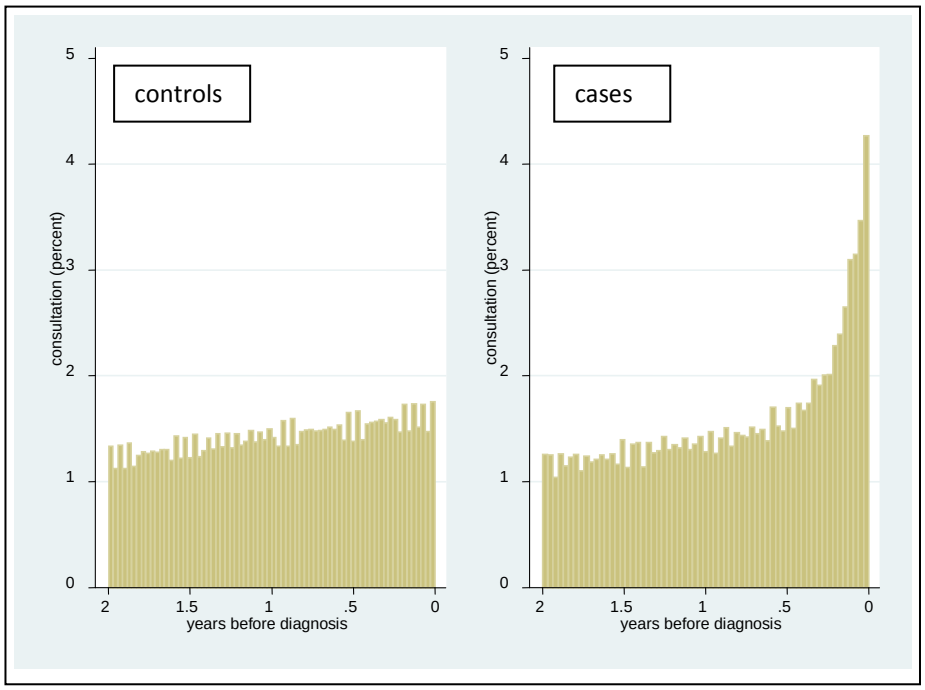


Figure 5.2: Plot of general consultation by controls and lung cancer cases, 2 years before lung cancer diagnosis
(the height of the bars are scaled so that the sum of their height equals 100)

5.3.4 Timing of chest x-rays prior to lung cancer diagnosis

A plot of the frequency of chest x-ray investigations in cases and controls before lung cancer diagnosis showed a fairly similar chest x-ray frequency in both groups of patients up to the 4th month preceding lung cancer diagnosis (Figure 5.3). During the 4 months before lung cancer diagnosis, there was a steep increase for cases implying that investigations for lung cancer were initiated by GPs at this time. Based on this finding, it was considered logical to exclude all symptoms, blood tests and other consultations recorded within the 4 month period from further analyses, so that the early-stage factors associated with lung cancer could be determined.

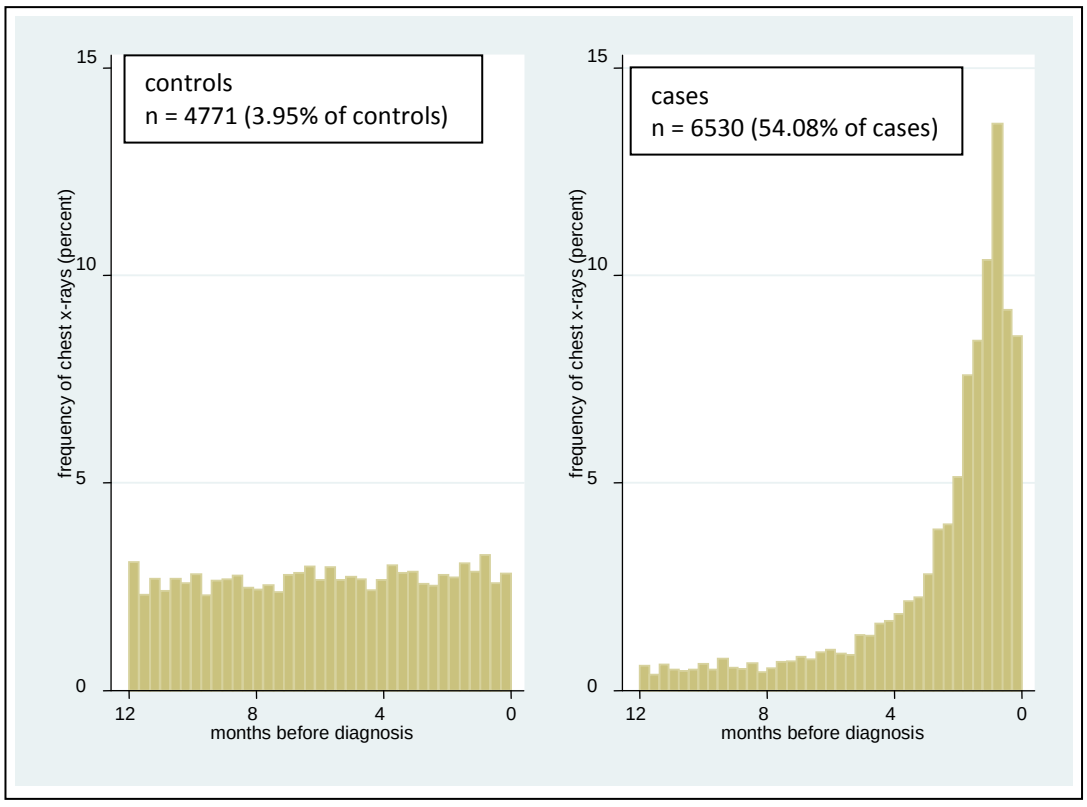


Figure 5.3. Plots showing the frequency distribution of chest x-rays in general practice prior to the diagnosis of lung cancer
(the height of the bars are scaled so that the sum of their height equals 100)
Odds ratio (95% CI) for chest x-ray within the 12 months among cases compared to controls was 28.63 (27.34-29.98); P<0.001

5.3.5 Clinical features associated with lung cancer

Analysis of the symptoms, diagnoses, blood tests and number of consultations within the 4-12 and 13-24 month periods preceding diagnosis (Table 5.2 & Table 5.3) showed that the symptoms with the highest frequency among cases were cough, non-specific chest infections, dyspnoea, chest pain and COPD. In the 4-12 month period before diagnosis, these symptoms were recorded among 16%, 12%, 9%, 8% and 8% of cases respectively compared to 6%, 4%, 2%, 4% and 1% of controls respectively. Although haemoptysis records were made for only 2% of cases within the 4-12 month period, the odds ratio for lung cancer among people who had haemoptysis within this period was 20.15 (95% CI 16.24-25.01).

Compared to the controls, cases consulted their GPs for other symptoms more often before diagnosis. Using fewer than 10 consultations as a reference value, the odds ratio for cases to consult their GPs 21 times or more was 3.56 (95% CI 3.41-3.73) in the 13-24 month period and 4.45 (95% CI 4.24-4.68) in the 4-12 month period. Depression was a commonly recorded symptom in the records of patients with lung cancer however, depression records within the 24 months before diagnosis was found not to be associated with lung cancer. There were also more blood investigations among cases than controls within the 4-12 and 13-24 month periods before diagnosis, with an increase in the number of normal and abnormal test results. The odds ratios for lung cancer were greater with all the symptoms recorded within the 4-12 month period than the 13-24 month period.

Table 5.2 : Symptoms, blood investigations and number of general practice consultations recorded among cases and controls within the 4 to 12 month period prior to lung cancer diagnosis

Variable in GP record 4-12 months before lung cancer diagnosis	Cases n(%) N=12,074	Controls n(%) N=120,731	Unadjusted OR for lung cancer	95% CI	p-value §
Cough	1,938 (16.05)	7,088 (5.87)	3.07	2.90-3.24	<0.001
Haemoptysis	247 (2.05)	125 (0.10)	20.15	16.24-25.01	<0.001
Chest/shoulder pain	1,002 (8.30)	4,880 (4.04)	2.15	2.00-2.31	<0.001
Voice hoarseness	66 (0.55)	219 (0.18)	3.02	2.30-3.99	<0.001
Dyspnoea	1,091 (9.04)	2,479 (2.05)	4.74	4.40-5.10	<0.001
Weight loss	197 (1.63)	323 (0.27)	6.18	5.17-7.39	<0.001
Constipation	423 (3.50)	1,469 (1.22)	2.95	2.64-3.29	<0.001
Depressive disorders	365 (3.02)	3,365 (2.79)	1.09	0.97-1.21	0.135
URTI	426 (3.53)	3,082 (2.55)	1.40	1.26-1.55	<0.001
LRTI	516 (4.27)	1,585 (1.31)	3.36	3.03-3.71	<0.001
Non-specific chest infections	1,398 (11.58)	4,350 (3.60)	3.50	3.29-3.73	<0.001
COPD	978 (8.10)	1,349 (1.12)	7.80	7.17-8.49	<0.001
Outcome of blood tests					
No blood test record	6,406 (53.06)	84,997 (70.40)	1.00		
Test without results	5,431 (44.98)	34,295 (28.41)	2.10	2.02-2.18	
Abnormal	107 (0.89)	528 (0.44)	2.69	2.18-3.31	<0.001
Normal	130 (1.08)	911 (0.75)	1.89	1.57-2.28	
Number of GP consultations					
0-10	4,316 (35.75)	77,720 (64.37)	1.00		
11-20	4,373 (36.22)	29,327 (24.29)	2.69	2.57-2.81	<0.001
21 or more	3,385 (28.04)	13,684 (11.33)	4.45	4.24-4.68	

§ p-values for binary variables were obtained using the Wald's test of significance. In variables with more than 2 categories, p-values were obtained from the likelihood ratio test

Table 5.3: Symptoms, blood investigations and number of general practice consultations recorded among cases and controls within the 13 to 24 month period prior to lung cancer diagnosis

Variable in GP record 13-24 months before lung cancer diagnosis	Cases n(%) N=12,074	Controls n(%) N=120,731	Unadjusted OR for lung cancer	95% CI	p-value §
Cough	1,774 (14.69)	9,087 (7.53)	2.12	2.00-2.24	<0.001
Haemoptysis	133 (1.10)	191 (0.16)	7.03	5.63-8.78	<0.001
Chest/shoulder pain	959 (7.94)	6,540 (5.42)	1.51	1.40-1.62	<0.001
Voice hoarseness	56 (0.46)	326 (0.27)	1.72	1.30-2.29	<0.001
Dyspnoea	992 (8.22)	3,047 (2.52)	3.46	3.21-3.72	<0.001
Weight loss	139 (1.15)	416 (0.34)	3.37	2.78-4.09	<0.001
Constipation	421 (3.49)	1,848 (1.53)	2.32	2.09-2.59	<0.001
Depressive disorders	449 (3.72)	4,705 (3.90)	0.95	0.86-1.05	0.333
URTI	497 (4.12)	4,274 (3.54)	1.17	1.06-1.29	<0.001
LRTI	566 (4.69)	2,218 (1.84)	2.63	2.39-2.89	<0.001
Non-specific chest infections	1,356 (11.23)	5,856 (4.85)	2.48	2.33-2.64	<0.001
COPD	1,024 (8.48)	1,553 (1.29)	7.11	6.56-7.71	<0.001
Outcome of blood tests					
No blood test record	6,136 (50.82)	79,446 (65.80)	1.00		
Test without results	5,632 (46.65)	39,255 (32.51)	1.86	1.79-1.93	<0.001
Abnormal	127 (1.05)	752 (0.62)	2.19	1.81-2.64	
Normal	179 (1.48)	1,278 (1.06)	1.81	1.55-2.13	
Number of GP consultations					
0-10	3,491 (28.91)	64,881 (53.74)	1.00		
11-20	3,492 (28.92)	29,296 (24.27)	2.22	2.11-2.33	<0.001
21 or more	5,091 (42.16)	26,554 (21.99)	3.56	3.41-3.73	

§ p-values for binary variables were obtained using the Wald's test of significance. In variables with more than 2 categories, p-values were obtained from the likelihood ratio test

In multivariate analysis (Table 5.4), age, sex, Townsend deprivation quintiles, smoking (status and highest daily cigarette consumption), number of general practice consultations as well as symptom presentations of cough, haemoptysis, dyspnoea, weight loss, LRTI, non-specific chest infections and COPD were independently associated with lung cancer up to 24 months before diagnosis. Chest pain, voice hoarseness and URTI were associated with lung cancer in the 4-12 months but not in the 13-24 months before diagnosis. Constipation, depression and blood tests were not independently associated with lung cancer in either the 4-12 or the 13-24 month periods.

Compared with the univariate model, the association of age, sex and smoking with lung cancer were almost unchanged in the multivariate model. The association with deprivation was slightly attenuated in the multivariate model but remained significantly associated with lung cancer. The odds of lung cancer increased with increasing number of daily cigarettes smoked and this effect was stronger among current than ex-smokers.

Table 5.4: Multivariate model of factors associated with lung cancer before diagnosis

Risk factor variable	13-24 months before diagnosis		4-12 months before diagnosis	
	Adjusted odds ratio (95% CI)	P-value §	Adjusted odds ratio (95% CI)	p-value §
Age at diagnosis (yrs)				
40-45	1.00		1.00	
45-50	2.55 (2.00-3.26)		2.50 (1.96-3.19)	
50-55	5.50 (4.40-6.88)		5.42 (4.34-6.78)	
55-60	10.88 (8.78-13.48)	<0.001	10.67 (8.61-13.22)	<0.001
60-65	20.74 (16.80-25.61)		19.59 (15.86-24.18)	
65-70	30.58 (24.78-37.74)		28.61 (23.17-35.32)	
70-75	47.87 (38.80-59.06)		44.74 (36.26-55.21)	
75-80	65.60 (53.13-80.99)		60.03 (48.62-74.12)	
>80	72.53 (58.76-89.53)		65.55 (53.10-80.93)	
Sex				
Male	1.59 (1.53-1.66)	<0.001	1.62 (1.55-1.69)	<0.001
Female	1.00			
Townsend score				
5(most deprived)	1.13 (1.05-1.21)		1.10 (1.02-1.18)	
4	1.14 (1.05-1.21)	0.0001	1.12 (1.05-1.20)	0.0017
3	1.07 (1.00-1.14)		1.07 (1.00-1.14)	
2	1.01 (0.94-1.08)		1.00 (0.93-1.07)	
1 (least deprived)	1.00		1.00	
Missing Townsend records	1.01 (0.90-1.13)		1.01 (0.90-1.13)	
Smoking status and qty				
Current V heavy (40+/d)	16.61 (14.53-18.98)		15.91 (13.90-18.21)	
Current Heavy (20-39/d)	13.65 (12.63-14.75)		13.45 (12.44-14.54)	
Current Mod (10-19/d)	9.85 (9.07-10.70)		9.82 (9.04-10.68)	
Current Light (1-9/d)	6.09 (5.46-6.79)	<0.001	5.98 (5.36-6.68)	<0.001
Current Trivial (<1/d)	2.64 (1.20-5.81)		2.68 (1.21-5.90)	
Current, no qty recorded	3.48 (3.09-3.92)		3.47 (3.08-3.91)	
Ex V Heavy (40+/d)	5.70 (4.80-6.76)		5.33 (4.48-6.35)	
Ex Heavy (20-39/d)	7.15 (6.50-7.86)		6.67 (6.06-7.35)	
Ex Mod (10-19/d)	4.72 (4.27-5.21)		4.50 (4.07-4.98)	
Ex Light (1-9/d)	3.75 (3.31-4.25)		3.54 (3.12-4.02)	
Ex Trivial (<1/d)	1.29 (0.73-2.28)		1.21 (0.68-2.17)	
Ex, no qty recorded	2.69 (2.49-2.91)		2.57 (2.38-2.78)	
Missing smoking records	2.56 (2.33-2.82)		2.70 (2.45-2.97)	
Non smoker	1.00		1.00	
Cough	1.22 (1.14-1.30)	<0.001	1.63 (1.53-1.75)	<0.001
Haemoptysis	3.40 (2.59-4.45)	<0.001	8.70 (6.75-11.20)	<0.001
Dyspnoea	1.18 (1.08-1.29)	<0.001	1.41 (1.29-1.55)	<0.001
Weight loss	1.78 (1.43-2.23)	<0.001	2.66 (2.16-3.29)	<0.001
LRTI	1.40 (1.26-1.53)	<0.001	1.56 (1.38-1.76)	<0.001
Chest infections	1.24 (1.15-1.33)	<0.001	1.55 (1.44-1.68)	<0.001
COPD	1.79 (1.63-1.97)	<0.001	1.61 (1.46-1.78)	<0.001
Chest/shoulder pain*			1.39 (1.28-1.51)	<0.001
Voice hoarseness*			1.79 (1.28-2.49)	0.001
URTI*			1.15 (1.02-1.30)	0.020
No. of GP consultations				
0-10	1.00		1.00	
11-20	1.14 (1.07-1.20)	<0.001	1.23 (1.16-1.29)	<0.001
21 or more	1.17 (1.10-1.24)		1.36 (1.28-1.44)	

§ p-values for binary variables were obtained using the Wald's test of significance. In variables with more than 2 categories, p-values were obtained from the likelihood ratio test

*symptoms not associated with lung cancer in the 13-24 month period before diagnosis

5.4 Discussion

5.4.1 Main findings

Similar to the earlier findings from the study using the case-control dataset that was matched on age and sex, this study has shown an increase in symptom reporting to GPs by patients up to 2 years before lung cancer diagnosis but the overall increase in consultation frequency by cases was shown to occur mostly within the year before diagnosis. There was an increase in the frequency of chest x-ray investigations in cases which occurred at about the 4th month before lung cancer diagnosis implying that investigations were initiated by GPs at this time.

After excluding symptoms recorded in the 4 month period prior to lung cancer diagnosis, symptoms that were more commonly reported in the 2 years before diagnosis were cough, non-specific chest infections, dyspnoea, chest pain and COPD. On taking account of the combined effects of patients' socio-demographic factors, smoking and number of consultations, the symptoms that were found to be independently associated with lung cancer within the 4-24 months before diagnosis were cough, haemoptysis, dyspnoea, weight loss, LRTI, chest infections and COPD. Chest pain, voice hoarseness and URTI remained associated with lung cancer only within the 4-12 months before diagnosis.

Socio-demographic characteristics found to be independently associated with lung cancer were age, sex, deprivation and smoking. These findings were comparable to findings in the UK national lung cancer audit database[84] and other populations[90]. Smoking is the most important risk factor for lung cancer[25 38] and this was reflected in the general practice population in THIN.

5.4.2 Comparison with other studies

This is the first large study that uses a combination of patients' socio-demographic characteristics and general practice records while excluding symptoms in the final months before diagnosis, to identify the early predictors of lung cancer. A few studies have explored the symptoms of lung cancer in general practice[112 114 134] but only one study so far has excluded symptoms in the final months before diagnosis as a means of identifying the early-stage symptoms associated with lung cancer[113]. This study of 247 lung cancer cases and 1235 controls explored symptoms of lung cancer but did not identify the socio-demographic characteristics associated with lung cancer[113]. A more recent study used a combination of baseline risk factors and primary care symptoms up to diagnosis to develop an algorithm to predict lung cancer[192]. In identifying lung cancer predictors for this algorithm, this study included symptoms up to diagnosis which GPs may already be investigating, so the algorithm developed using these predictors may not be able to predict lung cancer early enough to improve clinical outcomes.

Apart from COPD and chest infections (URTI, LRTI and non-specific chest infections), the symptoms which were found to be associated with lung cancer in this study are comparable to those found in the study by Hamilton et al.[113] as well as the NICE guideline recommendations[147]. The association between COPD and lung cancer in this study, although not investigated by Hamilton et al., is similar to the finding by Hippisley-Cox et al.[192]. However, in contrast to Hippisley-Cox et al. where dyspnoea, pneumonia and voice hoarseness were not associated with lung cancer, this study found dyspnoea and LRTI to be associated with lung cancer up to 24 months before diagnosis and voice hoarseness was associated with lung cancer up to 12 months before diagnosis.

The majority of these symptoms, except haemoptysis, can be found in benign conditions and present frequently in general practice[116 256]. The association

between URTI and lung cancer in the 4-12 months before diagnosis is likely explained by GPs making URTI diagnoses following complaints of cough and other non-specific respiratory symptoms in the year before diagnosis. Chest pain and voice hoarseness were also associated with lung cancer in the 4-12 months before diagnosis and this may be because these symptoms are indicative of intra-thoracic spread[110] and therefore characteristic of the later stages of disease.

5.4.3 Strengths and limitations

This study was done using lung cancer cases in the thesis dataset which had previously been validated against UK national lung cancer databases (Chapter 3), hence the results can be generalised to and are widely representative of the early interactions between GPs and lung cancer patients in the UK. As previously highlighted in chapter 2 (section 2.1.4), the database is large and therefore has considerable statistical power. Also, all records that have been used for analyses are routinely collected in general practice and therefore freely available to GPs.

A drawback in this study is the unavailability of information on cigarette pack-years for defining patients' lifetime cigarette exposure. As a proxy, patients' cigarette consumption were categorised using the highest recorded number of cigarettes smoked daily which provided the highest possible daily consumption for the patients. The results from analyses using these categories fit broadly with existing literature. In conducting these analyses, the recorded date of lung cancer diagnosis in THIN was assumed to be the patients' date of diagnosis. In practice however, lung cancer diagnosis is either made in general practice following investigations in primary care, or following diagnosis by a chest physician in secondary care. This leads to the possibility that a patients' actual date of lung cancer diagnosis is earlier than the diagnosis date recorded in the

GP's notes. Despite this limitation, the general characteristics and survival estimates for patients with lung cancer in THIN have been shown to be representative and highly comparable to the lung cancer population in the UK.

5.4.4 Conclusion

Although there is an increase in symptom reporting up to 2 years before lung cancer diagnosis, a considerable amount of consultation for these symptoms were made within the year before diagnosis and suggests the need for more efforts to educate the public and especially smokers on the key symptoms of lung cancer and the need to seek medical care as and when they have these symptoms. The warning symptoms identified 4-12 months and even 13-24 months before diagnosis were comparable to the NICE guideline recommendations[147] indicating that some patients with lung cancer could have been investigated or diagnosed earlier.

A combination of early-stage symptoms in general practice, smoking and socioeconomic characteristics were found to be associated with lung cancer and could be used to develop a predictive score to aid earlier identification of patients at increased lung cancer risk who will benefit from further investigations such as chest x-rays. However, in view of the fact that GPs start chest x-ray investigations for suspected lung cancer at about 4 months before diagnosis and since the increase in general consultation frequency by cases occurs mostly within the year before diagnosis, it seems logical that the optimal time period during which patients' general practice records could be used to reliably predict lung cancer in enough time to improve clinical outcomes is the 4-12 month period before diagnosis. The following chapter will use the variables which were independently associated with lung cancer in the 4-12 months before diagnosis

to develop and validate a lung cancer prediction score for use in general practice.

Chapter 6. The derivation and validation of a general practice risk prediction model for lung cancer

In the previous chapter, the socio-demographic and early clinical features independently associated with lung cancer in general practice were identified. This chapter describes the use of these variables to develop a risk prediction model for lung cancer as well the validation of this model in an independent THIN dataset.

6.1 Introduction

As previously discussed in chapter 1 (section 1.6.2), there are several risk prediction models which have been developed to estimate the risk of lung cancer[189-192 194]. Only one risk-prediction algorithm so far has been developed using a combination of patients' baseline risk factors and symptoms recorded in primary care[192]. This model development incorporated patients' symptoms up to the period immediately before lung cancer diagnosis when GPs will be investigating for suspected lung cancer and may therefore not detect lung cancer early enough to improve clinical outcomes.

Using the combination of patients' early pre-diagnostic symptoms and features in general practice, smoking and socioeconomic characteristics which had been shown in the previous chapter to be independently associated with lung cancer, this chapter aims to develop and validate a lung cancer risk-prediction model that could be used to aid earlier diagnosis of lung cancer in general practice.

6.2 Methods

6.2.1 Derivation of the risk model

In the previous chapter, the socio-demographic and clinical features independently associated with lung cancer in the 4-12 and the 13-24 month periods before lung cancer diagnosis were determined using multivariate logistic regression analyses. These variables were derived using the unmatched dataset developed in chapter 5, hereafter referred to as the derivation dataset.

Based on the findings that the increase in frequency of general consultations among cases occurs within the first year before lung cancer diagnoses and that chest x-rays are initiated by GPs at about 4 months before diagnosis, the 4-12 month period was decided to be an optimal period during which variables independently associated with lung cancer could be used to develop a lung cancer prediction model that would reliably predict lung cancer and aid earlier diagnosis of cases.

In developing the risk probabilities for lung cancer, the method used to develop the Thoracic Surgery Scoring System (Thoracoscore)[257] was applied by assigning the β-Coefficient values (log odds ratio) from multivariate logistic regression model to the respective variables, as a means of ensuring that variables were weighted according to the strength of their association with lung cancer in the model. Aggregate scores were then computed for individual patients in the dataset.

6.2.2 Validation cohort

The last date of data collection in the version of THIN which has been used so far in this thesis was July 28th 2009. To ensure that the risk-prediction model is validated using a dataset of patients in THIN with records spanning over a different period from the derivation period, a more recent version of the dataset which had records of patients up to a last data collection date of September 22nd 2010, was obtained for this purpose.

Since the last date of data collection in the derivation dataset was the 28th of July 2009, the 29th of July 2009 was taken as the date following which the outcome - incidence of lung cancer, could be used to assess the model validation in the validation cohort. The validation cohort comprised of all patients in THIN who were aged 39 years of age or older and free from lung cancer on the 29th of July 2009, the validation start date. Eligibility in this cohort was limited to patients who had at least one year of general practice follow-up before and after the 29th of July 2009.

6.2.3 Validation of the risk model

A lung cancer risk probability score was computed for all patients in the dataset on the basis of the socio-demographic characteristics and symptoms in their records. The β-Coefficient values (log odds ratio) derived from multivariate logistic regression modelling were used to compute aggregate risk probabilities for individual patients in the dataset using the equation:

Risk score = constant + sum of β coefficients at different values of the exposure variables.

The actual number of incident lung cancer cases within the year after the 29th of July 2009 were identified and then the performance of the model was assessed by comparing the sensitivity and specificity at different cut-offs. Additionally, a comparison of the sensitivity and specificity of the model with those of the NICE

guideline symptoms was made. The discriminatory power of the model was assessed by means of a receiver operating characteristic (ROC) curve and then the area under the curve (AUC) statistic was calculated.

6.3 Results

6.3.1 Risk prediction model for lung cancer

Based on the analyses and results in the previous chapter, variables that were found to be independently associated with lung cancer in the 4-12 month period and therefore included in the final model were age, sex, Townsend deprivation quintiles, smoking (status and highest record of cigarettes smoked daily), number of other GP consultations as well as symptom presentations of cough, haemoptysis, dyspnoea, weight loss, LRTI, non-specific chest infections, COPD, chest/shoulder pain, voice hoarseness and URTI (Table 6.1). The odds of lung cancer increased with increasing age, male sex, greater socioeconomic deprivation and higher daily cigarette consumption. Haemoptysis and weight loss were the symptoms associated with the greatest risk of lung cancer. For example, a general practice record of haemoptysis was associated with an 8.7 fold higher risk of lung cancer (odds ratio 8.70; 95% CI 6.75-11.20) and weight loss was associated with a 2.7 fold higher risk of lung cancer (odds ratio 2.66; 95% CI 2.16-3.29).

The β-Coefficient values derived from multivariate logistic regression modelling in the derivation dataset are also detailed in Table 6.1.

Table 6.1. Factors independently associated with lung cancer in the derivation dataset, 4 to 12 months before diagnosis (n=132,805)

Risk factor variable	Adjusted odds ratio (95% CI)	p-value	β coefficient
Age at diagnosis (yrs)			
40-45	1.00		
45-50	2.50 (1.96-3.19)		0.9164
50-55	5.42 (4.34-6.78)		1.6900
55-60	10.67 (8.61-13.22)	<0.001	2.3669
60-65	19.59 (15.86-24.18)		2.9746
65-70	28.61 (23.17-35.32)		3.3534
70-75	44.74 (36.26-55.21)		3.8006
75-80	60.03 (48.62-74.12)		4.0944
>80	65.55 (53.10-80.93)		4.1828
Sex			
Male	1.62 (1.55-1.69)	<0.001	0.4805
Female			
Townsend score			
5(most deprived)	1.10 (1.02-1.18)		0.0932
4	1.12 (1.05-1.20)	0.0017	0.1157
3	1.07 (1.00-1.14)		0.0640
2	1.00 (0.93-1.07)		-0.0009
1 (least deprived)	1.00		
Missing Townsend records	1.01 (0.90-1.13)		0.0099
Smoking status and 6m qty			
Current V heavy (40+/d)	15.91 (13.90-18.21)		2.7664
Current Heavy (20-39/d)	13.45 (12.44-14.54)		2.5984
Current Mod (10-19/d)	9.82 (9.04-10.68)		2.2845
Current Light (1-9/d)	5.98 (5.36-6.68)	<0.001	1.7885
Current Trivial (<1/d)	2.68 (1.21-5.90)		0.9851
Current, no qty recorded	3.47 (3.08-3.91)		1.2432
Ex V Heavy (40+/d)	5.33 (4.48-6.35)		1.6742
Ex Heavy (20-39/d)	6.67 (6.06-7.35)		1.8980
Ex Mod (10-19/d)	4.50 (4.07-4.98)		1.5045
Ex Light (1-9/d)	3.54 (3.12-4.02)		1.2636
Ex Trivial (<1/d)	1.21 (0.68-2.17)		0.1943
Ex, no qty recorded	2.57 (2.38-2.78)		0.9455
Missing smoking records	2.70 (2.45-2.97)		0.9922
Non smoker	1.00		
Cough	1.63 (1.53-1.75)	<0.001	0.4915
Haemoptysis	8.70 (6.75-11.20)	<0.001	2.1630
Dyspnoea	1.41 (1.29-1.55)	<0.001	0.3449
Weight loss	2.66 (2.16-3.29)	<0.001	0.9794
LRTI	1.56 (1.38-1.76)	<0.001	0.4414
Chest infections	1.55 (1.44-1.68)	<0.001	0.4393
COPD	1.61 (1.46-1.78)	<0.001	0.4786
Chest/shoulder pain	1.39 (1.28-1.51)	<0.001	0.3296
Voice hoarseness	1.79 (1.28-2.49)	0.001	0.5806
URTI	1.15 (1.02-1.30)	0.020	0.1417
No. of GP consultations			
0-10	1.00		
11-20	1.23 (1.16-1.29)	<0.001	0.2032
21 or more	1.36 (1.28-1.44)		0.3069
Logistic regression constant			-7.2295

§ p-values for binary variables were obtained using the Wald's test of significance. In variables with more than 2 categories, p-values were obtained from the likelihood ratio test

6.3.2 Model validation in an independent THIN dataset

There were 1,897,742 patients in THIN who had no history of lung cancer up to the 29th of July 2009 and with at least one year of follow-up data after the 29th of July 2009. A total of 71,449 patients had less than one year of follow-up in their general practices before the 29th of July 2009 and were excluded. The final validation cohort therefore comprised of 1,826,293 patients which was made up of 939,299 females (51.4%) and 886,994 males (48.6%). There were 1,728 incident diagnoses of lung cancer (0.09% of the cohort) identified during the one-year of follow-up from the 29th of July 2009.

Risk probability scores were computed for all patients in the validation dataset using the β-coefficient values in Table 6.1 and the number of patients identified by the score as well as the sensitivity and specificity of the risk model at different cut-off values are shown in Table 6.2.

Table 6.2. Performance of the risk model at different cut-off values in the validation population (n=1,826,293)

Cut-off value	Patients at risk of lung cancer based on risk model	Patients not requiring a chest x-ray based on risk model	Number of True positives	Number of True negatives	Sensitivity*	Specificity§
-3	737,390	1,088,903	1,624	1,088,799	93.98%	59.67%
-2.5	541,074	1,285,219	1,526	1,285,017	88.31%	70.43%
-2	388,040	1,438,253	1,375	1,437,900	79.57%	78.81%
-1.5	255,788	1,570,505	1,182	1,569,959	68.40%	86.05%
-1.25	192,433	1,633,860	1,063	1,633,195	61.52%	89.51%
-1	144,523	1,681,770	917	1,680,959	53.07%	92.13%
-0.5	72,883	1,752,292	610	1,752,292	35.30%	96.04%
0	30,994	1,795,299	367	1,793,938	21.24%	98.32%
0.5	11,860	1,814,433	174	1,812,879	10.07%	99.36%

* sensitivity = True positives/(true positives + false negatives)
§specificity = true negatives/ (true negatives + false positives)

Table 6.3. Sensitivity and specificity of NICE guideline symptoms alone in the validation population (n=1,826,293)

symptom	Patients requiring a chest x-ray based on NICE guideline	Patients not requiring a chest x-ray based on NICE guideline	Number of True positives	Number of True negatives	Sensitivity	Specificity
Haemoptysis	1843	1,824,450	24	1,822,746	1.39%	99.90%
Cough	175,290	1,651,003	413	1,649,688	23.90%	90.42%
Chest/shoulder pain	107,753	1,718,540	192	1,717,004	11.11%	94.10%
Dyspnoea	61,631	1,764,662	315	1,763,249	18.23%	96.64%
Weight loss	7,679	1,818,614	26	1,816,912	1.50%	99.58%
Voice hoarseness	5,209	1,821,084	9	1,819,365	0.52%	99.72%

Using only the NICE guidelines symptoms, the number of patients who will be identified to be at risk of lung cancer and hence require a chest x-ray, the number of true positives and the sensitivity and specificity of the guideline symptoms in predicting lung cancer risk are shown in Table 6.3. Using Haemoptysis alone as a trigger for chest x-rays, only 24 cases of lung cancer in the cohort population can be detected. Using the most commonly reported symptom cough as a trigger for investigations, 175,290 patients are identified to be at risk of lung cancer and 413 of these will be diagnosed with lung cancer.

Using the NICE symptoms therefore to identify a comparable number of true positives as the lung cancer risk model, a higher number of patients are required to undergo chest x-rays than the risk model. For example, at a cut-off to identify 610 cases of lung cancer in the validation cohort, the risk model identified 72,883 patients at high risk of lung cancer for whom chest x-ray investigations are indicated, yet using a weighted combination of all the NICE symptoms, a total of 305,137 patients will have to undergo chest-ray investigations to identify 724 cases of lung cancer.

The receiver operating characteristic curve obtained from the application of the risk model in the validation cohort is shown in Figure 6.1. The area under the curve (AUC) is 0.88. Using a weighted combination of the NICE guideline symptoms to identify patients at high risk of lung cancer, the area under the receiver operating characteristic curve is 0.64 (Figure 6.2).

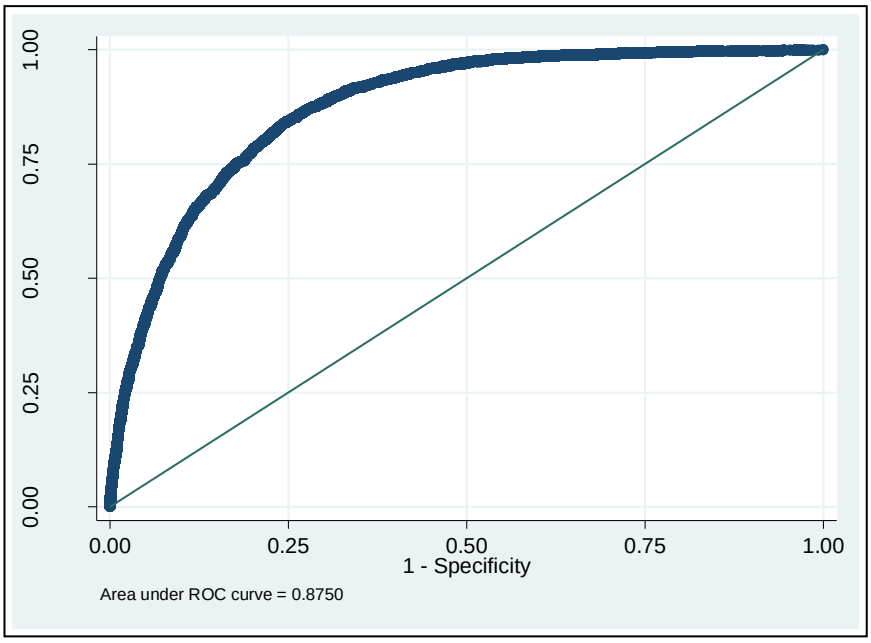


Figure 6.1. Receiver operating characteristic curve for the lung cancer risk prediction model.
The area under the curve is 0.88. the diagonal line represents the discrimination expected by chance alone

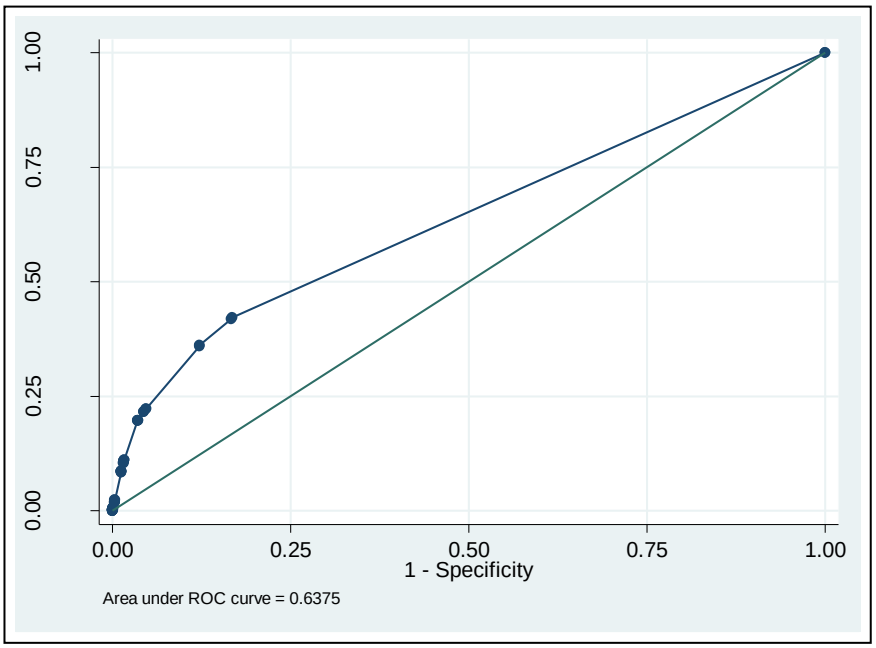


Figure 6.2. Receiver operating characteristic curve for a lung cancer risk model developed using a weighted combination of the NICE guideline symptoms
The area under the curve is 0.64

6.4 Discussion

6.4.1 Main findings from study

In this chapter, a lung cancer risk prediction model was developed using a combination of patients' socio-demographic and clinical records which were found to be independently associated with lung cancer in the 4-12 month period before diagnosis. On validating this model in an independent dataset, it performed well and showed good discrimination with an area under the receiver operating characteristic curve of 0.88.

6.4.2 Strengths and limitations

Strengths of THIN database such as the large size and representativeness of the UK lung cancer population are as discussed in previous sections. The information which have been incorporated into the risk model development are readily available to GPs so the application of the score in practice will be relatively easy and at no extra cost to GPs. In developing this risk model, records made in the 4 months before diagnosis were excluded to avoid the inclusion of symptoms, diagnoses and investigations attributable to lung cancer instead of predictive of it. This ensures that the model can aid earlier diagnosis and improve clinical outcomes for people with lung cancer.

Limitations of this study include the lack of relevant information for example family history of lung cancer and occupational exposure to carcinogens such as asbestos, which were unavailable in THIN and so could not be included in the model. Although inclusion of these variables may improve the performance of the model, the validation analyses using the currently available variables have shown good discrimination and the model performed substantially better than the current NICE guidelines[147] when validated in an independent dataset.

In prospective analyses to assess the model performance in the validation cohort, patients' clinical data were collected at different time periods during the year before the 28th of July 2009 and the outcome of lung cancer incidence was measured at different time periods after the 29th of July 2009. By doing this, it was not possible to identify and exclude an appropriate 4 month period before lung cancer diagnosis in the validation cohort. However, further analyses was done to assess the model performance after excluding 580 incident lung cancer diagnosis made during the 4 months after the 29th of July 2009 and the results showed a similar model discrimination with an area under the ROC curve of 0.88.

Analysis of the risk model in the validation cohort showed that a considerable number of patients need to undergo chest x-ray investigations to diagnose lung cancer cases. The positive predictive value of the model is therefore not as high as the positive predictive value of the NICE guideline symptoms, as previously reported by a study[142]. This is unsurprising considering that lung cancer was rare in the population and was only diagnosed in 1,728 patients (0.09% of the population). As previously noted, positive predictive values are not good measures of model accuracy particularly with rare outcomes as they are usually low even with good sensitivity and specificity[158]. A similar finding was shown in the randomised Danish lung cancer screening trial where 980 CT scans were done in order to identify 69 lung cancer cases[258]. This model however compared quite favourably with the NICE guideline symptoms, with less than a quarter of chest x-rays required to detect a comparable number of lung cancers even than a weighted combination of the NICE guideline symptoms.

6.4.3 Comparison with other studies

As previously discussed in chapter 1, a number of models including the Bach[189], Spitz[194] and the Liverpool Lung Project (LLP)[191] have been developed to predict the risk of lung using patients' baseline risk factors. A study which compared the

discriminatory power of these three models found an AUC statistic of 0.69 for both the Spitz and LLP models and AUC of 0.66 for the Bach model[195] and these are substantially lower than the AUC statistic value of 0.88 in this model. The LLP model is currently being used to select individuals who have a 5% risk of developing lung cancer over 5 years for inclusion into the UK lung screen (UKLS) trial of low dose CT screening for lung cancer[128]. However, at a cut-off to capture 62% of cases of lung cancer, the LLP model falsely identifies 30% of non-lung cancer controls and does not perform as well as this risk model which for accurately identifying 79.6% of lung cancer cases gives a false positive rate of 21.2%.

The only other model which uses primary care data to predict lung cancer was developed using patient records up to a certain point to establish baseline risk, after which incident diagnoses of lung cancer over the subsequent 2 years were predicted. The model appears to have a good discriminatory power with ROC values of 0.92 for males and females, however all GP records of patients recorded in the period leading up to lung cancer diagnosis were included in the algorithm development so it is likely that many symptoms and smoking records included were those after the point at which clinical lung cancer investigations were already underway and a diagnosis of lung cancer were actively being sought by the GPs. The study in this thesis has shown that in the 4 month period leading up to lung cancer diagnosis, the majority of patients with lung cancer start undergoing investigations in general practice. It follows therefore that the model developed by Hippisley-Cox et al.[192] will be predicting lung cancer in patients that are already being investigated in general practice and hence it is of limited value in diagnosing lung cancer at an earlier stage.

6.5 Conclusion

A combination of the early-stage symptoms of lung cancer presented in general practice, smoking status and socioeconomic characteristics associated with lung cancer appear to aid earlier identification of patients who are at an increased risk of lung cancer and who will benefit from further investigations such as chest x-rays. The weighting and inclusion of socio-demographic variables - age, sex, socioeconomic status and smoking, as well as the weighting and inclusion of other clinical diagnoses - upper respiratory tract infections, lower respiratory tract infections, non-specific chest infections, COPD and the frequency of general practice consultations make this model a huge improvement on the NICE list[147] of symptoms. Evidence from past research has shown that a delay of 18 to 131 days (median of 54 days) between diagnosis and curative treatment for lung cancer was associated with an increase in cross-sectional tumour size and an increased risk of the cancer becoming incurable[133]. The outcomes of lung cancer are likely to be better in patients referred earlier and whose disease is diagnosed earlier because they may have earlier stage disease and better performance status. Earlier identification of lung cancer would consequently avert disease progression and metastases and lead to improved prognosis for people with lung cancer. A clinical trial perhaps in conjunction with a screening trial, is needed to fully quantify the benefit of the model in practice.

Chapter 7 Conclusions and recommendations for future research

7.1 Summary of main findings

The main findings of the studies in this thesis are as follows:

- The characteristics of patients with lung cancer in The Health Improvement Network (THIN) are comparable to patient characteristics in two reliable UK national lung cancer databases - The UK National Cancer Registry[230] and the National Lung Cancer Audit Database (LUCADA)[231]. The database was found to capture a high proportion of incident lung cancer cases from cancer registries. THIN is therefore highly representative of the national UK population of patients with lung cancer and is a potentially valuable tool for lung cancer research in the UK.
- Experian's Mosaic Public Sector™ classification tool linked with patients' records in THIN identified wider variations in lung cancer incidence across different types and groups than the more widely used socio-economic classification marker - Townsend deprivation quintiles. In doing this, it was able to identify the specific sectors of the UK population where the incidence of lung cancer was highest.
- There is no trend of increasing smoking-associated risk of lung cancer with deprivation and therefore no evidence to suggest that more deprived individuals are more vulnerable to the harmful effects of cigarette smoke than less deprived individuals. Although the risk of lung cancer is greater among individuals of lower socioeconomic groups compared to individuals of higher socioeconomic groups, this is largely due to smoking.
- Depression is associated with an increased risk of lung cancer and this is largely explained by the higher prevalence of cigarette smoking among people with depression compared to people with no depression. There is

no difference in the smoking-associated risk of lung cancer between depressed and non-depressed individuals, hence depression does not make individuals more vulnerable to the carcinogenic effects of smoking.

- There is an increase in the frequency of general consultations as well as specific consultations for lung cancer symptoms up to two years before lung cancer is diagnosed in patients in general practice. A combination of patients' early symptoms in general practice, smoking and sociodemographic features was found to be independently associated with lung cancer.
- A lung cancer risk prediction model was developed using the socio-demographic and clinical records that were independently associated with lung cancer in the 4 to 12 month period before lung cancer diagnosis. This model showed good discrimination when validated in an independent dataset and it performed better than a combination of the NICE guideline symptoms alone and also out-performed existing models for lung cancer.

7.2 Clinical implications

The studies in this thesis provide substantial evidence that can inform the care pathway of patients who may be at risk of lung cancer in general practice or in the general population.

The Experian's Mosaic public sector ^{TM} classification is a useful tool which if applied outside general practice, will enable tailored public health lung cancer campaigns and interventions to be more effectively targeted to specific groups of people in society.

Although this research provides support to the existing body of evidence on the increased risk of lung cancer among individuals of lower socioeconomic status as

well as individuals with a history of depression, there was no evidence to suggest that these individuals are more susceptible to the effects of smoking compared to individuals of higher socioeconomic status and individuals with no history of depression respectively. Since the prevalence of cigarette smoking is typically higher in individuals of lower socioeconomic status as well as depressed individuals, there is a pressing need for smoking cessation programs to be specifically targeted at deprived communities as well as the incorporation of smoking cessation interventions into the NICE guidelines for the management of patients with depression.

Findings from this research showed that there is an increase in symptom presentation and other clinical activity such as general consultations and blood test investigations up to two years before lung cancer diagnosis and especially within the year leading up to diagnosis. Most patients with lung cancer do not start to present frequently with symptoms until the year before diagnosis and it is likely that the majority of them have symptoms for a considerable period of time before they present to the GP. To get the maximum benefit from any general practice predictive score for lung cancer, it is essential that patients present early with their symptoms. There is a public health need therefore for more efforts to educate the public and especially smokers on the key symptoms of lung cancer and the need to seek medical care as and when they have symptoms. During the one year to four months before diagnosis, the symptoms which were found to be independently associated with lung cancer were similar to the symptoms in the NICE guideline[147] and are common symptom presentations in general practice.

The weighting and inclusion of patients' socio-demographic features in addition to clinical symptoms performed better than the NICE guidelines symptoms alone and will not only aid earlier recognition by GPs of patients at high risk of lung cancer who will benefit from further investigations and/or earlier specialist

referral, but it will also enable earlier intervention, avert disease progression, prevent a substantial number of early lung cancer deaths and consequently improve survival of lung cancer. In applying this model clinically, the aim is to incorporate the algorithm into GP software so that at a certain threshold, a hint is offered to the GP to investigate the patient for possible lung cancer and as such, these would not need to be directly calculated by GPs. Similar methods are already being used for the calculation of cardiovascular disease risk and the benefits of this as opposed to GPs working out the score for individual patients is that rather than making a risk estimation based on information collected by the GP during a consultation, the system takes account of all previous recorded data for patients including records entered during consultation with other GPs in the same practice .

7.3 Suggestions for further research

7.3.1 The use of Experian's Mosaic tool to target lung cancer public health services

Experian's Mosaic public sector ™ variable has been shown to be a useful tool to aid more precise targeting of lung cancer-related public health interventions to specific sectors of society. A more detailed knowledge of the information used to derive the Mosaic groups and types is needed to assess the validity of the tool and quantify its benefits over the Townsend deprivation quintiles. Since the Mosaic categories that had the highest incidence of lung cancer comprised of the elderly and deprived individuals in society, it will therefore be useful to know the additional benefits of the Mosaic over an age-adjusted Townsend measure of socioeconomic status. To adequately assess the benefit of Mosaic in identifying the particular sectors of society where lung cancer interventions are most required, a practical implementation of this tool to deliver tailored lung cancer

interventions is needed. One way in which Mosaic can be assessed in practice is to run pilot schemes of health promotion strategies in geographical locations, some of which would incorporate Mosaic in the planning stage as a means of tailoring specific interventions to specific sectors of society. An evaluation of the effectiveness of the tool in practice can then be assessed and quantified.

7.3.2 Smoking-associated risk of lung cancer in deprived individuals

The lack of variation in the risk of lung cancer between individuals of lower socioeconomic status and individuals of higher socioeconomic status, though not entirely surprising, raises the need for further research using general practice data to investigate and possibly quantify the contribution of other risk factors such as occupational exposure, diet and alcohol consumption to the association between socioeconomic status and lung cancer. Lifestyle risk factors are not readily available in general practice database, however a possible approach that will enable a comprehensive study of risk factor exposures among individuals from different socioeconomic groups and an outcome of lung cancer would be a nested case control study using a cohort study database which measures lifestyle risk factors as the primary exposure with information on other confounders such as occupational and environmental factors.

7.3.3 Smoking-associated risk of lung cancer in depressed compared to non-depressed individuals

The finding that there is no difference in the smoking-associated risk of lung cancer between individuals with and without a previous history of depression is an important addition to the evidence base of lung cancer risk in depressed people. The depression records analysed in this thesis were however diagnosis

made by GPs during routine consultation in general practice and were not based on standardised psychiatric assessments. Although the results from this study are applicable in UK general practices, a study which explores the association between depressed patients based on psychiatric criteria and the risk of lung cancer will be worthwhile and will provide definitive evidence of the true difference in lung cancer risk among smokers with depression and smokers with no history of depression.

7.3.4 Validation of the general practice prediction model for lung cancer

Validation of the lung cancer risk-prediction model that was developed in this thesis was done using a cohort of 1,826,293 patients with only 1,728 incident lung cancer diagnosis identified during the follow-up period from July 2009 to September 2010 and this was fairly small compared to the model derivation population where there were 12,074 incident cases of lung cancer identified over a 10 year period from 2000 to 2009. Despite the fact that the validation in an independent THIN dataset showed good discrimination of the model with an area under the ROC curve of 0.88, validation of the risk model may have been better assessed using a larger validation cohort with data collected over a longer time period. In practice however, the best way to assess the accuracy as well as quantify the benefit of the risk-prediction model would be to do a clinical trial using the model to identify patients at risk of lung cancer at a defined time period and then to measure an outcome of lung cancer 4 to 12 months afterwards.

7.3.5 Proportion of patients with lung cancer diagnosed following urgent general practice referral

A Study over a 2 year period at the Bradford hospitals NHS trust had shown that only 23% of patients with lung cancer were referred urgently by their GPs[142] and another study of all 246 patients with primary lung cancers in Exeter, UK showed that 45% of the patients were diagnosed following referral to hospital respiratory departments for specialist investigation[141]. There is the need for a large scale study to estimate the exact proportion of patients with lung cancer nationally who are diagnosed following the general practice urgent referral route. One way to achieve this would be a study using a national hospital dataset which provides information on whether individual patients have been referred from general practice and whether the referral was urgent or non-urgent.

7.4 Conclusion

The studies in this thesis have demonstrated the usefulness of general practice data from THIN for studies to explore the early interaction between GPs and patients before lung cancer is diagnosed. Although there is an increase in clinical activity of patients before lung cancer is diagnosed, a considerable amount of these occur within the year before diagnosis and suggests the need for further efforts to educate the public especially smokers and those from the sectors of society where lung cancer incidence is highest, on the need to seek medical care when they have symptoms. Using the early features of patients within the year up to 4 months before diagnosis, this thesis has been able to develop a risk prediction score which has not only out-performed existing scores but compares quite favourably with the NICE guidelines and can aid earlier diagnosis of lung cancer in future cases.

Appendix I: List of Read codes

a) List of Read codes for lung cancer

Read code	Description
B22..00	Malignant neoplasm of trachea, bronchus and lung
B220.00	Malignant neoplasm of trachea
B220z00	Malignant neoplasm of trachea NOS
B221.00	Malignant neoplasm of main bronchus
B221000	Malignant neoplasm of carina of bronchus
B221100	Malignant neoplasm of hilus of lung
B221z00	Malignant neoplasm of main bronchus NOS
B222.00	Malignant neoplasm of upper lobe, bronchus or lung
B222.11	Pancoast's syndrome
B222000	Malignant neoplasm of upper lobe bronchus
B222100	Malignant neoplasm of upper lobe of lung
B222z00	Malignant neoplasm of upper lobe, bronchus or lung NOS
B223.00	Malignant neoplasm of middle lobe, bronchus or lung
B223000	Malignant neoplasm of middle lobe bronchus
B223100	Malignant neoplasm of middle lobe of lung
B223z00	Malignant neoplasm of middle lobe, bronchus or lung NOS
B224.00	Malignant neoplasm of lower lobe, bronchus or lung
B224000	Malignant neoplasm of lower lobe bronchus
B224100	Malignant neoplasm of lower lobe of lung
B224z00	Malignant neoplasm of lower lobe, bronchus or lung NOS
B225.00	Malignant neoplasm of overlapping lesion of bronchus & lung
B22y.00	Malignant neoplasm of other sites of bronchus or lung
B22z.00	Malignant neoplasm of bronchus or lung NOS
B22z.11	Lung cancer
B26..00	Malignant neoplasm, overlap lesion of resp & intrathor orgs
B2zz.00	Malignant neoplasm of respiratory tract NOS
B551100	Malignant neoplasm of chest wall NOS
B551z00	Malignant neoplasm of thorax NOS
Byu2.00	[X]Malignant neoplasm of respiratory and intrathoracic orga
Byu2000	[X]Malignant neoplasm of bronchus or lung, unspecified
Byu2400	[X]Malignant neoplasm/ill-defined sites within resp system

b) List of Read codes for histology

Read code	Description
BB08.00	[M]Malignant tumour, small cell type
BB17.00	[M]Large cell carcinoma NOS
BB1J.00	[M]Small cell carcinoma NOS
BB1L.00	[M]Small cell carcinoma, fusiform cell type
BB1M.00	[M]Small cell carcinoma, intermediate cell
BB1N.00	[M]Small cell-large cell carcinoma
BB2..00	[M]Papillary and squamous cell neoplasms
BB2..12	[M]Squamous cell neoplasms
BB25.00	[M]Squamous cell papilloma
BB26.00	[M]Papillary squamous cell carcinoma
BB29.00	[M]Squamous cell carcinoma in situ NOS
BB29.13	[M]Intraepithelial squamous cell carcinoma
BB2A.00	[M]Squamous cell carcinoma NOS
BB2B.00	[M]Squamous cell carcinoma, metastatic NOS
BB2C.00	[M]Squamous cell carcinoma, keratinising type NOS
BB2D.00	[M]Squamous cell carcinoma, large cell, non-keratinising
BB2E.00	[M]Squamous cell carcinoma, small cell, non-keratinising
BB2F.00	[M]Squamous cell carcinoma, spindle cell type
BB2G.00	[M]Adenoid squamous cell carcinoma
BB2H.00	[M]Squamous cell ca-in-situ, questionable stromal invasion
BB2J.00	[M]Squamous cell carcinoma, microinvasive
BB2z.00	[M]Papillary or squamous cell neoplasm NOS
BB35.00	[M]Basosquamous carcinoma
BB5..00	[M]Adenomas and adenocarcinomas
BB5..11	[M]Adenocarcinomas
BB51.00	[M]Adenocarcinoma in situ
BB51000	[M]Adenocarcinoma in situ in villous adenoma
BB51100	[M]Adenocarcinoma in situ in tubulovillous adenoma
BB52.00	[M]Adenocarcinoma NOS
BB52000	[M]Adenocarcinoma in tubulovillous adenoma
BB53.00	[M]Adenocarcinoma, metastatic, NOS
BB54.00	[M]Scirrhous adenocarcinoma
BB56.00	[M]Superficial spreading adenocarcinoma
BB5F.00	[M]Trabecular adenocarcinoma
BB5J.11	[M]Cylindroid adenocarcinoma
BB5M.00	[M]Tubular adenomas and adenocarcinomas
BB5M100	[M]Tubular adenocarcinoma
BB5Mz00	[M]Tubular adenoma or adenocarcinoma NOS
BB5R800	[M]Adenocarcinoid tumour
BB5S.00	[M]Respiratory tract adenomas and adenocarcinomas
BB5S200	[M]Bronchiolo-alveolar adenocarcinoma
BB5S400	[M]Alveolar adenocarcinoma
BB5Sz00	[M]Respiratory tract adenoma or adenocarcinoma NOS
BB5T.00	[M]Papillary adenomas and adenocarcinomas
BB5T100	[M]Papillary adenocarcinoma NOS
BB5Tz00	[M]Papillary adenoma or adenocarcinoma NOS
BB5U.00	[M]Villous adenomas and adenocarcinomas
BB5U100	[M]Adenocarcinoma in villous adenoma
BB5U200	[M]Villous adenocarcinoma
BB5Uz00	[M]Villous adenoma or adenocarcinoma NOS
BB5W.00	[M]Oxyphilic adenomas and adenocarcinomas
BB5W100	[M]Oxyphilic adenocarcinoma
BB5W111	[M]Hurthle cell adenocarcinoma
BB5W112	[M]Oncytic adenocarcinoma
BB5Wz00	[M]Oxyphilic adenoma or adenocarcinoma NOS
BB5X.00	[M]Clear cell adenomas and adenocarcinomas
BB5X100	[M]Clear cell adenocarcinoma NOS
BB5Xz00	[M]Clear cell adenoma or adenocarcinoma NOS
BB5c200	[M]Water-clear cell adenocarcinoma
BB5d.00	[M]Mixed cell adenoma and adenocarcinoma
BB5d100	[M]Mixed cell adenocarcinoma
BB5dz00	[M]Mixed cell adenoma or adenocarcinoma NOS
BB5f600	[M]Papillary and follicular adenocarcinoma
BB5y.00	[M]Adenoma and adenocarcinoms OS
BB5y000	[M]Basal cell adenocarcinoma

BB5z.00	[M]Adenoma or adenocarcinoma NOS
BB82.00	[M]Mucinous adenoma and adenocarcinoma
BB82100	[M]Mucinous adenocarcinoma
BB82111	[M]Colloid adenocarcinoma
BB82112	[M]Gelatinous adenocarcinoma
BB82113	[M]Mucoid adenocarcionoma
BB82114	[M]Mucous adenocarcinoma
BB82z00	[M]Mucinous adenoma or adenocarcinoma NOS
BB84.00	[M]Mucin-producing adenocarcinoma
BB91000	[M]Intraductal papillary adenocarcinoma with invasion
BB96.00	[M]Noninfiltrating intraductal papillary adenocarcinoma
BBB0.00	[M]Adenosquamous carcinoma
BBB2.00	[M]Adenocarcinoma with squamous metaplasia
BBB4.00	[M]Adenocarcinoma with spindle cell metaplasia
BBB5.00	[M]Adenocarcinoma with apocrine metaplasia
H58y400	Squamous metaplasia of lung

c) List of Read codes for Chronic Obstructive Pulmonary Disease

Read code	Description
66YI.00	COPD self-management plan given
66YL.00	Chronic obstructive pulmonary disease follow-up
66YL.11	COPD follow-up
66YL.12	COAD follow-up
66YM.00	Chronic obstructive pulmonary disease annual review
8H2R.00	Admit COPD emergency
14B3.00	history of COPD
H3...00	Chronic obstructive pulmonary disease
H3...11	Chronic obstructive airways disease
H31..00	Chronic bronchitis
H310.00	Simple chronic bronchitis
H310000	Chronic catarrhal bronchitis
H310z00	Simple chronic bronchitis NOS
H311.00	Mucopurulent chronic bronchitis
H311000	Purulent chronic bronchitis
H311100	Fetid chronic bronchitis
H311z00	Mucopurulent chronic bronchitis NOS
H312.00	Obstructive chronic bronchitis
H312100	Emphysematous bronchitis
H312200	Acute exacerbation of chronic obstructive airways disease
H312z00	Obstructive chronic bronchitis NOS
H313.00	Mixed simple and mucopurulent chronic bronchitis
H31y.00	Other chronic bronchitis
H31y100	Chronic tracheobronchitis
H31yz00	Other chronic bronchitis NOS
H31z.00	Chronic bronchitis NOS
H32..00	Emphysema
H320.00	Chronic bullous emphysema
H320000	Segmental bullous emphysema
H320100	Zonal bullous emphysema
H320200	Giant bullous emphysema
H320300	Bullous emphysema with collapse
H320z00	Chronic bullous emphysema NOS
H321.00	Panlobular emphysema
H322.00	Centrilobular emphysema
H32y.00	Other emphysema
H32y000	Acute vesicular emphysema
H32y100	Atrophic (senile) emphysema
H32y111	Acute interstitial emphysema
H32y200	MacLeod's unilateral emphysema
H32yz00	Other emphysema NOS
H32z.00	Emphysema NOS
H36..00	Mild chronic obstructive pulmonary disease
H37..00	Moderate chronic obstructive pulmonary disease
H38..00	Severe chronic obstructive pulmonary disease
H3y..00	Other specified chronic obstructive airways disease
H3y..11	Other specified chronic obstructive pulmonary disease
H3z..00	Chronic obstructive airways disease NOS
H3z..11	Chronic obstructive pulmonary disease NOS
Hyu3000	[X]Other emphysema
Hyu3100	[X]Other specified chronic obstructive pulmonary disease
H312000	Chronic asthmatic bronchitis
H312011	Chronic wheezy bronchitis
H312300	Bronchiolitis obliterans
H320311	Tension pneumatocoele
H32yz11	Sawyer - Jones syndrome
H3y0.00	Chronic obstruct pulmonary disease with acute lower resp infection
H3y1.00	Chronic obstruct pulmonary dis wth acute exacerbation, unspecified

d) List of smoking status Read codes

Read code	Description	Smoking category
137..00	Tobacco consumption	see AHD
137..11	Smoker - amount smoked	Current
1371.00	Never smoked tobacco	Never
1371.11	Non-smoker	see AHD
1372.00	Trivial smoker - < 1 cig/day	Current
1372.11	Occasional smoker	Current
1373.00	Light smoker - 1-9 cigs/day	Current
1374.00	Moderate smoker - 10-19 cigs/d	Current
1375.00	Heavy smoker - 20-39 cigs/day	Current
1376.00	Very heavy smoker - 40+cigs/d	Current
1377.00	Ex-trivial smoker (<1/day)	Ex
1378.00	Ex-light smoker (1-9/day)	Ex
1379.00	Ex-moderate smoker (10-19/day)	Ex
137A.00	Ex-heavy smoker (20-39/day)	Ex
137B.00	Ex-very heavy smoker (40+/day)	Ex
137C.00	Keeps trying to stop smoking	Current
137D.00	Admitted tobacco cons untrue ?	Unknown
137E.00	Tobacco consumption unknown	Unknown
137F.00	Ex-smoker - amount unknown	Ex
137G.00	Trying to give up smoking	Current
137H.00	Pipe smoker	Current
137J.00	Cigar smoker	Current
137K.00	Stopped smoking	Ex
137L.00	Current non-smoker	see AHD
137M.00	Rolls own cigarettes	Current
137N.00	Ex pipe smoker	Ex
137O.00	Ex cigar smoker	Ex
137P.00	Cigarette smoker	Current
137P.11	Smoker	Current
137Q.00	Smoking started	Current
137Q.11	Smoking restarted	Current
137R.00	Current smoker	Current
137S.00	Ex smoker	Ex
137T.00	Date ceased smoking	Ex
137V.00	Smoking reduced	Current
137X.00	Cigarette consumption	see AHD
137Y.00	Cigar consumption	see AHD
137Z.00	Tobacco consumption NOS	see AHD
137a.00	Pipe tobacco consumption	see AHD
137b.00	Ready to stop smoking	Current
137c.00	Thinking about stopping smoking	Current
137d.00	Not interested in stopping smoking	Current
137e.00	Smoking restarted	Current
137f.00	Reason for restarting smoking	Current
137g.00	Cigarette pack-years	Unknown
137h.00	Minutes from waking to first tobacco consumption	Current
13p..00	Smoking cessation milestones	Unknown
13p0.00	Negotiated date for cessation of smoking	Current
13p1.00	Smoking status at 4 weeks	Unknown
13p2.00	Smoking status between 4 and 52 weeks	Unknown
13p3.00	Smoking status at 52 weeks	Unknown
13p4.00	Smoking free weeks	Unknown
13p5.00	Smoking cessation programme start date	Current
13p6.00	Carbon monoxide reading at 4 weeks	Unknown
4I90.00	Expired carbon monoxide concentration	Unknown
6791.00	Health ed. - smoking	Current
67A3.00	Pregnancy smoking advice	Current
67H1.00	Lifestyle advice regarding smoking	Current
6893.00	Tobacco usage screen	see AHD
68T..00	Tobacco usage screen	see AHD
745H.00	Smoking cessation therapy	Unknown
745H000	Nicotine replacement therapy using nicotine patches	Current
745H100	Nicotine replacement therapy using nicotine gum	Current
745H200	Nicotine replacement therapy using nicotine inhalator	Current

745H300	Nicotine replacement therapy using nicotine lozenges	Current
745H400	Smoking cessation drug therapy	Current
745Hy00	Other specified smoking cessation therapy	Current
745Hz00	Smoking cessation therapy NOS	Unknown
8B2B.00	Nicotine replacement therapy	Current
8B3Y.00	Over the counter nicotine replacement therapy	Current
8B3f.00	Nicotine replacement therapy provided free	Current
8BP3.00	Nicotine replacement therapy provided by community pharmacist	Current
8CAL.00	Smoking cessation advice	Current
8CAg.00	Smoking cessation advice provided by community pharmacist	Current
8H7i.00	Referral to smoking cessation advisor	Current
8HTK.00	Referral to stop-smoking clinic	Current
8I2I.00	Nicotine replacement therapy contraindicated	Current
8I39.00	Nicotine replacement therapy refused	Current
9N2k.00	Seen by smoking cessation advisor	Unknown
9N4M.00	DNA - Did not attend smoking cessation clinic	Unknown
9OO..00	Anti-smoking monitoring admin.	Unknown
9OO..11	Stop smoking clinic admin.	Unknown
9OO..12	Stop smoking monitoring admin.	Unknown
9OO1.00	Attends stop smoking monitor.	Unknown
9OO2.00	Refuses stop smoking monitor	Unknown
9OO3.00	Stop smoking monitor default	Unknown
9OO4.00	Stop smoking monitor 1st letter	Unknown
9OO5.00	Stop smoking monitor 2nd letter	Unknown
9OO6.00	Stop smoking monitor 3rd letter	Unknown
9OO7.00	Stop smoking monitor verb.inv.	Current
9OO8.00	Stop smoking monitor phone inv	Current
9OO9.00	Stop smoking monitoring delete	Unknown
9OOA.00	Stop smoking monitor. check done	Unknown
9OOZ.00	Stop smoking monitor admin.NOS	Unknown
9hG..00	Exception reporting: smoking quality indicators	Exception
9hG0.00	Excepted from smoking quality indicators: Patient unsuitable	Exception
9hG1.00	Excepted from smoking quality indicators: Informed dissent	Exception
E023.00	Nicotine withdrawal	Unknown
E251.00	Tobacco dependence	Current
E251100	Tobacco dependence, continuous	Current
E251300	Tobacco dependence in remission	Ex
E251z00	Tobacco dependence NOS	Current
ZG23300	Advice on smoking	Current
ZRBm200	Fagerstrom test for nicotine dependence	Current
ZRBm211	FTND - Fagerstrom test for nicotine dependence	Current
ZRaM.00	Motives for smoking scale	Current
ZRaM.11	MFS - Motives for smoking scale	Current
ZRao.00	Occasions for smoking scale	Current
ZRh4.00	Reasons for smoking scale	Current
ZRh4.11	RFS - Reasons for smoking scale	Current
ZV11600	[V]Personal history of tobacco abuse	Unknown
ZV4K000	[V]Tobacco use	see AHD
ZV6D800	[V]Tobacco abuse counselling	Current
137j.00	Ex-cigarette smoker	Ex

e) List of Read codes for records of quantity of cigarettes smoked

Read code	Description	Smoking category
1374.00	Moderate smoker - 10-19 cigs/d	current/moderate
1373.00	Light smoker - 1-9 cigs/day	current/light
1375.00	Heavy smoker - 20-39 cigs/day	current/heavy
1372.00	Trivial smoker - < 1 cig/day	current/trivial
1376.00	Very heavy smoker - 40+cigs/d	current/very heavy
1379.00	Ex-moderate smoker (10-19/day)	Ex/moderate
1378.00	Ex-light smoker (1-9/day)	Ex/light
137A.00	Ex-heavy smoker (20-39/day)	Ex/heavy
1377.00	Ex-trivial smoker (<1/day)	Ex/trivial
137B.00	Ex-very heavy smoker (40+/day)	Ex/very heavy
137..00	Tobacco consumption	see AHD
137Z.00	Tobacco consumption NOS	see AHD
137a.00	Pipe tobacco consumption	see AHD
137Y.00	Cigar consumption	see AHD
137X.00	Cigarette consumption	see AHD
ZV4K000	[V]Tobacco use	see AHD

f) List of Read codes for Depression

Read code	description
1B17.00	Depressed
1B17.11	C/O - feeling depressed
1B1U.00	Symptoms of depression
1B1U.11	Depressive symptoms
1BT..00	Depressed mood
1BT..11	Low mood
2257.00	O/E - depressed
62T1.00	Puerperal depression
6G00.00	Postnatal depression counselling
E11..12	Depressive psychoses
E112.00	Single major depressive episode
E112.11	Agitated depression
E112.12	Endogenous depression first episode
E112.13	Endogenous depression first episode
E112.14	Endogenous depression
E112000	Single major depressive episode, unspecified
E112100	Single major depressive episode, mild
E112200	Single major depressive episode, moderate
E112300	Single major depressive episode, severe, without psychosis
E112400	Single major depressive episode, severe, with psychosis
E112z00	Single major depressive episode NOS
E113.00	Recurrent major depressive episode
E113.11	Endogenous depression - recurrent
E113000	Recurrent major depressive episodes, unspecified
E113100	Recurrent major depressive episodes, mild
E113200	Recurrent major depressive episodes, moderate
E113300	Recurrent major depressive episodes, severe, no psychosis
E113400	Recurrent major depressive episodes, severe, with psychosis
E113700	Recurrent depression
E113z00	Recurrent major depressive episode NOS
E118.00	Seasonal affective disorder
E11y200	Atypical depressive disorder
E11z200	Masked depression
E130.00	Reactive depressive psychosis
E130.11	Psychotic reactive depression
E135.00	Agitated depression
E200300	Anxiety with depression
E204.00	Neurotic depression reactive type
E204.11	Postnatal depression
E290.00	Brief depressive reaction
E290z00	Brief depressive reaction NOS
E291.00	Prolonged depressive reaction
E2B..00	Depressive disorder NEC
E2B0.00	Post-viral depression
E2B1.00	Chronic depression
Eu32.00	[X]Depressive episode
Eu32.11	[X]Single episode of depressive reaction
Eu32.12	[X]Single episode of psychogenic depression
Eu32.13	[X]Single episode of reactive depression
Eu32000	[X]Mild depressive episode
Eu32100	[X]Moderate depressive episode
Eu32200	[X]Severe depressive episode without psychotic symptoms
Eu32211	[X]Single episode agitated depression w'out psychotic symptoms
Eu32212	[X]Single episode major depression w'out psychotic symptoms
Eu32300	[X]Severe depressive episode with psychotic symptoms
Eu32311	[X]Single episode of major depression and psychotic symptoms
Eu32312	[X]Single episode of psychogenic depressive psychosis
Eu32313	[X]Single episode of psychotic depression
Eu32314	[X]Single episode of reactive depressive psychosis
Eu32400	[X]Mild depression
Eu32y00	[X]Other depressive episodes
Eu32y11	[X]Atypical depression
Eu32z00	[X]Depressive episode, unspecified
Eu32z11	[X]Depression NOS
Eu32z12	[X]Depressive disorder NOS

Eu32z13	[X]Prolonged single episode of reactive depression
Eu32z14	[X] Reactive depression NOS
Eu33.00	[X]Recurrent depressive disorder
Eu33.11	[X]Recurrent episodes of depressive reaction
Eu33.12	[X]Recurrent episodes of psychogenic depression
Eu33.13	[X]Recurrent episodes of reactive depression
Eu33.14	[X]Seasonal depressive disorder
Eu33.15	[X]SAD - Seasonal affective disorder
Eu33000	[X]Recurrent depressive disorder, current episode mild
Eu33100	[X]Recurrent depressive disorder, current episode moderate
Eu33200	[X]Recurrent depressive disorder cur epi severe without psyc sympt
Eu33211	[X]Endogenous depression without psychotic symptoms
Eu33212	[X]Major depression, recurrent without psychotic symptoms
Eu33300	[X]Recurrent depress disorder cur epi severe with psychotic symptoms
Eu33311	[X]Endogenous depression with psychotic symptoms
Eu33313	[X]Recurr severe episodes/major depression+psychotic symptom
Eu33314	[X]Recurr severe episodes/psychogenic depressive psychosis
Eu33315	[X]Recurrent severe episodes of psychotic depression
Eu33316	[X]Recurrent severe episodes/reactive depressive psychosis
Eu33y00	[X]Other recurrent depressive disorders
Eu33z00	[X]Recurrent depressive disorder, unspecified
Eu33z11	[X]Monopolar depression NOS
Eu34100	[X]Dysthymia
Eu34111	[X]Depressive neurosis
Eu34113	[X]Neurotic depression
Eu34114	[X]Persistant anxiety depression
Eu3y111	[X]Recurrent brief depressive episodes
Eu41200	[X]Mixed anxiety and depressive disorder
Eu41211	[X]Mild anxiety depression
Eu53011	[X]Postnatal depression NOS
Eu53012	[X]Postpartum depression NOS
R007z13	[D]Postoperative depression

g) List of Read codes for cough

Read code	description
171..00	Cough
171..11	C/O - cough
1712.00	Dry cough
1713.00	Productive cough -clear sputum
1714.00	Productive cough -green sputum
1715.00	Productive cough-yellow sputum
1716.00	Productive cough NOS
1716.11	Coughing up phlegm
1717.00	Night cough present
1719.00	Chesty cough
1719.11	Bronchial cough
171A.00	Chronic cough
171B.00	Persistent cough
171C.00	Morning cough
171D.00	Evening cough
171E.00	Unexplained cough
171F.00	Cough with fever
171G.00	Bovine cough
171H.00	Difficulty in coughing up sputum
171J.00	Reflux cough
171K.00	Barking cough
171Z.00	Cough symptom NOS
173B.00	Nocturnal cough / wheeze
H310100	Smokers' cough
R062.00	[D]Cough

h) List of Read codes for haemoptysis

Read code	description
172..00	Blood in sputum - haemoptysis
172..11	Blood in sputum - symptom
172..12	Haemoptysis - symptom
4E24.00	Sputum: contains blood
4E35.00	Sputum: blood cells present
R063.00	[D]Haemoptysis
R063000	[D]Cough with haemorrhage
R063z00	[D]Haemoptysis NOS

i) List of Read codes for Dyspnoea

Read code	description
173..00	Breathlessness
173..11	Breathlessness symptom
173..12	Dyspnoea - symptom
173..13	Shortness of breath symptom
1732.00	Breathless - moderate exertion
1733.00	Breathless - mild exertion
1734.00	Breathless - at rest
1735.00	Breathless - lying flat
1735.11	Orthopnoea symptom
1736.00	Paroxysmal nocturnal dyspnoea
1738.00	Difficulty breathing
1739.00	Shortness of breath
173C.00	Short of breath on exertion
173C.11	Dyspnoea on exertion
173C.12	SOBOE
173D.00	Nocturnal dyspnoea
173F.00	Short of breath dressing/undressing
173G.00	Breathless - strenuous exertion
173I.00	MRC Breathlessness Scale: grade 2
173J.00	MRC Breathlessness Scale: grade 3
173K.00	MRC Breathlessness Scale: grade 4
173L.00	MRC Breathlessness Scale: grade 5
173N.00	Borg Breathlessness Score: 0.5 very, very slight
173P.00	Borg Breathlessness Score: 1 very slight
173Q.00	Borg Breathlessness Score: 2 slight
173R.00	Borg Breathlessness Score: 3 moderate
173S.00	Borg Breathlessness Score: 4 somewhat severe
173T.00	Borg Breathlessness Score: 5 severe
173V.00	Borg Breathlessness Score: 6 severe (+)
173W.00	Borg Breathlessness Score: 7 very severe
173X.00	Borg Breathlessness Score: 8 very severe (+)
173Y.00	Borg Breathlessness Score: 9 very, very sev (almost maximal)
173Z.00	Breathlessness NOS
173a.00	Borg Breathlessness Score: 10 maximal
173b.00	Unable to complete a sentence in one breath
2322.00	O/E - dyspnoea
2323.00	O/E - orthopnoea
2324.00	O/E - respiratory distress
2327.00	O/E - accessory resp.m's.used
232D.00	O/E - sternal recession
232E.00	O/E - intercostal recession
232F.00	O/E - subcostal recession
232G.00	O/E - suprasternal recession
R060600	[D]Respiratory distress
R060700	[D]Respiratory insufficiency
R060800	[D]Shortness of breath
R060A00	[D]Dyspnoea
R060D00	[D]Breathlessness

j) List of Read codes for weight loss

Read code	description
1623.00	Weight decreasing
1625.00	Abnormal weight loss
1625.11	Abnormal weight loss - symptom
1D1A.00	Complaining of weight loss
22A8.00	Weight loss from baseline weight
R032.00	[D]Abnormal loss of weight

k) List of Read codes for Lower Respiratory Tract Infections (LRTI)

Read code	description
H06..00	Acute bronchitis and bronchiolitis
H060.00	Acute bronchitis
H060.11	Acute wheezy bronchitis
H060000	Acute fibrinous bronchitis
H060100	Acute membranous bronchitis
H060200	Acute pseudomembranous bronchitis
H060300	Acute purulent bronchitis
H060400	Acute croupous bronchitis
H060500	Acute tracheobronchitis
H060600	Acute pneumococcal bronchitis
H060700	Acute streptococcal bronchitis
H060800	Acute haemophilus influenzae bronchitis
H060900	Acute neisseria catarrhalis bronchitis
H060A00	Acute bronchitis due to mycoplasma pneumoniae
H060B00	Acute bronchitis due to coxsackievirus
H060C00	Acute bronchitis due to parainfluenza virus
H060D00	Acute bronchitis due to respiratory syncytial virus
H060E00	Acute bronchitis due to rhinovirus
H060F00	Acute bronchitis due to echovirus
H060v00	Subacute bronchitis unspecified
H060w00	Acute viral bronchitis unspecified
H060x00	Acute bacterial bronchitis unspecified
H060z00	Acute bronchitis NOS
H061.00	Acute bronchiolitis
H061000	Acute capillary bronchiolitis
H061100	Acute obliterating bronchiolitis
H061200	Acute bronchiolitis with bronchospasm
H061300	Acute exudative bronchiolitis
H061400	Obliterating fibrous bronchiolitis
H061500	Acute bronchiolitis due to respiratory syncytial virus
H061600	Acute bronchiolitis due to other specified organisms
H061z00	Acute bronchiolitis NOS
H062.00	Acute lower respiratory tract infection
H06z.00	Acute bronchitis or bronchiolitis NOS
H06z100	Lower resp tract infection
H06z112	Acute lower respiratory tract infection
H2...00	Pneumonia and influenza
H20..00	Viral pneumonia
H20..11	Chest infection - viral pneumonia
H200.00	Pneumonia due to adenovirus
H201.00	Pneumonia due to respiratory syncytial virus
H202.00	Pneumonia due to parainfluenza virus
H20y.00	Viral pneumonia NEC
H20z.00	Viral pneumonia NOS
H21..00	Lobar (pneumococcal) pneumonia
H21..11	Chest infection - pneumococcal pneumonia
H22..00	Other bacterial pneumonia
H22..11	Chest infection - other bacterial pneumonia
H220.00	Pneumonia due to klebsiella pneumoniae
H221.00	Pneumonia due to pseudomonas
H222.00	Pneumonia due to haemophilus influenzae
H222.11	Pneumonia due to haemophilus influenzae
H223.00	Pneumonia due to streptococcus
H223000	Pneumonia due to streptococcus, group B
H224.00	Pneumonia due to staphylococcus
H22y.00	Pneumonia due to other specified bacteria
H22y000	Pneumonia due to escherichia coli
H22y011	E.coli pneumonia
H22y100	Pneumonia due to proteus
H22y200	Pneumonia - Legionella
H22yX00	Pneumonia due to other aerobic gram-negative bacteria
H22yz00	Pneumonia due to bacteria NOS
H22z.00	Bacterial pneumonia NOS
H23..00	Pneumonia due to other specified organisms
H23..11	Chest infection - pneumonia organism OS

H230.00	Pneumonia due to Eaton's agent
H231.00	Pneumonia due to mycoplasma pneumoniae
H232.00	Pneumonia due to pleuropneumonia like organisms
H233.00	Chlamydial pneumonia
H23z.00	Pneumonia due to specified organism NOS
H24..00	Pneumonia with infectious diseases EC
H240.00	Pneumonia with measles
H241.00	Pneumonia with cytomegalic inclusion disease
H242.00	Pneumonia with ornithosis
H243.00	Pneumonia with whooping cough
H243.11	Pneumonia with pertussis
H244.00	Pneumonia with tularaemia
H245.00	Pneumonia with anthrax
H246.00	Pneumonia with aspergillosis
H247.00	Pneumonia with other systemic mycoses
H247000	Pneumonia with candidiasis
H247100	Pneumonia with coccidioidomycosis
H247200	Pneumonia with histoplasmosis
H247z00	Pneumonia with systemic mycosis NOS
H24y.00	Pneumonia with other infectious diseases EC
H24y000	Pneumonia with actinomycosis
H24y100	Pneumonia with nocardiasis
H24y200	Pneumonia with pneumocystis carinii
H24y300	Pneumonia with Q-fever
H24y400	Pneumonia with salmonellosis
H24y500	Pneumonia with toxoplasmosis
H24y600	Pneumonia with typhoid fever
H24y700	Pneumonia with varicella
H24yz00	Pneumonia with other infectious diseases EC NOS
H24z.00	Pneumonia with infectious diseases EC NOS
H25..00	Bronchopneumonia due to unspecified organism
H25..11	Chest infection - unspecified bronchopneumonia
H26..00	Pneumonia due to unspecified organism
H26..11	Chest infection - pnemonia due to unspecified organism
H260.00	Lobar pneumonia due to unspecified organism
H261.00	Basal pneumonia due to unspecified organism
H262.00	Postoperative pneumonia
H270.00	Influenza with pneumonia
H270.11	Chest infection - influenza with pneumonia
H270000	Influenza with bronchopneumonia
H270100	Influenza with pneumonia, influenza virus identified
H270z00	Influenza with pneumonia NOS
H28..00	Atypical pneumonia
H2y..00	Other specified pneumonia or influenza
H2z..00	Pneumonia or influenza NOS
H30..00	Bronchitis unspecified
H30..11	Chest infection - unspecified bronchitis
H30..12	Recurrent wheezy bronchitis
H300.00	Tracheobronchitis NOS
H302.00	Wheezy bronchitis
H30z.00	Bronchitis NOS
H31..00	Chronic bronchitis
H310.00	Simple chronic bronchitis
H310000	Chronic catarrhal bronchitis
H310z00	Simple chronic bronchitis NOS
H311.00	Mucopurulent chronic bronchitis
H311000	Purulent chronic bronchitis
H311100	Fetid chronic bronchitis
H311z00	Mucopurulent chronic bronchitis NOS
H312.00	Obstructive chronic bronchitis
H312000	Chronic asthmatic bronchitis
H312011	Chronic wheezy bronchitis
H312100	Emphysematous bronchitis
H312300	Bronchiolitis obliterans
H312z00	Obstructive chronic bronchitis NOS
H313.00	Mixed simple and mucopurulent chronic bronchitis
H31y.00	Other chronic bronchitis
H31y000	Chronic tracheitis

H31y100	Chronic tracheobronchitis
H31yz00	Other chronic bronchitis NOS
H31z.00	Chronic bronchitis NOS
H530200	Gangrenous pneumonia
H530300	Abscess of lung with pneumonia
H540000	Hypostatic pneumonia
H540100	Hypostatic bronchopneumonia
H564.00	Bronchiolitis obliterans organising pneumonia
H56y.00	Other alveolar and parietoalveolar disease
H56y000	Endogenous lipoid pneumonia
H56y100	Interstitial pneumonia
H571.00	Rheumatic pneumonia
Hyu0800	[X]Other viral pneumonia
Hyu0900	[X]Pneumonia due to other aerobic gram-negative bacteria
Hyu0A00	[X]Other bacterial pneumonia
Hyu0B00	[X]Pneumonia due to other specified infectious organisms
Hyu0C00	[X]Pneumonia in bacterial diseases classified elsewhere
Hyu0D00	[X]Pneumonia in viral diseases classified elsewhere
Hyu0E00	[X]Pneumonia in mycoses classified elsewhere
Hyu0F00	[X]Pneumonia in parasitic diseases classified elsewhere
Hyu0G00	[X]Pneumonia in other diseases classified elsewhere
Hyu0H00	[X]Other pneumonia, organism unspecified
Hyu1.00	[X]Other acute lower respiratory infections
Hyu1000	[X]Acute bronchitis due to other specified organisms
Hyu1100	[X]Acute bronchiolitis due to other specified organisms
Hyu3.00	[X]Chronic lower respiratory diseases

l) List of Read codes for Upper Respiratory Tract Infections (URTI)

Read code	description
H00..00	Acute nasopharyngitis
H04..00	Acute laryngitis and tracheitis
H05..00	Other acute upper respiratory infections
H050.00	Acute laryngopharyngitis
H051.00	Acute upper respiratory tract infection
H052.00	Pharyngotracheitis
H053.00	Tracheopharyngitis
H054.00	Recurrent upper respiratory tract infection
H055.00	Pharyngolaryngitis
H05y.00	Other upper respiratory infections of multiple sites
H05z.00	Upper respiratory infection NOS
H05z.11	Upper respiratory tract infection NOS
H05z.12	Viral upper respiratory tract infection NOS
H1...00	Other upper respiratory tract diseases
H12..00	Chronic pharyngitis and nasopharyngitis
H13..00	Chronic sinusitis
H13..11	Chronic rhinosinusitis
H14..00	Chronic tonsil and adenoid disease
H14..11	Adenoid disease - chronic
H14..12	Tonsil disease - chronic
H15..00	Peritonsillar abscess - quinsy
H15..11	Quinsy
H16..00	Chronic laryngitis and laryngotracheitis
H17..00	Allergic rhinitis
H17..11	Perennial rhinitis
H17..12	Allergic rhinosinusitis
H18..00	Vasomotor rhinitis
H1y..00	Other specified diseases of upper respiratory tract
H1y1.12	Nasal vestibulitis
H1y2.00	Other pharyngeal disease NEC
H1y2.11	Other nasopharyngeal disease NEC
H1y7.00	Other diseases of larynx NEC
H1yz.00	Other upper respiratory tract diseases NOS
H1yz000	Abscess of trachea
H1yzz00	Other upper respiratory tract disease NOS
H1z..00	Upper respiratory tract disease NOS
H271000	Influenza with laryngitis
H271100	Influenza with pharyngitis
H301.00	Laryngotracheobronchitis
Hyu0.00	[X]Acute upper respiratory infections
Hyu0000	[X]Other acute sinusitis
Hyu0100	[X]Acute pharyngitis due to other specified organisms
Hyu0200	[X]Acute tonsillitis due to other specified organisms
Hyu0300	[X]Other acute upper respiratory infections/multiple sites
Hyu0400	[X]Flu+oth respiratory manifestations,'flu virus identified
Hyu0500	[X]Influenza+other manifestations,influenza virus identified
Hyu0600	[X]Influenza+oth respiratory manifestatns,virus not identifd
Hyu0700	[X]Influenza+other manifestations, virus not identified
Hyu2.00	[X]Other diseases of the upper respiratory tract
Hyu2000	[X]Other seasonal allergic rhinitis
Hyu2100	[X]Other allergic rhinitis
Hyu2200	[X]Other chronic sinusitis
Hyu2500	[X]Other chronic diseases of tonsils and adenoids
Hyu2700	[X]Other diseases of larynx
Hyu2800	[X]Other abscess of pharynx
Hyu2900	[X]Other diseases of pharynx
Hyu2A00	[X]Other specified diseases of upper respiratory tract

m) List of Read codes for non-specific chest infections

Read code	description
H0...00	Acute respiratory infections
H06z000	Chest infection NOS
H06z011	Chest infection
H06z111	Respiratory tract infection
H06z200	Recurrent chest infection
H07..00	Chest cold
H0y..00	Other specified acute respiratory infections
H0z..00	Acute respiratory infection NOS
H20y000	Severe acute respiratory syndrome
H24..11	Chest infection with infectious disease EC
H271.00	Influenza with other respiratory manifestation
H271z00	Influenza with respiratory manifestations NOS
H27y.00	Influenza with other manifestations
H5yy.11	Respiratory infection NOS

n) List of Read codes for chest/shoulder pain

Read code	description
182..00	Chest pain
1822.00	Central chest pain
1823.00	Precordial pain
1824.00	Anterior chest wall pain
1825.00	Pleuritic pain
1826.00	Parasternal pain
1827.00	Painful breathing -pleurodynia
1828.00	Atypical chest pain
1829.00	Retrosternal pain
182A.00	Chest pain on exertion
182B.00	Rib pain
182B000	Costal margin chest pain
182C.00	Chest wall pain
182Z.00	Chest pain NOS
1D22000	Chest wall tenderness
8HTG.00	Referred to acute chest pain clinic
8HTJ.00	Referral to rapid access chest pain clinic
9N0f.00	Seen in rapid access chest pain clinic
G33z400	Ischaemic chest pain
N094111	Shoulder joint pain
N245.17	Shoulder pain
N245700	Shoulder pain
R065.00	[D]Chest pain
R065000	[D]Chest pain, unspecified
R065011	[D] Retrosternal chest pain
R065200	[D]Anterior chest wall pain
R065300	[D]Painful respiration NOS
R065400	[D]Pleuritic pain
R065600	[D]Chest discomfort
R065700	[D]Chest pressure
R065800	[D]Chest tightness
R065900	[D]Parasternal chest pain
R065A00	[D]Musculoskeletal chest pain
R065B00	[D]Non cardiac chest pain
R065B14	[D]Non-cardiac chest pain
R065C00	[D]Retrosternal chest pain
R065D00	[D]Central chest pain
R065z00	[D]Chest pain NOS
Ryu0400	[X]Other chest pain

o) List of Read codes for voice hoarseness

Read code	description
1CA..00	Hoarseness symptom
1CA..11	Hoarseness - throat symptom
1CA2.00	Hoarse
1CA2.11	Voice hoarseness
1CAZ.00	Hoarseness symptom NOS
2DE4.00	O/E - hoarseness
2DE5.00	O/E - dysphonia
R044300	[D]Change in voice
R044400	[D]Dysphonia
R044500	[D]Hoarseness
ZS2..00	Disorder of voice
ZS21.00	Dysphonia
ZT15.00	Change in voice

p) List of Read codes for chest x-rays

Read code	description
535..00	Standard chest X-ray
535..11	Chest X-ray - routine
5351.00	Standard chest X-ray requested
5352.00	Standard chest X-ray normal
5352.11	Chest X-ray normal
5353.00	Standard chest X-ray abnormal
535Z.00	Standard chest X-ray NOS
536..00	Soft tissue X-ray chest
5361.00	Soft tiss.X-ray chest normal
5362.00	Soft tiss.X-ray chest abnormal
5363.00	X-ray larynx/trachea
5363.11	Larynx soft tis. X-ray
5363.12	Trachea soft tis. X-ray
5364.00	Soft tiss.X-ray lung/bronchus
5364.11	Bronchus soft tis.X-ray
5364.12	Lung soft tis. X-ray
5365.00	Soft tissue X-ray chest wall
536Z.00	Soft tissue X-ray chest NOS
545..11	Bronchography
5451.00	Bronchography requested
5452.00	Bronchography normal
5453.00	Bronchography abnormal
5454.00	Contrast radiog.larynx/trachea
5454.11	Larynx - contrast radiography
5454.12	Trachea - contrast radiography
5455.00	Bilat.transglot.bronchography
5456.00	Bilat.transcric bronchography
5457.00	Selective bronchography
545Z.00	Resp.contrast radiogr.NOS
5661.00	Serial radiography of lungs
68C1.00	Screening chest X-ray
68C1.11	CXR - screening
7P04200	Plain x-ray of chest
7P04y00	Other specified diagnostic imaging of chest
7P04z00	Diagnostic imaging of chest NOS
ZV72511	[V]Routine chest X-ray

q) List of Read codes for blood investigations

Read code	description	Blood test status
4131.00	Blood test requested	blood test
4142.00	Blood sample -> Haematology Lab	blood test
4143.00	Blood sample -> Biochemistry Lab	blood test
4144.00	Blood sample -> Microbiology Lab	blood test
4145.00	Blood sample -> Lab NOS	blood test
41D0.00	Blood sample taken	blood test
421..00	Haematology - general	blood test
4212.00	Haematology test performed	blood test
4213.00	Haematology test requested	blood test
4214.00	Blood sent for haematological test	blood test
4217.00	Haematology res. not back yet	blood test
4218.00	Haematology result normal	normal
4219.00	Haematology result abnormal	abnormal
421A.00	Haematology result borderline	normal
423..00	Haemoglobin estimation	blood test
423..11	Hb estimation	blood test
4232.00	Haemoglobin requested	blood test
4233.00	Haemoglobin - sample sent	blood test
4234.00	Haemoglobin very low	abnormal
4235.00	Haemoglobin low	abnormal
4236.00	Haemoglobin borderline low	abnormal
4237.00	Haemoglobin normal	normal
4238.00	Haemoglobin borderline high	abnormal
4239.00	Haemoglobin high	abnormal
423A.00	Haemoglobin very high	abnormal
423B.00	Haemoglobin abnormal	abnormal
423Z.00	Haemoglobin estimation NOS	blood test
424..00	Full blood count - FBC	blood test
4241.00	Full blood count normal	normal
4242.00	Full blood count borderline	normal
4243.00	Full blood count abnormal	abnormal
424Z.00	Full blood count NOS	blood test
425..00	Haematocrit - PCV	blood test
425..11	Packed cell volume - PCV	blood test
4251.00	Haematocrit - PCV - normal	normal
4252.00	Haematocrit - borderline high	abnormal
4253.00	Haematocrit - PCV - high	abnormal
4254.00	Haematocrit - PCV - low	abnormal
4255.00	Haematocrit - borderline low	abnormal
4256.00	Haematocrit - PCV - abnormal	abnormal
4257.00	Packed cell volume	blood test
4258.00	Haematocrit	blood test
425Z.00	Haematocrit - PCV - NOS	blood test
426..00	Red blood cell (RBC) count	blood test
4261.00	RBC count normal	normal
4262.00	RBC count borderline low	abnormal
4263.00	RBC count low	abnormal
4264.00	RBC count raised	abnormal
4265.00	RBC count borderline raised	abnormal
4266.00	Nucleated red blood cell count	blood test
4267.00	RBC count abnormal	abnormal
426Z.00	RBC count NOS	blood test
428..00	Mean corpusc. haemoglobin(MCH)	blood test
428..11	Mean cell haemoglobin	blood test
4281.00	MCH - normal	normal
4282.00	MCH - borderline low	abnormal
4283.00	MCH - low	abnormal
4284.00	MCH - raised	abnormal
4285.00	MCH - borderline raised	abnormal
4286.00	MCH - abnormal	abnormal
428Z.00	MCH - NOS	blood test
429..00	Mean corpuscular Hb. conc. (MCHC)	blood test
4291.00	MCHC - normal	normal
4292.00	MCHC - borderline low	abnormal
4293.00	MCHC - low	abnormal

4294.00	MCHC - raised	abnormal
4295.00	MCHC - borderline raised	abnormal
429Z.00	MCHC - NOS	blood test
42A..00	Mean corpuscular volume (MCV)	blood test
42A..11	Mean cell volume	blood test
42A1.00	MCV - normal	normal
42A2.00	MCV - borderline raised	abnormal
42A3.00	MCV - raised	abnormal
42A4.00	MCV - low	abnormal
42A5.00	MCV - borderline low	abnormal
42AZ.00	MCV - NOS	blood test
42B..00	Plasma viscosity	blood test
42B..11	Plasma viscosity - PV	blood test
42B1.00	Plasma viscosity normal	normal
42B2.00	Plasma visc. borderline raised	abnormal
42B3.00	Plasma viscosity raised	abnormal
42B4.00	Plasma viscosity low	abnormal
42B5.00	Plasma visc. borderline low	abnormal
42B6.00	Erythrocyte sedimentation rate	blood test
42B6000	ESR abnormal	abnormal
42B6100	ESR low	abnormal
42B6200	ESR normal	normal
42B6300	ESR raised	abnormal
42B6z00	Erythrocyte sediment rate NOS	blood test
42BZ.00	Plasma viscosity NOS	blood test
42C..00	RBC - red blood cell size	blood test
42C1.00	Red blood cell size normal	normal
42C2.00	RBC's - microcytic	abnormal
42C3.00	RBC's - macrocytic	abnormal
42CZ.00	Red blood cell size NOS	blood test
42D..00	RBC - red blood cell shape	blood test
42D1.00	Red blood cell shape - normal	normal
42G..00	Red blood cell enzymes	blood test
42G1.00	Red blood cell enzymes normal	normal
42G2.00	RBC enzymes abnormal	abnormal
42H..00	Total white cell count	blood test
42H..11	White blood count	blood test
42H..12	White cell count	blood test
42H1.00	White cell count normal	normal
42H2.00	Leucopenia - low white count	abnormal
42H2.11	Leucopenia	abnormal
42H3.00	Leucocytosis -high white count	abnormal
42H3.11	Leucocytosis	abnormal
42H4.00	Agranulocytosis	abnormal
42H5.00	White cell count abnormal	abnormal
42H6.00	Polymorphonuclear leukocyte count	blood test
42H7.00	Total white blood count	blood test
42H8.00	Total WBC (IMM)	blood test
42HZ.00	Total white cell count NOS	blood test
42I..00	Differential white cell count	blood test
42I..11	WCC - differential	blood test
42I1.00	Diff. white cell count normal	normal
42I2.00	Diff. white count abnormal	abnormal
42IZ.00	Diff. white cell count NOS	blood test
42J..00	Neutrophil count	blood test
42J..11	Granulocyte count	blood test
42J1.00	Neutrophil count normal	normal
42J2.00	Neutropenia	abnormal
42J3.00	Neutrophilia	abnormal
42J4.00	Neutrophil count abnormal	abnormal
42JZ.00	Neutrophil count NOS	blood test
42K..00	Eosinophil count	blood test
42K1.00	Eosinophil count normal	normal
42K2.00	Eosinopenia	abnormal
42K3.00	Eosinophil count raised	abnormal
42KZ.00	Eosinophil count NOS	blood test
42L..00	Basophil count	blood test
42L1.00	Basophil count normal	normal

42L2.00	Basophilia	abnormal
42L3.00	Basophil count abnormal	abnormal
42LZ.00	Basophil count NOS	blood test
42M..00	Lymphocyte count	blood test
42M1.00	Lymphocyte count normal	normal
42M2.00	Lymphocytosis - absolute	blood test
42M3.00	Lymphocytosis - relative	blood test
42M4.00	Abnormal lymphocytes	abnormal
42M5.00	Lymphocyte count abnormal	abnormal
42M6.00	Total T lymphocyte count	blood test
42M7.00	T cell subsets	blood test
42M8.00	Total lymphocyte count (IMM)	blood test
42M9.00	Total B lymphocyte count	blood test
42MA.00	Lymphocyte subsets	blood test
42MB.00	Natural killer cell level	blood test
42MC.00	Prolymphocyte count	blood test
42MD.00	Reactive lymphocyte count	blood test
42ME.00	Hairy cell markers	blood test
42MF.00	Lymphocyte function test	blood test
42MG.00	Leucocyte count	blood test
42MH.00	Population gated lymphocytes	blood test
42MZ.00	Lymphocyte count NOS	blood test
42N..00	Monocyte count	blood test
42N1.00	Monocyte count normal	normal
42N2.00	Monocyte count raised	abnormal
42N3.00	Monocytopenia	abnormal
42N4.00	Abnormal monocytes	abnormal
42N5.00	Monocyte count abnormal	abnormal
42N6.00	Absolute atypical mononuclear cell count	blood test
42N7.00	Percentage atypical mononuclear cell count	blood test
42NZ.00	Monocyte count NOS	blood test
42O..00	Immature white blood cells	blood test
42O1.00	Immature WBC's - non present	normal
42P..00	Platelet count	blood test
42P1.00	Platelet count normal	normal
42P2.00	Thrombocytopenia	abnormal
42P2.11	Auto-immune thrombocytopenia	abnormal
42P3.00	Thrombocythaemia	abnormal
42P4.00	Platelet count abnormal	abnormal
42P5.00	Platelet distribution width	blood test
42P6.00	Platelet/neutrophil ratio	blood test
42P7.00	Percentage reticulated platelet count	blood test
42P8.00	Heparin induced thrombocytopenia screening test	blood test
42P9.00	Plateletcrit	blood test
42PZ.00	Platelet count NOS	blood test
42Q..00	Coagulation/bleeding tests	blood test
42Q..11	Bleeding tests	blood test
42Q..12	Clotting tests	blood test
42Q..13	Coagulation tests	blood test
42Q1.00	Coag./bleeding tests normal	normal
42Q2.00	Coag./bleeding tests abnormal	abnormal
42Q3.00	Bleeding time	blood test
42Q4.00	Whole blood clotting time	blood test
42Q5.00	Prothrombin time	blood test
42Q5000	Prothrombin time abnormal	abnormal
42Q5100	Prothrombin time low	abnormal
42Q5200	Prothrombin time normal	normal
42Q6.00	Partial thromboplastin time	blood test
42Q7.00	Heparin assay	blood test
42Q8.00	Thrombin time	blood test
42Q8000	Thrombin time normal	normal
42Q8100	Thrombin time abnormal	abnormal
42Q9.00	Fibrinogen assay/titre	blood test
42QA.00	Fibrinogen degradation products	blood test
42QB.00	Factor VIII assay	blood test
42QB.11	Plasma factor VIII level	blood test
42QC.00	Factor IX assay	blood test
42QD.00	Serum vitamin K	blood test

42QE.00	International normalised ratio	blood test
42QE000	INR - international normal ratio normal	normal
42QE100	INR - international normal ratio abnormal	abnormal
42QF.00	Plasma total protein S level	blood test
42QG.00	Plasma free:total protein S ratio	blood test
42QH.00	Plasma free protein S level	blood test
42QI.00	Plasma ristocetin cofactor level	blood test
42QI.11	Plasma von Willebrand factor level	blood test
42QJ.00	Plasma antithrombin III level	blood test
42QK.00	Plasma plasminogen level	blood test
42QL.00	Plasma factor VIII related antigen test	blood test
42QM.00	Plasma factor XII level	blood test
42QN.00	Plasma factor XI level	blood test
42QO.00	Plasma factor X level	blood test
42QP.00	Plasma factor VII level	blood test
42QQ.00	Plasma antithrombin III antigen level	blood test
42QR.00	Plasma factor V level	blood test
42QS.00	Clotting screen	blood test
42QT.00	Plasma factor XIII screening test	blood test
42QU.00	Euglobulin clot lysis time	blood test
42QV.00	Thrombophilia screen	blood test
42QW.00	Kaolin cephalin clotting time	blood test
42QX.00	Dilute Russell viper venom ratio	blood test
42QY.00	Ivy bleeding time	blood test
42QZ.00	Coag./bleeding test NOS	blood test
42Qa.00	Protein C function estimate	blood test
42Qb.00	Protein S function estimate	blood test
42Qc.00	Plasma activated protein C resistance	blood test
42Qd.00	Plasma protein C antigen level	blood test
42Qe.00	Factor V Leiden genotype	blood test
42Qf.00	D-Dimer level	blood test
42Qg.00	Factor II level	blood test
42Qh.00	Factor IX inhibitor activity	blood test
42Qi.00	Factor IX related antigen level	blood test
42Qj.00	Factor VIII inhibitor activity	blood test
42Qk.00	Factor VIII related antigen level	blood test
42Ql.00	Factor VIII von Willebrands Factor ratio	blood test
42Qm.00	Factor VIIIc level	blood test
42Qn.00	Fibrinogen level	blood test
42Qo.00	High molecular weight kininogen level	blood test
42Qp.00	Prekallikrein level	blood test
42Qq.00	Protein C level	blood test
42Qr.00	Prothrombin consumption	blood test
42Qs.00	von Willebrand factor level	blood test
42Qt.00	Partial thromboplastin time ratio	blood test
42Qu.00	Activated partial thromboplastin time ratio	blood test
42Qv.00	Prothrombin time - reference	blood test
42Qw.00	APTT - reference	blood test
42Qx.00	von Willebrand factor activity	blood test
42Qy.00	Thrombin time reference	blood test
42Qz.00	APTR actin FSL ratio	blood test
42R..00	Serum iron tests	blood test
42R..11	Serum iron level	blood test
42R1.00	Serum iron normal	normal
42R2.00	Serum iron low	abnormal
42R3.00	Serum iron raised	abnormal
42R4.00	Serum ferritin	blood test
42R4.11	Ferritin - serum	blood test
42R4.12	TIBC - serum	blood test
42R4100	Ferritin level low	abnormal
42R4200	Serum ferritin normal	normal
42R4300	Serum ferritin high	abnormal
42R5.00	Serum TIBC	blood test
42R5000	TIBC - Total iron binding capacity normal	normal
42R5100	TIBC - Total iron binding capacity low	abnormal
42R6.00	Serum iron abnormal	abnormal
42R7.00	Serum iron level	blood test
42R8.00	Unsaturated iron binding capacity	blood test

42R9.00	Saturation of iron binding capacity	blood test
42RA.00	Percentage iron saturation	blood test
42RZ.00	Serum iron tests NOS	blood test
42S..00	Iron kinetics	blood test
42S1.00	Iron kinetics normal	normal
42S2.00	Iron kinetics abnormal	abnormal
42S3.00	Iron absorption	blood test
42S4.00	Iron clearance	blood test
42S5.00	Iron utilisation	blood test
42SZ.00	Iron kinetics NOS	blood test
42T..00	Serum vitamin B12	blood test
42T1.00	Serum vitamin B12 normal	normal
42T2.00	Serum vitamin B12 low	abnormal
42T3.00	Serum vit B12 borderline	normal
42TZ.00	Serum vitamin B12 NOS	blood test
42U..00	Blood folate	blood test
42U..11	Folate blood level	blood test
42U1.00	Serum folate normal	normal
42U2.00	Serum folate low	abnormal
42U3.00	Serum folate borderline	normal
42U4.00	Red blood cell folate	blood test
42U4.11	Folate - RBC	blood test
42U5.00	Serum folate	blood test
42U6.00	Whole blood folate	blood test
42U7.00	RBC folate normal	normal
42U8.00	RBC folate low	abnormal
42U9.00	RBC folate borderline	normal
42UA.00	Whole blood folate normal	normal
42UB.00	Whole blood folate low	abnormal
42UC.00	Whole blood folate borderline	normal
42UD.00	RBC folate abnormal	abnormal
42UE.00	Plasma folate level	blood test
42UZ.00	Blood folate NOS	blood test
42V..00	Haemoglobin variants	blood test
42V1.00	Haemoglobin electrophoresis	blood test
42V1.11	Electrophoresis - Hb	blood test
42ZZ.00	Haematology NOS	blood test
42a..00	Plasma cell count	blood test
42a0.00	Percentage plasma cell count	blood test
42b..00	Percentage cell count	blood test
42b0.00	Percentage neutrophils	blood test
42b1.00	Percentage lymphocytes	blood test
42b2.00	Percentage monocytes	blood test
42b3.00	Percentage basophils	blood test
42b4.00	Percentage metamyelocytes	blood test
42b5.00	Percentage blast cells	blood test
42b6.00	Percentage smear cells	blood test
42b7.00	Percentage granulocytes	blood test
42b8.00	Percentage nucleated Red Blood Cells	blood test
42b9.00	Percentage eosinophils	blood test
42bA.00	Percentage myelocyte count	blood test
42bB.00	Percentage promyelocyte count	blood test
42bC.00	Percentage reticulocyte count	blood test
42bD.00	T cell total %	blood test
42bE.00	Percentage hypochromic cells	blood test
42f..00	Hess test	blood test
42g..00	Haematology test	blood test
42g0.00	Whole blood viscosity	blood test
43F..00	Rheumatoid factor	blood test
43F..11	Latex test	blood test
43F..12	Rose Waaler test	blood test
43F1.00	Rheumatoid factor positive	abnormal
43F2.00	Rheumatoid factor negative	normal
43F3.00	R.A. latex test	blood test
43F4.00	Rose Waaler test - sheep cells	blood test
43F4000	Heterophile agglutin test normal	normal
43F4100	Heterophile agglutin test abnormal	abnormal
43F5.00	Serum rheumatoid antigen level	blood test

43F6.00	Fluid rheumatoid factor level	blood test
43F7.00	Rheumatoid factor screening test	blood test
43F8.00	Serum rheumatoid antibody level	blood test
43F9.00	Rheumatoid factor IgG level	blood test
43FA.00	Rheumatoid factor IgM level	blood test
43FB.00	IgA rheumatoid factor level	blood test
43FZ.00	Rheumatoid factor NOS	blood test
43G..00	Autoantibody titres	blood test
43G1.00	Anti-nuclear factor	blood test
43G1.11	Anti-nuclear antibody	blood test
43G1000	Anti-nuclear factor positive	abnormal
43G1011	Anti-nuclear antibody positive	abnormal
43G1100	Anti-nuclear factor negative	normal
43G1111	Anti-nuclear antibody negative	normal
43G1200	Anti-nuclear factor weakly positive	abnormal
43G1211	Anti-nuclear antibody weakly positive	abnormal
43G2.00	Antimitochondrial autoantibod.	blood test
43G3.00	Anti smooth muscle autoantibod	blood test
43G3000	Smooth muscle antibodies negative	normal
43G3100	Smooth muscle antibodies positive	abnormal
43G3200	Smooth muscle antibodies weakly positive	abnormal
43G4.00	Parietal cell autoantibodies	blood test
43G4000	Parietal cell antibodies negative	normal
43G4100	Parietal cell antibodies positive	abnormal
43G4200	Parietal cell antibodies weakly positive	abnormal
44...00	Blood chemistry	blood test
441..00	Blood chemistry - general	blood test
4411.00	Blood sent for chemistry	blood test
4412.00	Blood chemistry normal	normal
4412000	Urea and electrolytes normal	normal
4412100	Urea and electrolytes abnormal	abnormal
4413.00	Blood chemistry abnormal	abnormal
441Z.00	Blood chemistry - general NOS	blood test
44D..00	Liver function tests - general	blood test
44D..11	Liver function tests	blood test
44D1.00	Liver function tests normal	normal
44D2.00	Liver function tests abnormal	abnormal
44D6.00	Liver function test	blood test
44DZ.00	Liver function tests NOS	blood test
44E..00	Serum bilirubin level	blood test
44E1.00	Serum bilirubin normal	normal
44E2.00	Serum bilirubin raised	abnormal
44E3.00	Total bilirubin	blood test
44E4.00	Direct (conjugated) bilirubin	blood test
44E5.00	Indirect (unconj.) bilirubin	blood test
44E6.00	Serum bilirubin borderline	normal
44E7.00	Serum conjugated:total bilirubin ratio	blood test
44E8.00	Plasma conjugated bilirubin level	blood test
44E9.00	Plasma total bilirubin level	blood test
44EA.00	Plasma unconjugated bilirubin level	blood test
44EB.00	Serum conjugated bilirubin level	blood test
44EC.00	Serum total bilirubin level	blood test
44ED.00	Serum unconjugated bilirubin level	blood test
44EZ.00	Serum bilirubin NOS	blood test
44F..00	Serum alkaline phosphatase	blood test
44F1.00	Serum alk. phos. normal	normal
44F2.00	Serum alk. phos. raised	abnormal
44F3.00	Total alkaline phosphatase	blood test
44F4.00	Alk. phos. - liver isoenzyme	blood test
44F5.00	Alk. phos. - bone isoenzyme	blood test
44F5000	Alkaline phosphatase bone isoenzyme raised	abnormal
44F6.00	Alk. phos. - bile isoenzyme	blood test
44F7.00	Alkaline phosphatase isoenzyme studies	blood test
44F8.00	Plasma alkaline phosphatase bile isoenzyme level	blood test
44F9.00	Plasma alkaline phosphatase bone isoenzyme level	blood test
44FA.00	Plasma alkaline phosphatase liver isoenzyme level	blood test
44FB.00	Serum alkaline phosphatase bile isoenzyme level	blood test
44FC.00	Serum alkaline phosphatase bone isoenzyme level	blood test

44FD.00	Serum alkaline phosphatase liver isoenzyme level	blood test
44FE.00	Serum alkaline phosphatase electrophoresis	blood test
44FG.00	Alkaline phosphatase - bile isoenzyme level	blood test
44FH.00	Alkaline phosphatase liver isoenzyme level	blood test
44FI.00	Alkaline phosphatase - bone isoenzyme level	blood test
44FJ.00	Heat stable alkaline phosphatase measurement	blood test
44FZ.00	Serum alkaline phosphatase NOS	blood test
44G..00	Liver enzymes	blood test
44G..11	ALT - blood level	blood test
44G..12	SGPT - blood level	blood test
44G1.00	Liver enzymes normal	normal
44G2.00	Liver enzymes abnormal	abnormal
44G3.00	ALT/SGPT serum level	blood test
44G3000	ALT/SGPT level normal	normal
44G3100	ALT/SGPT level abnormal	abnormal
44G4.00	Gamma - G.T. level	blood test
44G4000	Gamma glutamyl transferase level normal	normal
44G4100	Gamma glutamyl transferase level abnormal	abnormal
44G5.00	Serum 5 - nucleotidase	blood test
44G5000	Serum 5-nucleotidase level normal	normal
44G5100	Serum 5-nucleotidase level low	abnormal
44G5200	Serum 5-nucleotidase level raised	abnormal
44G6.00	Plasma hydroxybutyrate dehydrogenase level	blood test
44G7.00	Plasma gamma-glutamyl transferase level	blood test
44G8.00	Serum hydroxybutyrate dehydrogenase level	blood test
44G9.00	Serum gamma-glutamyl transferase level	blood test
44GA.00	Plasma alanine aminotransferase level	blood test
44GB.00	Serum alanine aminotransferase level	blood test
44GZ.00	Liver enzymes NOS	blood test
44H..00	Cardiac enzymes	blood test
44H1.00	Blood sent: cardiac enzymes	blood test
44H2.00	Cardiac enzymes normal	normal
44H3.00	Cardiac enzymes abnormal	abnormal
44H3000	Cardiac enzymes abnormal - first set	abnormal
44H4.00	CK - creatine kinase level	blood test
44H4.11	Creatine phosphokinase level	blood test
44H5.00	AST - aspartate transam.(SGOT)	blood test
44H5.11	AST serum level	blood test
44H5.12	SGOT serum level	blood test
44H5000	AST/SGOT level normal	normal
44H5100	AST/SGOT level abnormal	abnormal
44H5200	AST/SGOT level raised	abnormal
44H6.00	LDH (HBD) level	blood test
44H6.11	LDH blood level	blood test
44H6.12	Serum total lactate dehydrogenase level	blood test
44H7.00	Cardiac enzymes equivocal	blood test
44H8.00	Serum creatinine phosphokinase MB isoenzyme level	blood test
44H9.00	Total lactic dehydrogenase	blood test
44HA.00	Serum total lactate dehydrogenase level	blood test
44HB.00	AST serum level	blood test
44HB.11	SGOT serum level	blood test
44HC.00	Plasma aspartate transaminase level	blood test
44HD.00	Plasma lactate dehydrogenase level	blood test
44HE.00	Plasma creatine kinase level	blood test
44HF.00	Serum lactate dehydrogenase level	blood test
44HG.00	Serum creatine kinase level	blood test
44HH.00	LDH blood level	blood test
44I..00	Serum electrolytes	blood test
44I1.00	Blood sent for electrolytes	blood test
44I2.00	Electrolytes normal	normal
44I3.00	Electrolytes abnormal	abnormal
44I4.00	Serum potassium	blood test
44I4000	Normal serum potassium level	normal
44I4100	Raised serum potassium level	abnormal
44I4200	Low serum potassium level	abnormal
44I5.00	Serum sodium	blood test
44I5000	Serum sodium level normal	normal
44I5100	Serum sodium level abnormal	abnormal

44I6.00	Serum chloride	blood test
44I6000	Serum chloride level normal	normal
44I6100	Serum chloride level abnormal	abnormal
44I7.00	Serum bicarbonate	blood test
44I7000	Serum bicarbonate level normal	normal
44I7100	Serum bicarbonate level abnormal	abnormal
44I8.00	Serum calcium	blood test
44I8000	Normal serum calcium level	normal
44I8100	Raised serum calcium level	abnormal
44I9.00	Serum inorganic phosphate	blood test
44I9000	Serum phosphate level normal	normal
44I9100	Serum phosphate level abnormal	abnormal
44IA.00	Plasma anion gap	blood test
44IB.00	Serum anion gap	blood test
44IC.00	Corrected serum calcium level	blood test
44ID.00	Serum ionised calcium level	blood test
44IE.00	Serum ionized calcium (pH 7.4) level	blood test
44IZ.00	Serum electrolytes NOS	blood test
44J..11	Urea - blood	blood test
44J..12	Urea and electrolytes	blood test
44J..13	Serum urea level	blood test
44J1.00	Blood urea normal	normal
44J2.00	Blood urea abnormal	abnormal
44J3.00	Serum creatinine	blood test
44J3000	Serum creatinine abnormal	abnormal
44J3100	Serum creatinine low	abnormal
44J3200	Serum creatinine normal	normal
44J3300	Serum creatinine raised	abnormal
44J3z00	Serum creatinine NOS	blood test
44J4.00	Serum osmolality	blood test
44J8.00	Blood urea	blood test
44J8.11	Urea - blood	blood test
44J9.00	Serum urea level	blood test
44JA.00	Plasma urea level	blood test
44JB.00	Urea and electrolytes	blood test
44JH.00	Plasma osmolality	blood test
44JZ.00	Blood urea/renal function NOS	blood test
44K..00	Blood urate	blood test
44K..11	Serum uric acid	blood test
44K1.00	Blood urate normal	normal
44K2.00	Blood urate raised	abnormal
44K2.11	Hyperuricaemia	abnormal
44K3.00	Blood urate level borderline	normal
44K4.00	Blood urate abnormal	abnormal
44M..00	Serum / plasma proteins	blood test
44M1.00	Serum proteins normal	normal
44M2.00	Serum proteins low	abnormal
44M3.00	Serum total protein	blood test
44M3000	Serum total protein normal	normal
44M3100	Serum total protein abnormal	abnormal
44Y..00	Blood gases	blood test
44Y1.00	Blood gases normal	normal
44Y2.00	Blood arterial pH	blood test
44Y2000	Blood pH normal	normal
44Y2100	Blood pH abnormal	abnormal
44Y3.00	Blood venous pH	blood test
44Y4.00	Arterial oxygen level	blood test
44Y4000	Blood oxygen level normal	normal
44Y4100	Blood oxygen level abnormal	abnormal
44Y5.00	Mixed venous oxygen level	blood test
44Y6.00	Arterial carbon dioxide	blood test
44Y6000	Arterial carbon dioxide normal	normal
44Y6100	Arterial carbon dioxide abnormal	abnormal
44Y7.00	Blood gases abnormal	abnormal
44Y8.00	Arterial blood gas analysis	blood test
44Y9.00	Blood oxygen saturation (calculated)	blood test
44YA.00	Blood oxygen saturation	blood test
44YB.00	Mixed venous carbon dioxide level	blood test

44YC.00	Mixed venous oxygen saturation	blood test
44YD.00	Hydrogen ion concentration	blood test
44YZ.00	Blood gases NOS	blood test
44Z..00	Blood chemistry NOS	blood test
44Z2.00	Bone profile	blood test
44ZR.00	Calcium profile	blood test
44h..00	Blood electrolyte levels	blood test
44h0.00	Blood potassium level	blood test
44h1.00	Blood sodium level	blood test
44h2.00	Blood chloride level	blood test
44h3.00	Blood bicarbonate level	blood test
44h4.00	Blood calcium level	blood test
44h5.00	Blood inorganic phosphate level	blood test
44h6.00	Plasma sodium level	blood test
44h7.00	Plasma calcium level	blood test
44h8.00	Plasma potassium level	blood test
44h9.00	Plasma corrected calcium level	blood test
44hA.00	Blood total carbon dioxide (calculated)	blood test
44hB.00	Actual bicarbonate level	blood test
44hC.00	Standard bicarbonate level	blood test
44i..00	Plasma electrolyte levels	blood test
44i0.00	Plasma bicarbonate level	blood test
44i1.00	Plasma chloride level	blood test
44i2.00	Plasma inorganic phosphate level	blood test

Appendix II: Most commonly recorded symptoms and conditions and their frequency in the medical records of patients with lung cancer

Most commonly recorded symptoms and diagnosis and their frequency in the medical records of patients with lung cancer

Read code description	frequency
[D]Sleep disturbances	1002
Knee pain	1005
Cellulitis NOS	1018
Shortness of breath symptom	1018
Resp. system examined - NAD	1021
Oedema	1026
Osteoporosis	1029
Hip joint pain	1030
Dermatitis NOS	1051
Blood sample -> Biochem Lab	1054
Letter invite to screening	1063
Acute myocardial infarction	1077
Glaucoma	1092
[D]Cough	1092
Diabetic on diet only	1106
Duodenal ulcer - (DU)	1106
Foot pain	1110
Nausea	1114
Influenza vaccination declined	1120
Intramuscular injection of vitamin B12	1121
[D]Haemoptysis	1130
Standard chest X-ray	1146
Wheezing	1156
Skin lesion	1162
Arthritis	1196
O/E - dry skin	1200
Atrial fibrillation	1205
C/O - low back pain	1207
Seen in rheumatology clinic	1220
Wound dressing NOS	1224
Seen in dermatology clinic	1229
Rheumatoid arthritis	1231
Throat soreness	1235
Dysuria	1255
[D]Vertigo NOS	1257
Chronic obstructive pulmonary disease monitoring	1258
Diarrhoea symptoms	1262
Immunisations	1266
Psoriasis unspecified	1275
Asthma annual review	1298
Intermittent claudication	1313
Vomiting	1314
ECG	1332
Type 2 diabetes mellitus	1334
Geriatric screen - seen	1334
Sinusitis	1339
Pure hypercholesterolaemia	1342
Acute conjunctivitis	1350
Conjunctivitis	1357
Dizziness symptom	1358
Haemoptysis - symptom	1365
Examination of patient	1427
Seen in cardiac clinic	1430
Gout	1437
Physiotherapy	1440
C/O: a pain	1463
Warfarin monitoring	1464
Health ed. - alcohol	1470
Haematuria	1472
Telephone triage encounter	1503
Seen in ENT clinic	1506
Pain relief	1515
Seen in diabetic clinic	1516

Seen by practice nurse	1531
Breathlessness	1553
Follow-up diabetic assessment	1560
Upper respiratory tract infection NOS	1560
Hypertension screen	1564
O/E - foot	1570
Urinary tract infection, site not specified	1572
Hip pain	1585
Cataract	1593
Acute exacerbation of chronic obstructive airways disease	1600
Cystitis	1640
Seen in oncology clinic	1648
Repeat prescription monitoring	1684
Emergency hospital admission	1691
[D]Dizziness	1698
C/O: a rash	1736
Epigastric pain	1753
Body Mass Index	1755
Leg pain	1763
Fall - accidental	1799
Anxiety states	1859
Laboratory test requested	1887
Leg ulcer NOS	1926
Diarrhoea	1934
Seen by respiratory physician	1938
Diabetic monitoring	1947
Chesty cough	1973
[D]Rash and other nonspecific skin eruption NOS	1981
Seen in chest clinic	2000
Backache, unspecified	2019
Seen in urology clinic	2034
Back pain without radiation NOS	2039
Depression screening using questions	2042
Intramuscular injection	2062
Constipation	2114
Otitis externa NOS	2122
[D]Insomnia NOS	2150
Knee joint pain	2233
[D]Abdominal pain	2234
Eczema NOS	2237
Headache	2260
Diabetic on oral treatment	2261
Seen in orthopaedic clinic	2273
[D]Shortness of breath	2279
Constipation symptom	2279
Refer for X-Ray	2318
Sciatica	2412
Wax in ear	2476
Influenza vaccination invitation letter sent	2497
X-ray report received	2520
Bronchitis unspecified	2527
Patient informed - test result	2531
O/E - BP reading	2585
Osteoarthritis	2603
Chronic obstructive airways disease	2647
Abdominal pain	2698
Shortness of breath	2719
Blood sample taken	2770
Backache	2780
Urinary tract infection, site not specified NOS	2809
Acute bronchitis	2916
Feet examination	2994
Pain	3019
Ischaemic heart disease	3162
Diabetes mellitus	3237
Cervicalgia - pain in neck	3329
Hormone replacement therapy	3370
Seen in hospital casualty	3388

Injection given	3429
Respiratory tract infection	3491
CHD monitoring	3776
Upper respiratory infection NOS	3941
Geriatric screening	3942
Shoulder pain	3950
Low back pain	3956
Weight loss advised	4082
Depressive disorder NEC	4176
Dyspepsia	4285
Syringe ear to remove wax	4538
O/E - blood pressure reading	4594
Asthma	4811
Angina pectoris	4886
Chronic obstructive pulmonary disease	5200
Hypertensive disease	5322
Chest infection NOS	5400
Cardiac disease monitoring	5420
C/O - cough	5576
Influenza vaccination	5989
Smoking cessation advice	6119
Hypertension monitoring	6217
Diabetes monitoring admin.	6251
Asthma monitoring	6528
Hypertension monitoring	6887
Essential hypertension	7451
Chest pain	8236
Cough	9976
Health ed. - smoking	10609
Blood sample -> Lab NOS	11400
Chest infection	12167

References

1. Muers M. Lung cancer. *Medicine* 2003;31(11):28-37.
2. WHO International Agency for Research on Cancer. GLOBOCAN 2008: Lung Cancer Incidence, Mortality and Prevalence Worldwide in 2008. Lyon, 2010. Available at <http://globocan.iarc.fr>.
3. Hunt I, Muers,M, Treasure,T, editor. *ABC of Lung Cancer*. West Sussex: Blackwell Publishing Ltd, 2009.
4. Ferlay J, Parkin DM, Steliarova-Foucher E. Estimates of cancer incidence and mortality in Europe in 2008. *Eur J Cancer* 2010;46(4):765-81.
5. Cancer Research UK. UK Lung cancer incidence statistics, 2010. Available at <http://info.cancerresearchuk.org/cancerstats/types/lung/incidence/index.htm>.
6. Office for National Statistics. Cancer registrations in England, 2009, 2011. Available at <http://www.ons.gov.uk/ons/rel/vsob1/cancer-registrations-in-england/2009/index.html>.
7. Information Services Division Scotland. Cancer Incidence in Scotland, 2009., 2011. Available at <http://www.isdscotland.org/Health-Topics/Cancer/Publications/2011-08-30/2011-08-30-Cancer-Incidence-Report.pdf>.
8. Welsh Cancer Intelligence and Surveillance Unit. Cancer Incidence in Wales 2005-2009, 2011. Available at <http://www.wales.nhs.uk/sites3/Documents/242/incpub2011.pdf>.
9. Northern Ireland Cancer Registry. Cancer incidence and mortality 2009, 2010. Available at <http://www.qub.ac.uk/research-centres/nicr/CancerData/OnlineStatistics/TracheaBronchusLung/>.
10. Office for National Statistics. Deaths registered in England and Wales in 2010, by cause, 2011. Available at <http://www.ons.gov.uk/ons/dcp171778_239518.pdf>.
11. Information Services Division; NHS National Services Scotland. Cancer mortality in Scotland (2010)

2011. Available at <http://www.isdscotland.org/Health-Topics//Cancer/Publications/index.asp#630>

12. Northern Ireland Cancer Registry. Number of cancer deaths and mortality rates in 2010, by sex, 2012. Available at <http://www.qub.ac.uk/research-centres/nicr/CancerData/OnlineStatistics/TracheaBronchusLung/>.

13. Cancer Research UK. UK Cancer incidence for common cancers, 2012. Available at <http://info.cancerresearchuk.org/cancerstats/incidence/commoncancers/#Twenty>.
14. Cancer Research UK. Deaths from common cancers in the UK, 2012. Available at <http://info.cancerresearchuk.org/cancerstats/mortality/cancerdeaths/>.
15. Tyczynski JE, Bray F, Parkin DM. Lung cancer in Europe in 2000: epidemiology, prevention, and early detection. *Lancet Oncology* 2003;4(1):45-55.
16. Office for National Statistics (ONS). Cancer survival in England - Patients diagnosed 2005-2009 and followed up to 2010, Available at <http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-239726>.
17. Janssen-Heijnen MLG, Gatta G, Forman D, Capocaccia R, Coebergh JWW, Grp EW. Variation in survival of patients with lung cancer in Europe, 1985-1989. *European Journal of Cancer* 1998;34(14):2191-96.
18. Holmberg L, Sandin F, Bray F, Richards M, Spicer J, Lambe M, et al. National comparisons of lung cancer survival in England, Norway and Sweden 2001-2004: differences occur early in follow-up. *Thorax* 2010;65(5):436-41.
19. Imperatori A, Harrison RN, Leitch DN, Rovera F, Lepore G, Dionigi G, et al. Lung cancer in Teesside (UK) and Varese (Italy): a comparison of management and survival. *Thorax* 2006;61(3):232-9.
20. Verdecchia A, Francisci S, Brenner H, Gatta G, Micheli A, Mangone L, et al. Recent cancer survival in Europe: a 2000-02 period analysis of EUROCARE-4 data. *Lancet Oncol* 2007;8(9):784-96.
21. Levin ML, Goldstein H, Gerhardt PR. Cancer and tobacco smoking; a preliminary report. *J Am Med Assoc* 1950;143(4):336-8.
22. Wynder EL, Graham EA. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. *J Am Med Assoc* 1950;143(4):329-36.
23. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. *Br Med J* 1950;2(4682):739-48.
24. Doll R, Hill AB. A study of the aetiology of carcinoma of the lung. *Br Med J* 1952;2(4797):1271-86.

25. Doll R, Hill AB. Lung cancer and other causes of death in relation to smoking; a second report on the mortality of British doctors. *Br Med J* 1956;2(5001):1071-81.
26. Parkin DM. 2. Tobacco-attributable cancer burden in the UK in 2010. *Br J Cancer* 2011;105 Suppl 2:S6-S13.
27. Edwards R. The problem of tobacco smoking. *BMJ* 2004;328(7433):217-9.
28. Lung and Asthma Information Agency. Trends in Smoking. Available at <http://www.laia.ac.uk/factsheets/982.pdf>.
29. Office for National Statistics. A report on the 2010 General Lifestyle Survey., 2011. Available at <http://www.ons.gov.uk/ons/rel/ghs/general-lifestyle-survey/2010/general-lifestyle-survey-overview-report-2010.pdf>.
30. Wood H, Cooper N, Rowan S, Quinn M. Lung. *Cancer atlas of the United Kingdom and Ireland 1991-2000*: Macmillan, 2005.
31. Doll R, Peto R, Boreham J, Sutherland I. Mortality from cancer in relation to smoking: 50 years observations on British doctors. *Br J Cancer* 2005;92(3):426-9.
32. Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ* 2000;321(7257):323-9.
33. Law MR, Morris JK, Watt HC, Wald NJ. The dose-response relationship between cigarette consumption, biochemical markers and risk of lung cancer. *Br J Cancer* 1997;75(11):1690-3.
34. Lubin JH, Caporaso NE. Cigarette smoking and lung cancer: modeling total exposure and intensity. *Cancer Epidemiol Biomarkers Prev* 2006;15(3):517-23.
35. Lubin JH, Alavanja MC, Caporaso N, Brown LM, Brownson RC, Field RW, et al. Cigarette smoking and cancer risk: modeling total exposure and intensity. *Am J Epidemiol* 2007;166(4):479-89.
36. Taylor R, Cumming R, Woodward A, Black M. Passive smoking and lung cancer: a cumulative meta-analysis. *Aust N Z J Public Health* 2001;25(3):203-11.
37. Taylor R, Najafi F, Dobson A. Meta-analysis of studies of passive smoking and lung cancer: effects of study type and continent. *Int J Epidemiol* 2007;36(5):1048-59.
38. Alberg AJ, Ford JG, Samet JM. Epidemiology of lung cancer - ACCP evidence-based clinical practice guidelines (2nd edition). *Chest* 2007;132(3):29s-55s.

39. van Loon AJ, Kant IJ, Swaen GM, Goldbohm RA, Kremer AM, van den Brandt PA. Occupational exposure to carcinogens and risk of lung cancer: results from The Netherlands cohort study. *Occup Environ Med* 1997;54(11):817-24.
40. Schoenberg JB, Stemhagen A, Mason TJ, Patterson J, Bill J, Altman R. Occupation and lung cancer risk among New Jersey white males. *J Natl Cancer Inst* 1987;79(1):13-21.
41. Vineis P, Thomas T, Hayes RB, Blot WJ, Mason TJ, Pickle LW, et al. Proportion of lung cancers in males, due to occupation, in different areas of the USA. *Int J Cancer* 1988;42(6):851-6.
42. Doll R, Peto J. Asbestos: Effects on health of exposure to asbestos: Health and Safety Commission, 1985. Available at <http://www.hse.gov.uk/asbestos/exposure.pdf>.
43. Rushton L, Bagga S, Bevan R, Brown TP, Cherrie JW, Holmes P, et al. Occupation and cancer in Britain. *Br J Cancer* 2010;102(9):1428-37.
44. Parkin DM. 14. Cancers attributable to occupational exposures in the UK in 2010. *Br J Cancer* 2011;105 Suppl 2:S70-2.
45. Lubin JH, Boice JD, Jr., Edling C, Hornung RW, Howe GR, Kunz E, et al. Lung cancer in radon-exposed miners and estimation of risk from indoor exposure. *J Natl Cancer Inst* 1995;87(11):817-27.
46. World Health Organisation. Radon and cancer: Key facts, 2012. Available at <http://www.who.int/mediacentre/factsheets/fs291/en/index.html>.
47. Parkin DM, Darby SC. 12. Cancers in 2010 attributable to ionising radiation exposure in the UK. *Br J Cancer* 2011;105 Suppl 2:S57-65.
48. Darby S, Hill D, Auvinen A, Barros-Dios JM, Baysson H, Bochicchio F, et al. Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. *BMJ* 2005;330(7485):223.
49. Nitadori J, Inoue M, Iwasaki M, Otani T, Sasazuki S, Nagai K, et al. Association between lung cancer incidence and family history of lung cancer: data from a large-scale population-based cohort study, the JPHC study. *Chest* 2006;130(4):968-75.
50. Cassidy A, Myles JP, Duffy SW, Liloglou T, Field JK. Family history and risk of lung cancer: age-at-diagnosis in cases and first-degree relatives. *British Journal of Cancer* 2006;95(9):1288-90.
51. Lorigan P, Radford J, Howell A, Thatcher N. Lung cancer after treatment for Hodgkin's lymphoma: a systematic review. *Lancet Oncol* 2005;6(10):773-9.

52. Das P, Ng AK, Stevenson MA, Mauch PM. Clinical course of thoracic cancers in Hodgkin's disease survivors. *Ann Oncol* 2005;16(5):793-7.
53. Mudie NY, Swerdlow AJ, Higgins CD, Smith P, Qiao Z, Hancock BW, et al. Risk of second malignancy after non-Hodgkin's lymphoma: a British Cohort Study. *J Clin Oncol* 2006;24(10):1568-74.
54. Raymond JS, Hogue CJ. Multiple primary tumours in women following breast cancer, 1973-2000. *Br J Cancer* 2006;94(11):1745-50.
55. Travis LB, Fossa SD, Schonfeld SJ, McMaster ML, Lynch CF, Storm H, et al. Second cancers among 40,576 testicular cancer patients: focus on long-term survivors. *J Natl Cancer Inst* 2005;97(18):1354-65.
56. Johnson BE. Second lung cancers in patients after treatment for an initial lung cancer. *J Natl Cancer Inst* 1998;90(18):1335-45.
57. Brody JS, Spira A. State of the art. Chronic obstructive pulmonary disease, inflammation, and lung cancer. *Proc Am Thorac Soc* 2006;3(6):535-7.
58. Kiri VA, Soriano J, Visick G, Fabbri L. Recent trends in lung cancer and its association with COPD: an analysis using the UK GP Research Database. *Primary Care Respiratory Journal* 2010;19(1):57-61.
59. Wasswa-Kintu S, Gan WQ, Man SF, Pare PD, Sin DD. Relationship between reduced forced expiratory volume in one second and the risk of lung cancer: a systematic review and meta-analysis. *Thorax* 2005;60(7):570-5.
60. Purdue MP, Gold L, Jarvholm B, Alavanja MC, Ward MH, Vermeulen R. Impaired lung function and lung cancer incidence in a cohort of Swedish construction workers. *Thorax* 2007;62(1):51-6.
61. Eriksson NE, Holmen A, Hogstedt B, Mikoczy Z, Hagmar L. A prospective study of cancer incidence in a cohort examined for allergy. *Allergy* 1995;50(9):718-22.
62. El-Zein M, Parent ME, Ka K, Siemiatycki J, St-Pierre Y, Rousseau MC. History of asthma or eczema and cancer risk among men: a population-based case-control study in Montreal, Quebec, Canada. *Ann Allergy Asthma Immunol* 2010;104(5):378-84.
63. Santillan AA, Camargo CA, Jr., Colditz GA. A meta-analysis of asthma and risk of lung cancer (United States). *Cancer Causes Control* 2003;14(4):327-34.
64. Fan YG, Jiang Y, Chang RS, Yao SX, Jin P, Wang W, et al. Prior lung disease and lung cancer risk in an occupational-based cohort in Yunnan, China. *Lung Cancer* 2011;72(2):258-63.

65. Liang H, Guan P, Yin Z, Li X, He Q, Zhou B. Risk of lung cancer following nonmalignant respiratory conditions among nonsmoking women living in Shenyang, Northeast China. *J Womens Health (Larchmt)* 2009;18(12):1989-95.
66. Hubbard R, Venn A, Lewis S, Britton J. Lung cancer and cryptogenic fibrosing alveolitis. A population-based cohort study. *Am J Respir Crit Care Med* 2000;161(1):5-8.
67. Le Jeune I, Gribbin J, West J, Smith C, Cullinan P, Hubbard R. The incidence of cancer in patients with idiopathic pulmonary fibrosis and sarcoidosis in the UK. *Respiratory Medicine* 2007;101(12):2534-40.
68. Daniels CE, Jett JR. Does interstitial lung disease predispose to lung cancer? *Curr Opin Pulm Med* 2005;11(5):431-7.
69. Vineis P, Hoek G, Krzyzanowski M, Vigna-Taglianti F, Veglia F, Airoldi L, et al. Lung cancers attributable to environmental tobacco smoke and air pollution in non-smokers in different European countries: a prospective study. *Environ Health* 2007;6:7.
70. Ramanakumar AV, Parent ME, Siemiatycki J. Risk of lung cancer from residential heating and cooking fuels in Montreal, Canada. *Am J Epidemiol* 2007;165(6):634-42.
71. Lissowska J, Bardin-Mikolajczak A, Fletcher T, Zaridze D, Szeszenia-Dabrowska N, Rudnai P, et al. Lung cancer and indoor pollution from heating and cooking with solid fuels: the IARC international multicentre case-control study in Eastern/Central Europe and the United Kingdom. *Am J Epidemiol* 2005;162(4):326-33.
72. Steindorf K, Friedenreich C, Linseisen J, Rohrmann S, Rundle A, Veglia F, et al. Physical activity and lung cancer risk in the European Prospective Investigation into Cancer and Nutrition Cohort. *Int J Cancer* 2006;119(10):2389-97.
73. Sinner P, Folsom AR, Harnack L, Eberly LE, Schmitz KH. The association of physical activity with lung cancer incidence in a cohort of older women: the Iowa Women's Health Study. *Cancer Epidemiol Biomarkers Prev* 2006;15(12):2359-63.
74. Alfano CM, Klesges RC, Murray DM, Bowen DJ, McTiernan A, Vander Weg MW, et al. Physical activity in relation to all-site and lung cancer incidence and mortality in current and former smokers. *Cancer Epidemiol Biomarkers Prev* 2004;13(12):2233-41.
75. Holick CN, Michaud DS, Stolzenberg-Solomon R, Mayne ST, Pietinen P, Taylor PR, et al. Dietary carotenoids, serum beta-carotene, and retinol

and risk of lung cancer in the alpha-tocopherol, beta-carotene cohort study. *Am J Epidemiol* 2002;156(6):536-47.

76. Rylander R, Axelsson G. Lung cancer risks in relation to vegetable and fruit consumption and smoking. *Int J Cancer* 2006;118(3):739-43.
77. Galeone C, Negri E, Pelucchi C, La Vecchia C, Bosetti C, Hu J. Dietary intake of fruit and vegetable and lung cancer risk: a case-control study in Harbin, northeast China. *Ann Oncol* 2007;18(2):388-92.
78. Mahabir S, Spitz MR, Barrera SL, Beaver SH, Etzel C, Forman MR. Dietary zinc, copper and selenium, and risk of lung cancer. *Int J Cancer* 2007;120(5):1108-15.
79. Parkin DM, Boyd L. 4. Cancers attributable to dietary factors in the UK in 2010. I. Low consumption of fruit and vegetables. *Br J Cancer* 2011;105 Suppl 2:S19-23.
80. Freudenheim JL, Ritz J, Smith-Warner SA, Albanes D, Bandera EV, van den Brandt PA, et al. Alcohol consumption and risk of lung cancer: a pooled analysis of cohort studies. *Am J Clin Nutr* 2005;82(3):657-67.
81. Rohrmann S, Linseisen J, Boshuizen HC, Whittaker J, Agudo A, Vineis P, et al. Ethanol intake and risk of lung cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC). *Am J Epidemiol* 2006;164(11):1103-14.
82. Thun MJ, Hannan LM, DeLancey JO. Alcohol consumption not associated with lung cancer mortality in lifelong nonsmokers. *Cancer Epidemiol Biomarkers Prev* 2009;18(8):2269-72.
83. Bagnardi V, Rota M, Botteri E, Scotti L, Jenab M, Bellocco R, et al. Alcohol consumption and lung cancer risk in never smokers: a meta-analysis. *Ann Oncol* 2011;22(12):2631-9.
84. Rich AL, Tata LJ, Stanley RA, Free CM, Peake MD, Baldwin DR, et al. Lung cancer in England: information from the National Lung Cancer Audit (LUCADA). *Lung Cancer* 2011;72(1):16-22.
85. Cancer Research UK. Lung cancer - UK statistics. Trends over time, 2011. Available at <http://info.cancerresearchuk.org/cancerstats/types/lung/incidence/uk-lung-cancer-incidence-statistics#trends>.
86. Brownson RC, Chang JC, Davis JR. Gender and histologic type variations in smoking-related risk of lung cancer. *Epidemiology* 1992;3(1):61-4.
87. Zang EA, Wynder EL. Differences in lung cancer risk between men and women: examination of the evidence. *J Natl Cancer Inst* 1996;88(3-4):183-92.

88. Bain C, Feskanich D, Speizer FE, Thun M, Hertzmark E, Rosner BA, et al. Lung cancer rates in men and women with comparable histories of smoking. *J Natl Cancer Inst* 2004;96(11):826-34.
89. Freedman ND, Leitzmann MF, Hollenbeck AR, Schatzkin A, Abnet CC. Cigarette smoking and subsequent risk of lung cancer in men and women: analysis of a prospective cohort study. *Lancet Oncol* 2008;9(7):649-56.
90. Sidorchuk A, Agardh EE, Aremu O, Hallqvist J, Allebeck P, Moradi T. Socioeconomic differences in lung cancer incidence: a systematic review and meta-analysis. *Cancer Causes & Control* 2009;20(4):459-71.
91. Dalton SO, Steding-Jessen M, Engholm G, Schuz J, Olsen JH. Social inequality and incidence of and survival from lung cancer in a population-based study in Denmark, 1994-2003. *European Journal of Cancer* 2008;44(14):1989-95.
92. Shack L, Jordan C, Thomson CS, Mak V, Moller H. Variation in incidence of breast, lung and cervical cancer and malignant melanoma of skin by socioeconomic group in England. *BMC Cancer* 2008;8:271.
93. Menvielle G, Boshuizen H, Kunst AE, Dalton SO, Vineis P, Bergmann MM, et al. The Role of Smoking and Diet in Explaining Educational Inequalities in Lung Cancer Incidence. *Journal of the National Cancer Institute* 2009;101(5):321-30.
94. Mignot A. Smoking Behavior Partially Explains Socioeconomic Inequities in Lung Cancer Incidence. *Journal of the National Cancer Institute* 2009;101(5):-.
95. Menvielle G, Boshuizen H, Kunst AE, Vineis P, Dalton SO, Bergmann MM, et al. Occupational exposures contribute to educational inequalities in lung cancer incidence among men: Evidence from the EPIC prospective cohort study. *International Journal of Cancer* 2010;126(8):1928-35.
96. Shack L, Jordan C, Thomson CS, Mak V, Moller H, Registries UAC. Variation in incidence of breast, lung and cervical cancer and malignant melanoma of skin by socioeconomic group in England. *BMC Cancer* 2008;8:-.
97. Mao Y, Hu J, Ugnat AM, Semenciw R, Fincham S. Socioeconomic status and lung cancer risk in Canada. *Int J Epidemiol* 2001;30(4):809-17.
98. van Loon AJ, Goldbohm RA, van den Brandt PA. Lung cancer: is there an association with socioeconomic status in The Netherlands? *J Epidemiol Community Health* 1995;49(1):65-9.

99. Fidler JA, Jarvis MJ, Mindell J, West R. Nicotine intake in cigarette smokers in England: distribution and demographic correlates. *Cancer Epidemiol Biomarkers Prev* 2008;17(12):3331-6.
100. Jarvis MJ, Wardle J. *Social patterning of individual health behaviours: the case of cigarette smoking. In: Marmot M, Wilkinson RG, editors.* Oxford: Oxford University Press, 2006.
101. Experian. Multimedia guide to mosaic public sector, 2009.
102. Sharma A, Lewis S, Szatkowski L. Insights into social disparities in smoking prevalence using Mosaic, a novel measure of socioeconomic status: an analysis using a large primary care dataset. *Bmc Public Health* 2010;10:755.
103. Coulthard M, Farrell M, Singleton N, Meltzer H. Tobacco, alcohol and drug use and mental health. London: Office of National Statistics,, 2002.
104. Miller AH, Spencer RL, McEwen BS, Stein M. Depression, adrenal steroids, and the immune system. *Ann Med* 1993;25(5):481-7.
105. Penninx BW, Guralnik JM, Pahor M, Ferrucci L, Cerhan JR, Wallace RB, et al. Chronically depressed mood and cancer risk in older persons. *J Natl Cancer Inst* 1998;90(24):1888-93.
106. Dalton SO, Mellemkjaer L, Olsen JH, Mortensen PB, Johansen C. Depression and cancer risk: a register-based study of patients hospitalized with affective disorders, Denmark, 1969-1993. *Am J Epidemiol* 2002;155(12):1088-95.
107. Knekt P, Raitasalo R, Heliovaara M, Lehtinen V, Pukkala E, Teppo L, et al. Elevated lung cancer risk among persons with depressed mood. *Am J Epidemiol* 1996;144(12):1096-103.
108. Linkins RW, Comstock GW. Depressed mood and development of cancer. *Am J Epidemiol* 1990;132(5):962-72.
109. Oerlemans ME, van den Akker M, Schuurman AG, Kellen E, Buntinx F. A meta-analysis on depression and subsequent cancer risk. *Clin Pract Epidemiol Ment Health* 2007;3:29.
110. Beckles MA, Spiro SG, Colice GL, Rudd RM. Initial evaluation of the patient with lung cancer - Symptoms, signs, laboratory tests, and paraneoplastic syndromes. *Chest* 2003;123(1):97s-104s.
111. Koyi H, Hillerdal G, Branden E. A prospective study of a total material of lung cancer from a county in Sweden 1997-1999: gender, symptoms, type, stage, and smoking habits. *Lung Cancer* 2002;36(1):9-14.

112. Corner J, Hopkinson J, Fitzsimmons D, Barclay S, Muers M. Is late diagnosis of lung cancer inevitable? Interview study of patients' recollections of symptoms before diagnosis. *Thorax* 2005;60(4):314-9.
113. Hamilton W, Peters TJ, Round A, Sharp D. What are the clinical features of lung cancer before the diagnosis is made? A population based case-control study. *Thorax* 2005;60(12):1059-65.
114. Bjerager M, Palshof T, Dahl R, Vedsted P, Olesen F. Delay in diagnosis of lung cancer in general practice. *British Journal of General Practice* 2006;56(532):863-68.
115. Cromartie RS, 3rd, Parker EF, May JE, Metcalf JS, Bartles DM. Carcinoma of the lung: a clinical review. *Annals of Thoracic Surgery* 1980;30(1):30-5.
116. Okkes IM, Oskam SK, Lamberts H. The probability of specific diagnoses for patients presenting with common symptoms to Dutch family physicians. *Journal of Family Practice* 2002;51(1):31-36.
117. Jones R, Latinovic R, Charlton J, Gulliford MC. Alarm symptoms in early diagnosis of cancer in primary care: cohort study using General Practice Research Database. *BMJ* 2007;334(7602):1040.
118. Lovgren M, Levealahti H, Tishelman C, Runesdotter S, Hamberg K. Time spans from first symptom to treatment in patients with lung cancer--the influence of symptoms and demographic characteristics. *Acta Oncol* 2008;47(3):397-405.
119. Buccheri G, Ferrigno D. Lung cancer: clinical presentation and specialist referral time. *Eur Respir J* 2004;24(6):898-904.
120. Weiss W, Seidman H, Boucot KR. The Philadelphia Pulmonary Neoplasm Research Project. Symptoms in occult lung cancer. *Chest* 1978;73(1):57-61.
121. Smith RA, Glynn TJ. Epidemiology of lung cancer. *Radiol Clin North Am* 2000;38(3):453-70.
122. National Institute for Health and Clinical Excellence. Lung cancer. The diagnosis and treatment of lung cancer, 2011. Available at <http://www.nice.org.uk/nicemedia/live/13465/54202/54202.pdf>.
123. Pearson FG. Current status of surgical resection for lung cancer. *Chest* 1994;106(6 Suppl):337S-39S.
124. Mountain CF. Revisions in the International System for Staging Lung Cancer. *Chest* 1997;111(6):1710-7.
125. Melamed MR, Flehinger BJ, Zaman MB, Heelan RT, Perchick WA, Martini N. Screening for early lung cancer. Results of the Memorial Sloan-Kettering study in New York. *Chest* 1984;86(1):44-53.

126. Flehinger BJ, Kimmel M, Melamed MR. The effect of surgical treatment on survival from early lung cancer. Implications for screening. *Chest* 1992;101(4):1013-8.
127. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011;365(5):395-409.
128. Baldwin DR, Duffy SW, Wald NJ, Page R, Hansell DM, Field JK. 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):308-13.
129. Infante MV, Pedersen JH. Screening for lung cancer: are we there yet? *Curr Opin Pulm Med* 2010;16(4):301-6.
130. Pedersen JH, Ashraf H, Dirksen A, Bach K, Hansen H, Toennesen P, et al. The Danish randomized lung cancer CT screening trial--overall design and results of the prevalence round. *J Thorac Oncol* 2009;4(5):608-14.
131. Department of Health. The NHS Cancer Plan. London: Department of Health, 2000.
132. Robinson E, Mohilever J, Zidan J, Sapir D. Delay in diagnosis of cancer. Possible effects on the stage of disease and survival. *Cancer* 1984;54(7):1454-60.
133. O'Rourke N, Edwards R. Lung cancer treatment waiting times and tumour growth. *Clinical Oncology* 2000;12(3):141-44.
134. Hamilton W, Sharp D. Diagnosis of lung cancer in primary care: a structured review. *Fam Pract* 2004;21(6):605-11.
135. The Information Centre for Health and Social Care and RCP. National Lung Cancer Audit, Report for the audit period 2005 <http://www.ic.nhs.uk/statistics-and-data-collections/hospital-care/cancer/national-lung-cancer-audit-report-2006> December 2006
136. Read C, Janes S, George J, Spiro S. Early lung cancer: screening and detection. *Prim Care Respir J* 2006;15(6):332-6.
137. Devbhandari MP, Yang SS, Quennell P, Krysiak P, Shah R, Jones MT. Lung cancer resection rate in south Manchester: is it comparable to international standards? Results of a prospective tracking study. *Interact Cardiovasc Thorac Surg* 2007;6(6):712-4.
138. Department of Health. The New NHS: Modern and dependable. London, 1997. Available at <http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/Publica

tionsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4008869>.

139. Allgar VL, Neal RD. Delays in the diagnosis of six cancers: analysis of data from the National Survey of NHS Patients: Cancer. *British Journal of Cancer* 2005;92(11):1959-70.
140. Richards MA. The National Awareness and Early Diagnosis Initiative in England: assembling the evidence. *Br J Cancer* 2009;101 Suppl 2:S1-4.
141. Barrett J, Hamilton W. Pathways to the diagnosis of lung cancer in the UK: a cohort study. *BMC Fam Pract* 2008;9:31.
142. Allgar VL, Neal RD, Ali N, Leese B, Heywood P, Proctor G, et al. Urgent GP referrals for suspected lung, colorectal, prostate and ovarian cancer. *British Journal of General Practice* 2006;56(526):355-62.
143. Muers MF, Holmes WF, Littlewood C. Issues at the interface between primary and secondary care in the management of common respiratory disease . 1 - The challenge of improving the delivery of lung cancer care. *Thorax* 1999;54(6):540-43.
144. National Cancer Intelligence Network. Routes to Diagnosis - NCIN Data Briefing, 2010. Available at <http://www.ncin.org.uk/publications/data_briefings/routes_to_diagnosis.aspx>.
145. National Institute for Health and Clinical Excellence. About NICE, 2012. Available at <http://www.nice.org.uk/aboutnice/whoweare/who_we_are.jsp>.
146. Department of Health. Referral guidelines for suspected cancer. 2000. Available at <http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4014421.pdf>.
147. National Institute for Clinical Excellence. The diagnosis and treatment of lung cancer: Methods, evidence and guidance. <Available at www.**nice.org.uk**/nicemedia/pdf/cg024fullguideline.pdf >, 2005.
148. Okello C, Treasure T, Nicholson AG, Peto J, Moller H, Okello C, et al. Certified causes of death in patients with mesothelioma in South East England. *BMC Cancer* 2009;9:28.
149. Dunleavey R, Dunleavey R. Malignant mesothelioma: risk factors and current management. *Nurs Times* 2004;100(16):40-3.
150. Yang H, Testa JR, Carbone M, Yang H, Testa JR, Carbone M. Mesothelioma epidemiology, carcinogenesis, and pathogenesis. *Curr Treat Options Oncol* 2008;9(2-3):147-57.

151. Kurumatani N, Tomioka K, Kurumatani N, Tomioka K. [Epidemiology of pleural mesothelioma in Japan]. *Nippon Geka Gakkai Zasshi* 2009;110(6):320-5.
152. Cancer Research UK. Statistics and outlook for mesothelioma, 2009.
153. Collins GS, Altman DG. An independent external validation and evaluation of QRISK cardiovascular risk prediction: a prospective open cohort study. *British Medical Journal* 2009;339:-.
154. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;338:b375.
155. Johnston ME, Langton KB, Haynes RB, Mathieu A. Effects of computer-based clinical decision support systems on clinician performance and patient outcome. A critical appraisal of research. *Ann Intern Med* 1994;120(2):135-42.
156. Grady D, Berkowitz S. Why is a good clinical prediction rule so hard to find? *Arch Intern Med* 2011;171(19):1701-2.
157. Wyatt JC, Altman DG. Prognostic Models - Clinically Useful or Quickly Forgotten - Commentary. *British Medical Journal* 1995;311(7019):1539-41.
158. Freedman AN, Seminara D, Gail MH, Hartge P, Colditz GA, Ballard-Barbash R, et al. Cancer risk prediction models: a workshop on development, evaluation, and application. *J Natl Cancer Inst* 2005;97(10):715-23.
159. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;115(7):928-35.
160. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143(1):29-36.
161. Kannel WB, Schwartz MJ, Mcnamara PM. Blood Pressure and Risk of Coronary Heart Disease - Framingham Study. *Diseases of the Chest* 1969;56(1):43-&.
162. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *British Medical Journal* 2007;335(7611):136-41.
163. Woodward M, Brindle P, Tunstall-Pedoe H, Estimation SGR. Adding social deprivation and family history to cardiovascular risk assessment: the ASSIGN score from the Scottish Heart Health Extended Cohort (SHHEC). *Heart* 2007;93(2):172-76.
164. Lindstrom J, Tuomilehto J. The diabetes risk score - A practical tool to predict type 2 diabetes risk. *Diabetes Care* 2003;26(3):725-31.

165. Teasdale G, Jennett B. Assessment of Coma and Impaired Consciousness - Practical Scale. *Lancet* 1974;2(7872):81-84.
166. Knaus WA, Wagner DP, Lynn J. Short-term mortality predictions for critically ill hospitalized adults: science and ethics. *Science* 1991;254(5030):389-94.
167. Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA* 1993;270(24):2957-63.
168. Ottman R, Pike MC, King MC, Henderson BE. Practical guide for estimating risk for familial breast cancer. *Lancet* 1983;2(8349):556-8.
169. Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;81(24):1879-86.
170. Claus EB, Risch N, Thompson WD. The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast Cancer Res Treat* 1993;28(2):115-20.
171. Colditz GA, Rosner B. Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study. *Am J Epidemiol* 2000;152(10):950-64.
172. Tice JA, Cummings SR, Ziv E, Kerlikowske K. Mammographic breast density and the Gail model for breast cancer risk prediction in a screening population. *Breast Cancer Res Treat* 2005;94(2):115-22.
173. Parmigiani G, Berry D, Aguilar O. Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am J Hum Genet* 1998;62(1):145-58.
174. Vahteristo P, Eerola H, Tamminen A, Blomqvist C, Nevanlinna H. A probability model for predicting BRCA1 and BRCA2 mutations in breast and breast-ovarian cancer families. *Br J Cancer* 2001;84(5):704-8.
175. Fisher TJ, Kirk J, Hopper JL, Godding R, Burgemeister FC. A simple tool for identifying unaffected women at a moderately increased or potentially high risk of breast cancer based on their family history. *Breast* 2003;12(2):120-7.
176. Selvachandran SN, Hodder RJ, Ballal MS, Jones P, Cade D. Prediction of colorectal cancer by a patient consultation questionnaire and scoring system: a prospective study. *Lancet* 2002;360(9329):278-83.

177. Driver JA, Gaziano JM, Gelber RP, Lee IM, Buring JE, Kurth T. Development of a risk score for colorectal cancer in men. *Am J Med* 2007;120(3):257-63.
178. Freedman AN, Slattery ML, Ballard-Barbash R, Willis G, Cann BJ, Pee D, et al. Colorectal cancer risk prediction tool for white men and women without known susceptibility. *J Clin Oncol* 2009;27(5):686-93.
179. Park Y, Freedman AN, Gail MH, Pee D, Hollenbeck A, Schatzkin A, et al. Validation of a colorectal cancer risk prediction model among white patients age 50 years and older. *J Clin Oncol* 2009;27(5):694-8.
180. Optenberg SA, Clark JY, Brawer MK, Thompson IM, Stein CR, Friedrichs P. Development of a decision-making tool to predict risk of prostate cancer: the Cancer of the Prostate Risk Index (CAPRI) test. *Urology* 1997;50(5):665-72.
181. Thompson IM, Ankerst DP, Chi C, Goodman PJ, Tangen CM, Lucia MS, et al. Assessing prostate cancer risk: results from the Prostate Cancer Prevention Trial. *J Natl Cancer Inst* 2006;98(8):529-34.
182. Nam RK, Toi A, Klotz LH, Trachtenberg J, Jewett MA, Appu S, et al. Assessing individual risk for prostate cancer. *J Clin Oncol* 2007;25(24):3582-8.
183. Wu X, Lin J, Grossman HB, Huang M, Gu J, Etzel CJ, et al. Projecting individualized probabilities of developing bladder cancer in white individuals. *J Clin Oncol* 2007;25(31):4974-81.
184. Fears TR, Guerry Dt, Pfeiffer RM, Sagebiel RW, Elder DE, Halpern A, et al. Identifying individuals at high risk of melanoma: a practical predictor of absolute risk. *J Clin Oncol* 2006;24(22):3590-6.
185. Williams LH, Shors AR, Barlow WE, Solomon C, White E. Identifying Persons at Highest Risk of Melanoma Using Self-Assessed Risk Factors. *J Clin Exp Dermatol Res* 2011;2(6).
186. Andersen MR, Goff BA, Lowe KA, Scholler N, Bergan L, Drescher CW, et al. Use of a Symptom Index, CA125, and HE4 to predict ovarian cancer. *Gynecol Oncol* 2010;116(3):378-83.
187. Wang W, Chen S, Brune KA, Hruban RH, Parmigiani G, Klein AP. PancPRO: risk assessment for individuals with a family history of pancreatic cancer. *J Clin Oncol* 2007;25(11):1417-22.
188. Cai QC, Chen Y, Xiao Y, Zhu W, Xu QF, Zhong L, et al. A prediction rule for estimating pancreatic cancer risk in chronic pancreatitis patients with focal pancreatic mass lesions with prior negative EUS-FNA cytology. *Scand J Gastroenterol* 2011;46(4):464-70.

189. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, et al. Variations in lung cancer risk among smokers. *J Natl Cancer Inst* 2003;95(6):470-8.
190. Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Dong Q, et al. A risk model for prediction of lung cancer. *J Natl Cancer Inst* 2007;99(9):715-26.
191. Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, et al. The LLP risk model: an individual risk prediction model for lung cancer. *British Journal of Cancer* 2008;98(2):270-6.
192. Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2011;61(592):715-23.
193. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, et al. Lung Cancer Risk Prediction: Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial Models and Validation. *J Natl Cancer Inst* 2011.
194. Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, et al. An expanded risk prediction model for lung cancer. *Cancer Prev Res (Phila)* 2008;1(4):250-4.
195. D'Amelio AM, Jr., Cassidy A, Asomaning K, Raji OY, Duffy SW, Field JK, et al. Comparison of discriminatory power and accuracy of three lung cancer risk models. *Br J Cancer* 2010;103(3):423-9.
196. Smith SM, Campbell NC, MacLeod U, Lee AJ, Raja A, Wyke S, et al. Factors contributing to the time taken to consult with symptoms of lung cancer: a cross-sectional study. *Thorax* 2009;64(6):523-31.
197. Bowen EF, Rayner CFJ. Patient and GP led delays in the recognition of symptoms suggestive of lung cancer. *Lung Cancer* 2002;37(2):227-28.
198. Hubbard RB, Baldwin DR. Diagnosing lung cancer earlier in the UK. *Thorax* 2010;65(9):756-8.
199. CSD EPIC. THIN data from EPIC: A guide for researchers. London, July 2009.
200. InPS. In Practice Systems.
201. Bourke A, Dattani H, Robinson M. Feasibility study and methodology to create a quality-evaluated database of primary care data. *Inform Prim Care* 2004;12(3):171-7.
202. Lis Y, Mann RD. The VAMP Research multi-purpose database in the U.K. *J Clin Epidemiol* 1995;48(3):431-43.

203. Dave S, Petersen I. Creating medical and drug code lists to identify cases in primary care databases. *Pharmacoepidemiol Drug Saf* 2009;18(8):704-7.
204. Lewis JD, Schinnar R, Bilker WB, Wang XM, Strom BL. Validation studies of the health improvement network (THIN) database for pharmacoepidemiology research. *Pharmacoepidem Dr S* 2007;16(4):393-401.
205. Langley TE, Szatkowski L, Gibson J, Huang Y, McNeill A, Coleman T, et al. Validation of The Health Improvement Network (THIN) primary care database for monitoring prescriptions for smoking cessation medications. *Pharmacoepidem Dr S* 2010;19(6):586-90.
206. Ruigomez A, Martin-Merino E, Rodriguez LA. Validation of ischemic cerebrovascular diagnoses in the health improvement network (THIN). *Pharmacoepidemiol Drug Saf* 2010;19(6):579-85.
207. Meal A, Leonardi-Bee J, Smith C, Hubbard R, Bath-Hextall F. Validation of THIN data for non-melanoma skin cancer. *Qual Prim Care* 2008;16(1):49-52.
208. Lo Re V, 3rd, Haynes K, Forde KA, Localio AR, Schinnar R, Lewis JD. Validity of The Health Improvement Network (THIN) for epidemiologic studies of hepatitis C virus infection. *Pharmacoepidemiol Drug Saf* 2009;18(9):807-14.
209. Haynes K, Forde KA, Schinnar R, Wong P, Strom BL, Lewis JD. Cancer incidence in The Health Improvement Network. *Pharmacoepidem Dr S* 2009;18(8):730-36.
210. Gulliford MC, Charlton J, Latinovic R. Risk of diabetes associated with prescribed glucocorticoids in a large population. *Diabetes Care* 2006;29(12):2728-9.
211. Rait G, Walters K, Bottomley C, Petersen I, Iliffe S, Nazareth I. Survival of people with clinical diagnosis of dementia in primary care: cohort study. *BMJ* 2010;341:c3584.
212. Fardet L, Petersen I, Nazareth I. Prevalence of long-term oral glucocorticoid prescriptions in the UK over the past 20 years. *Rheumatology* 2011;50(11):1982-90.
213. Kirby MG, Schnetzler G, Zou KH, Symonds T. Prevalence and detection rate of underlying disease in men with erectile dysfunction receiving phosphodiesterase type 5 inhibitors in the United Kingdom: a retrospective database study. *Int J Clin Pract* 2011;65(7):797-806.

214. Rodriguez LA, Cea-Soriano L, Martin-Merino E, Johansson S. Discontinuation of low dose aspirin and risk of myocardial infarction: case-control study in UK primary care. *BMJ* 2011;343:d4094.
215. Schlesselman J, Stolley P. *Case-control studies: Design, conduct, analysis*: Oxford University Press, 1982.
216. Maguire A, Blak BT, Thompson M. The importance of defining periods of complete mortality reporting for research using automated data from primary care. *Pharmacoepidemiol Drug Saf* 2009;18(1):76-83.
217. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, Brindle P. Performance of the QRISK cardiovascular risk prediction algorithm in an independent UK sample of patients from general practice: a validation study. *Heart* 2008;94(1):34-39.
218. Walley T, Mantgani A. The UK General Practice Research Database. *Lancet* 1997;350(9084):1097-9.
219. Morris R, Carstairs V. Which deprivation? A comparison of selected deprivation indexes. *J Public Health Med* 1991;13(4):318-26.
220. Vickers D, Rees P. Introducing the area classification of output areas. *Popul Trends* 2006(125):15-29.
221. Townsend P, Phillimore P, Beattie A. Health and deprivation: Inequality and the North. New York: Croom Helm, 1988.
222. Mosaic UK. Mosaic and its uses in research, 2007.
223. EPIC. THIN data from EPIC: A guide for researchers. London, 2009.
224. Ward PR, Javanparast S, Matt MA, Martini A, Tsourtos G, Cole S, et al. Equity of colorectal cancer screening: cross-sectional analysis of National Bowel Cancer Screening Program data for South Australia. *Aust N Z J Public Health* 2011;35(1):61-5.
225. Hakama M, Karjalainen S, Hakulinen T. Outcome-based equity in the treatment of colon cancer patients in Finland. *Int J Technol Assess Health Care* 1989;5(4):619-30.
226. Webber R. The relative power of geodemographics vis a vis person and household level demographic variables as discriminators of consumer behaviour. *Centre for Advanced Spatial Analysis (UCL)* 2004. Available from <http://eprints.ucl.ac.uk/202/>. (Accessed 29 September 2011).
227. de Gruchy J, Robinson J. Geodemographic profiling benefits stop-smoking service. *British Journal of Healthcare Computing and Information Management* 2007;24:29-31.
228. Powell J, Tapp A, Sparks E. Social marketing in action - geodemographics, alcoholic liver disease and heavy episodic drinking in Great Britain.

International Journal of Nonprofit and Voluntary Sector Marketing 2007;12:177-87.

229. Brambilla E, Travis WD, Colby TV, Corrin B, Shimosato Y. The new World Health Organization classification of lung tumours. *European Respiratory Journal* 2001;18(6):1059-68.
230. UK Cancer Information Service. Number of registrations per year and age-standardised incidence rates (ASR) per 100,000 European standard population, by area of residence; Cancer site C33-C34: Trachea, Bronchus and Lung; Period of diagnosis 2003-2007, May 2010 [obtained via personal communication, July 2010].
231. The NHS Information Centre for Health and Social Care and RCP. National Lung Cancer Audit 2009, Report for the audit period 2008
232. Haynes K, Bilker WB, Tenhave TR, Strom BL, Lewis JD. Temporal and within practice variability in the health improvement network. *Pharmacoepidemiol Drug Saf* 2011;20(9):948-55.
233. National Institute for Health and Clinical Excellence. Quality and Outcomes Framework, 2011. Available from <www.nice.org.uk/aboutnice/qof/qof.jsp>. Accessed 29 September 2011.
234. Office for National Statistics hwsgusPav. Cancer Registration Statistics: Registration of cancers diagnosed in 2008, England, 2010. Available at <http://www.statistics.gov.uk/statbase/Product.asp?vlnk=8843>.
235. Office for National Statistics. Smoking and drinking among adults, 2009. A report on the 2009 General Lifestyle Survey, 2011. Available at <http://www.ons.gov.uk/ons/rel/ghs/general-lifestyle-survey/2009-report/index.html>.
236. de Kok IM, van Lenthe FJ, Avendano M, Louwman M, Coebergh JW, Mackenbach JP. Childhood social class and cancer incidence: results of the globe study. *Soc Sci Med* 2008;66(5):1131-9.
237. Louwman WJ, van Lenthe FJ, Coebergh JW, Mackenbach JP. Behaviour partly explains educational differences in cancer incidence in the south-eastern Netherlands: the longitudinal GLOBE study. *Eur J Cancer Prev* 2004;13(2):119-25.
238. Hecht SS. Tobacco smoke carcinogens and lung cancer. *J Natl Cancer Inst* 1999;91(14):1194-210.
239. Glassman AH, Helzer JE, Covey LS, Cottler LB, Stetner F, Tipp JE, et al. Smoking, smoking cessation, and major depression. *JAMA* 1990;264(12):1546-9.

240. Lasser K, Boyd JW, Woolhandler S, Himmelstein DU, McCormick D, Bor DH. Smoking and mental illness: A population-based prevalence study. *JAMA* 2000;284(20):2606-10.
241. Lawrence D, Mitrou F, Zubrick SR. Smoking and mental illness: results from population surveys in Australia and the United States. *Bmc Public Health* 2009;9:285.
242. Salin-Pascual RJ, Rosas M, Jimenez-Genchi A, Rivera-Meza BL, Delgado-Parra V. Antidepressant effect of transdermal nicotine patches in nonsmoking patients with major depression. *J Clin Psychiatry* 1996;57(9):387-9.
243. Picciotto MR, Brunzell DH, Caldarone BJ. Effect of nicotine and nicotinic receptors on anxiety and depression. *Neuroreport* 2002;13(9):1097-106.
244. Dierker L, Donny E. The role of psychiatric disorders in the relationship between cigarette smoking and DSM-IV nicotine dependence among young adults. *Nicotine & Tobacco Research* 2008;10(3):439-46.
245. Anda RF, Williamson DF, Escobedo LG, Mast EE, Giovino GA, Remington PL. Depression and the dynamics of smoking. A national perspective. *JAMA* 1990;264(12):1541-5.
246. Weinberger AH, Pilver CE, Desai RA, Mazure CM, McKee SA. The relationship of major depressive disorder and gender to changes in smoking for current and former smokers: Longitudinal evaluation in the U.S. population. *Addiction* 2012.
247. Schleifer SJ, Keller SE, Meyerson AT, Raskin MJ, Davis KL, Stein M. Lymphocyte function in major depressive disorder. *Arch Gen Psychiatry* 1984;41(5):484-6.
248. Hopwood P, Stephens RJ. Depression in patients with lung cancer: prevalence and risk factors derived from quality-of-life data. *J Clin Oncol* 2000;18(4):893-903.
249. Friedman GD. Depression, smoking, and lung cancer. *Am J Epidemiol* 1996;144(12):1104-6.
250. Vreeburg SA, Hoogendijk WJ, van Pelt J, Derijk RH, Verhagen JC, van Dyck R, et al. Major depressive disorder and hypothalamic-pituitary-adrenal axis activity: results from a large cohort study. *Arch Gen Psychiatry* 2009;66(6):617-26.
251. National Institute for Health and Clinical Excellence. The treatment and management of depression in adults: NICE guideline, 2009. Available at <http://www.nice.org.uk/nicemedia/live/12329/45888/45888.pdf>.

252. Powell HA, Iyen-Omofoman B, Hubbard RB, Baldwin DR, Tata LJ. The association between smoking quantity and lung cancer in men and women. *Chest* 2012;(in press).
253. Huxley RR, Woodward M. Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies. *Lancet* 2011;378(9799):1297-305.
254. Risch HA, Howe GR, Jain M, Burch JD, Holowaty EJ, Miller AB. Are female smokers at higher risk for lung cancer than male smokers? A case-control analysis by histologic type. *Am J Epidemiol* 1993;138(5):281-93.
255. Kreuzer M, Boffetta P, Whitley E, Ahrens W, Gaborieau V, Heinrich J, et al. Gender differences in lung cancer risk by smoking: a multicentre case-control study in Germany and Italy. *Br J Cancer* 2000;82(1):227-33.
256. Linder JA, Sim I. Antibiotic treatment of acute bronchitis in smokers: a systematic review. *J Gen Intern Med* 2002;17(3):230-4.
257. Falcoz PE, Conti M, Brouchet L, Chocron S, Puyraveau M, Mercier M, et al. The Thoracic Surgery Scoring System (Thoracoscore): risk model for in-hospital death in 15,183 patients requiring thoracic surgery. *J Thorac Cardiovasc Surg* 2007;133(2):325-32.
258. Saghir Z, Dirksen A, Ashraf H, Bach KS, Brodersen J, Clementsen PF, et al. CT screening for lung cancer brings forward early disease. The randomised Danish Lung Cancer Screening Trial: status after five annual screening rounds with low-dose CT. *Thorax* 2012;67(4):296-301.