The chronic disease burden

An analysis of health risks and health care usage

Joanne Alder
Leslie Mayhew
Simon Moody
Richard Morris
Rajeev Shah

Foreword

The government's Choosing Health white Paper published in 2004 and Derek Wanless's report "Securing Our Future Health" published earlier the same year brought the nation's health to the forefront of public attention. From around age 50 people become far more susceptible to chronic diseases. As the population will age over the next two decades, it is inevitable that the numbers diagnosed with chronic disease will grow each year. Based on present demographic trends, there will be 26.2m people aged over 50 compared with 20.3m people today.

That people live longer than ever before is a testimony to the success of the NHS as well as to improvements in other areas that affect the quality of life. However, the management of chronic disease through health promotion, regular check-ups, medication, specialist care and spells in hospital, is expensive and likely to become increasingly so. Many people are diagnosed with more than one chronic disease, which raises the issue of how services can be customised for people with different and complex needs. Some diseases are risk factors for other diseases. Therefore it is important to understand how interventions or changes in diet or lifestyle can delay disease onset and hence subsequent disease pathways. We are accustomed to thinking about diseases individually, but this may be impeding our ability to think more widely on a 'whole person' basis about people's needs and strategies to improve overall health.

With these issues in mind I particularly welcome this piece of research, which is a collaborative effort between the actuarial profession, public health, academia and primary care. It deals with five of the main chronic diseases that affect quality of life and have significant resource implications for health services and the economy. I have long believed in the need for better information and intelligence about the factors affecting chronic disease, including socio-economic factors, how diseases interact over the life cycle and the consequences for health service organisation and planning.

This research presents results based on new data, which are analysed in a novel and innovative way using risk analysis and general linear modelling. The methods described give us potentially new ways of thinking about chronic disease and health care both at the local and national levels. It is clear that they will be of practical value to public health practitioners, health care commissioners, health insurers and employers. I look forward to seeing the fruits of further research in due course.

Fiona Adshead, Deputy Chief Medical Officer

About the authors

Joanne Alder is the lead consultant in the healthcare practice of Milliman in London and chair of the Chronic Disease Working Party. She specialises in projecting healthcare costs for private medical expense insurance as well as cost-benefit models for healthcare interventions. She qualified as an actuary in 1999 and is currently studying for a post-graduate diploma in Healthcare Economics at the University of York.

Professor Leslie Mayhew works in the Faculty of Actuarial Sciences and Statistics at Cass Business School, City University, London, where he is also director of The Risk Institute. He was formerly a senior civil servant at the Department of Health and Social Security, Department of Social Security, the Central Statistical Office and the Office for National Statistics. He is an Honorary Fellow of the Faculty of Public Health and of the Institute of Actuaries and director of Mayhew Associates Ltd.

Simon Moody is a Principal in the Health & Benefits practice at Mercer Human Resource Consulting. He advises employers on the financial impact of the implementation of employee benefit schemes, healthcare programmes and wellness initiatives. He qualified as an Actuary in 1998.

Richard Morris is a Business Manager with Munich Re (UK Life Branch), responsible for the client relationship with a number of major life and health insurers. He has also been a member of the Institute of Actuaries' Critical Illness Trends Working Party since its inception in 2001. He qualified as an actuary in 1999.

Rajeev Shah is a Partner in the Insurance Practice at Barnett Waddingham LLP. He specialises in forensic actuarial work and in the management of mortality and financial risks for insurers and qualified as an Actuary in 1998. He is Secretary to the Continuous Mortality Investigation (CMI), Income Protection Committee, and was previously Secretary to the CMI's Critical Illness Committee.

October 2005

Special thanks

The authors would like to acknowledge and express their thanks to the following, for their valuable contributions to this research:
The Actuarial Profession
Islington Primary Care Trust (PCT)
The London Borough of Islington
The Miller Practice (Islington)
St Paul's Road Medical Practice (Islington)
The Village Practice (Islington)
EPIC
Kukuwa Abba (PCT)
Tesfaye Gemechu (PCT)
Gillian Harper
Vladimir Kaishev (Cass)
Francesca Mills
Ganesh Sathyamoorthy (PCT)
Simon Wills (Practice Manager, St Paul's Road Medical Practice)
Graham Wilson
Dagmar Zeuner (PCT)

The Institute of Actuaries

The Institute of Actuaries in England and Wales and the Faculty of Actuaries in Scotland are the two chartered professional bodies for UK actuaries, working closely together across the UK as "The Actuarial Profession". Like most UK professional bodies, the Actuarial Profession has the twin roles of representing members to the outside world and regulating members for the benefit of the outside world. These aims include the regulation of members both in terms of an ethical code and technical standards; the education of new entrants who wish to become actuaries and continuing professional development of existing actuaries; co-operation with government, business, regulators and other professions; innovation through research and debate, to expand the horizons of actuarial knowledge; and promotion of the work of actuaries in general. The Actuarial Profession funds research into areas deemed to be of interest to actuaries. The Chronic Disease Working Party (CDWP) was formed to research issues around the prevalence of chronic diseases and co-morbidities, the probabilities of transition between different chronic diseases and the utilisation of medical services for people with these diseases. The work of the CDWP was carried out in conjunction with researchers at Cass Business School; it was funded by the Institute of Actuaries and Islington Primary Care Trust.

Contents

1 Introduction

1.1 BACKGROUND TO CHRONIC DISEASE

Chronic disease presents a significant cost burden for the UK economy and for the UK's healthcare system[1]. Collectively, coronary heart disease (CHD), cancer, renal disease, mental health services for adults and diabetes cover around 16% of total National Health Service (NHS) expenditure, 12% of morbidity (measured in terms of disability or use of health care services) and between 40% and 70% of mortality (depending on the age group considered)[2].

In July 2000 the UK Government published the NHS Plan[3]. The Plan outlined the programme for a radical and far reaching transformation of the NHS. The challenge was to make a faster, fairer and more convenient service for patients. A 10-year programme of modernisation was established, along with a commitment to a sustained increase in NHS spending.

The centre of the Plan's strategy for providing quality healthcare services is the development of National Service Frameworks (NSFs). These NSFs lay out blueprints for providing high quality integrated services in key areas. Four of these areas focused on specific chronic diseases: coronary heart disease (CHD), renal disease, mental health services for adults and diabetes. The NSFs set out the standards and services that should be available throughout England, specifying both the actions that should be implemented to help reduce incidence of disease, and the high quality treatment and care which should be available for those people who do become ill.

In addition to the development of NSFs, the Government has issued a commitment to enable everyone to live in better health if they chose to do so. In their White Paper "Saving Lives: Our Healthier Nation"[4] they committed to reducing the death rate from heart disease and related illnesses (such as stroke) in those aged under 75 by at least two fifths by 2010. Root causes of ill health are being tackled, including addressing poverty and unemployment, as well as the introduction of legislation aimed at improving lifestyle risk factors such as the use of tobacco and alcohol.

Chronic disease is especially prevalent at older ages and is most likely to afflict those with less healthy lifestyles. The Government Actuary's Department projects that by 2025 more than 20% of the UK population will be over 65. Hence the expectation is that chronic disease prevalence will become an increasingly important issue. In addition to the ageing of the population, there are increasing demands on those of working age and a prospect of having to work until older ages. The combination of these two facts put greater value on health and on leading a healthy lifestyle.

Against this background, the Government issued a White Paper in 2004, which set out the key principles for supporting the UK population to make healthier and more informed choices with regards to their health.[5] The priorities for action which were set out in the Choosing Health paper included:

- Reducing the numbers of smokers
- Reducing obesity and improving diet and nutrition
- Increasing exercise
- Encouraging and supporting sensible drinking
- Improving sexual health
- Improving mental health

The paper also commits to tackling these priorities in conjunction with the notion of promoting and maintaining improved health in the workplace.

1.2 WORKING PARTY OBJECTIVES

Wanless's final report "Securing Good Health for the Whole Population" was published in February 2004[6]. In his discussion of the New General Medical Services Contract, Wanless highlighted that one of its provisions is for Primary Care Trust (PCT) funding of information management and technology systems. These would "provide potential for the development of practice and PCT based patient registers that could be developed to record information on disease, medication and risk factors." Wanless went on to state how this information could be used to improve chronic disease management, as well as improve health and prevent disease.

In addition, Wanless's report provided recommendations for further research and evaluation programmes. He proposed that "an experiment should be established across primary care to assess the benefits of additional resource in information systems, in monitoring risk, and in services".

With this in mind the Chronic Disease Working Party (CDWP) was formed with the primary objective of using local PCT data, combined with data from Local Authorities (LAs) to perform some innovative analysis around risk factors and pathways of chronic diseases. Following on from an earlier project for the PCT on coronary heart disease, the CDWP's aim was to develop models that could be used to guide those seeking to manage the causes and treatment of chronic diseases, and inform

2 The chronic diseases

debate about the cost burden of these diseases. The CDWP's work was aided through access to a further data set which consisted of a sample of around four million historical and current GP records, with details of GP visits, prescriptions, diagnoses and in-patient spells, among other things.

Our key objectives:

- To learn about chronic diseases, their prevalence, and progression, including their impact on health care services
- To further actuarial knowledge in this area using new techniques to evaluate co-morbidity and risk
- To avoid the 'silo-mentality' that besets and constrains research in this area through a more integrated approach
- To provide new results and practical tools that can be applied locally and nationally

Availability of time and resource meant that we initially limited our focus to five chronic diseases: Coronary heart disease (CHD); stroke; hypertension; diabetes; and chronic obstructive pulmonary disease (COPD).

These five chronic diseases were chosen because they consume a large part of current healthcare resources in the UK.

2.1 OVERVIEW OF EXISTING LITERATURE

Given the vast body of available literature, it is impractical to provide a summary of all the research already carried out in these areas. We have instead provided a brief synopsis based on our review of some of this literature.

Research on particular socio-economic factors (such as social class and education), as risk factors for the above diseases is already well covered in existing literature. However, our review indicated that little research had yet been carried out on, for example, the specific impact of housing and household size. The Islington PCT data allowed analyses of these factors. Further detail and the results are given in section 4.

A key reason for carrying out this review was to identify the areas not yet covered in previous research, as well as areas that could benefit from further research. Our review highlighted the lack of existing literature on the impact of these diseases on the demand and costs of healthcare - where available, such literature has tended to focus on the demand and costs of hospital treatment. This encouraged us to analyse the consequent increases in demand for GP visits and medical prescriptions arising from the above diseases. The results of these analyses are covered in section 5.

A secondary reason for carrying out this review was to provide a comparison of the national picture of disease prevalence by risk factors with the prevalence in two important data sets. The first data set is based on GP medical records from a sample of three practices in Islington, London that were subsequently linked to data on housing. The second data set is the THIN (The Health Improvement Network) data set, supplied by EPIC[7]. Appendix 6 provides a brief example of the comparisons of the output from our analysis to the results reported in the existing literature.

The following part of this section gives an overview of the five chronic diseases we examined, focusing on the following aspects:

- Their prevalence in the UK
- The genetic, medical and lifestyle risk factors affecting the incidence of these diseases
- Their impact on mortality
- The co-morbidity of these chronic diseases
- The estimated burden on the NHS and UK economy

The chronic disease problem:

- The annual cost of treating CHD is put at £3.5bn for the UK with additional costs of £3.1bn due to lost working days
- The treatment of stroke is estimated to be in excess of £2.3bn each year
- Hypertension, although under-reported, costs £0.8bn a year to treat
- Diabetes results in long run complications that cost an estimated £1.3bn a year

2.2 CORONARY HEART DISEASE (CHD)

2.2.1 Prevalence

Coronary heart disease (also called ischaemic heart disease) is a disease in which the arteries supplying blood to the heart are seriously narrowed by atherosclerosis, causing angina and, sometimes, a heart attack. Prevalence rates increase with age, with around 1 in 4 men and 1 in 5 women aged 75 years and above living with CHD. For minority ethnic groups, prevalence rates are higher for South Asian males and lower for black Caribbean and Chinese males. Ethnic variation for females is lower with only Chinese females having significantly lower prevalence than the general population. Morbidity from CHD is rising. Approximately 2 million people in the UK suffer from angina and almost 260,000 people have a heart attack each year[8].

2.2.2 Risk factors

Genetic
CHD is known to run in families and certain ethnic minorities are more predisposed to this disease. The risk is higher for males than females but the differential reduces with age.

Medical
CHD is closely linked with diabetes, high cholesterol, hypertension and obesity. High cholesterol can increase the risk of CHD by up to 5 times and diabetes can increase CHD risk by up to 8 times[9]. Body shape is an important sub-factor for obesity with "apple-shaped" individuals with extra fat at the waistline facing a higher CHD risk than "pear-shaped" people with heavy hips and thighs[10].

Lifestyle
The primary lifestyle risk factors for CHD are smoking, the level of physical activity, diet and alcohol consumption. Smoking has the highest impact, increasing the risk of CHD by as much as 15 times[11]. However, the effects of smoking are not all permanent; the risk of CHD being halved one year after smoking is stopped.

Relative importance of risk factors
While the primary risk factors are genetic, lifestyle has a significant effect, particularly where the genetic or medical factors already predispose individuals to higher risk of contracting CHD.

2.2.3 Impact on mortality

CHD is the major cause of mortality for males in the UK. It kills more than 110,000 people a year and is the most common cause of premature deaths. CHD, in the absence of other diseases, increases mortality rates by between 100% and 450% for males. The impact is more severe for those affected by CHD at younger ages. CHD has a lower impact on mortality rates for females, increasing them by between 100% and 250%[12].

CHD highlights the social inequalities in health, with the premature CHD death rate for unskilled working men being 58% higher than for men in professional or managerial occupations. There are also ethnic variations with those born in the Indian sub-continent significantly more likely to die from heart disease than for the UK as a whole.

2.2.4 Co-morbidity

Hypertension and high cholesterol are common underlying medical risk factors for CHD, diabetes and strokes. The presence of CHD leads to a higher risk of contracting diabetes and/or strokes.

2.2.5 Burden

CHD costs the healthcare system in the UK around £3.5 billion a year, and a further £3.1 billion a year in economic costs (e.g. absence due to death, illness or caring for others with CHD)[13].

The Wanless Report estimated that implementing the CHD NSF and to go further in raising quality – for example implementation of the recommendations by the National Institute for Clinical Excellence - would cost an additional £2.4 billion a year by 2010/2011. This would mean roughly doubling the NHS's current expenditure on CHD[14].

2.3 STROKE

2.3.1 Prevalence

A stroke is an interruption of the blood supply to any part of the brain. A stroke is also known as 'cerebral infarction', 'cerebral haemorrhage' or simply 'brain attack'. Ischaemic stroke, the most common type, usually results from clogged arteries, a condition called atherosclerosis. Each year over 130,000 people in England and Wales have a stroke[15]. While the general prevalence of stroke at all ages, as reported by individuals in England, is low at 2.3% in men and 2.1% in women, it increases quickly with age so that by age 75, around 10% of men and women will have suffered a stroke[16]. Prevalence is higher for certain ethnic minority groups including South Asians, Africans and black Caribbeans.

2.3.2 Risk Factors

Genetic
Family history is known as a risk significant factor for strokes. Males are at a higher risk than females, particularly at younger ages.

Medical
Hypertension is the primary medical risk factor, with the risk of stroke increasing in line with the elevation of blood pressure. A history of previous strokes, CHD or diabetes also increases the risk of suffering strokes in the future. Obesity, while not as important a factor for strokes as compared with CHD, also increases the risk of strokes, especially if the obesity results in an "apple shaped" body.

Lifestyle
Smoking, physical inactivity, and high levels of alcohol consumption, can each double the risk of stroke. Diet is another lifestyle factor with the risk increasing in line with the intake of salts and fatty foods. Substance abuse, particularly cocaine and amphetamines, is known to increase the risk of strokes. For women, the use of oral contraceptives has also been linked to increased risk of strokes.

Relative importance of risk factors
Increasing age is the most important risk factor with 90% of strokes occurring in those aged above 55[17]. Other factors are less important but still significant.

2.3.3 Impact on mortality

In 2003 there were over 65,000 deaths from stroke in the UK.[18] Mortality rates of those affected by strokes are more than two times higher than for healthy lives[19].

2.3.4 Co-morbidity

Strokes are not a significant risk factor for other chronic diseases. However, the risk of stroke is more likely to be increased by the presence of other chronic diseases, especially at older ages.

2.3.5 Burden

The cost of stroke to the NHS is estimated to be over £2.3 billion each year[20] and the costs are expected to rise in real terms by 30% by the year 2023[21].

Many of the interventions and treatments proposed in the CHD NSF will also help reduce the incidence of stroke, as these diseases are closely linked.

2.4 HYPERTENSION

2.4.1 Prevalence

Hypertension is the result of persistently high arterial blood pressure. Hypertension may have no known cause (essential or idiopathic hypertension) or be associated with other primary diseases (known as secondary hypertension). Hypertension is also considered a risk factor for the development of heart disease, peripheral vascular disease, stroke and kidney disease.

In the UK there are about 16 million people with blood pressure higher than 140/90mmHg (the level used to diagnose high blood pressure)[22]. The prevalence of diagnosed hypertension in England is 34% in men and 31% in women. Less common in younger adults, prevalence rates increase with age, so that by age 75, around two in three men and three in four women are living with hypertension. Again, prevalence is higher for certain South Asians and black Caribbean ethnic groups[23].

2.4.2 Risk Factors

Genetic
Family history is known to be a significant risk factor for hypertension. Males are at a higher risk than females at younger ages, while females are at higher risk than males at the oldest age groups (over age 75).

Medical
The risk of hypertension is increased in the presence of diabetes, high blood cholesterol, and obesity. Body shape is an important sub-factor when considering the impact of obesity. Secondary hypertension can also develop

following diseases of the kidneys and adrenal glands, or from the use of medicines such as steroids to treat other conditions.

Lifestyle
A key lifestyle risk factor is diet, particularly the intake of salt and fat. High levels of alcohol consumption and smoking also increase the risk of hypertension. For women, the use of oral contraceptives can elevate blood pressure, resulting in hypertension.

Relative importance of risk factors
The key risk factors for hypertension are genetic predisposition and increasing age. However, lifestyle factors can act to significantly change this risk. The risk of secondary hypertension is low and underlies less than 5% of total cases of hypertension.

2.4.3 Impact on mortality

Though hypertension is not a large direct cause of mortality, by acting as a risk factor for other chronic diseases, it still significantly increases mortality rates. Mortality rates for people with hypertension are up to twice the level of those without hypertension, with the increased severity of hypertension leading to higher mortality rates[24].

2.4.4 Co-morbidity

Hypertension is a well-known common underlying medical risk factor for CHD, diabetes and stroke. A high proportion of people with hypertension will eventually suffer from CHD or stroke, even though this risk is reduced by suitable treatment for hypertension.

2.4.5 Burden

The cost to the NHS of prescriptions for anti-hypertensives was around £840m in 2001, nearly 15% of the total annual cost of all primary care drugs[25]. However it is widely believed that hypertension is hugely under-reported, and hence the actual cost to the NHS and society is likely to be much larger than this estimate.

2.5 DIABETES TYPE II

2.5.1 Prevalence

Type II diabetes is a metabolic disorder, which is often associated with obesity and stress and usually strikes adults. Unlike Type I diabetes, which often begins in childhood or the young adult years, Type II diabetes is non-insulin dependent and the disease may be controlled through diet and exercise. The prevalence of reported Type II diabetes in England is over 4% in men and 3% in women, but this is believed to understate true levels. Undiagnosed prevalence of diabetes is estimated at 3% for men and 0.7% for women aged over 35[26]. The prevalence of diabetes in men increases from less than 0.5% for ages 16-34 to over 10% for those aged above 75. Prevalence rates are higher for certain South Asian and black Caribbean ethnic groups.

2.5.2 Risk Factors

Genetic
Diabetes is known to run in families and certain ethnic minorities are more predisposed to this disease. Prevalence rates for women are slightly lower than for men at most ages but the age trends are similar.

Medical
High body weight is the key medical risk factor with 80% of diabetics being overweight. As for CHD, hypertension and stroke, body shape is an important sub-factor[27]. However, low birth weight is also a significant medical risk factor. Vascular disease including CHD also increases the risk of diabetes - many of the risk factors for diabetes are shared with CHD.

Lifestyle
Poor diet, particularly low-fibre and high fat content, and low levels of physical activity enhance the risk of diabetes, especially at younger ages.

Relative importance of risk factors
While genetic factors are the key determinant to the predisposition to diabetes, the age of onset is determined primarily by medical and lifestyle factors.

2.5.3 Impact on mortality

The presence of Type II diabetes increases mortality rates by between 50% and 300%. The impact on mortality rates increases with time since diagnosis[28].

2.5.4 Co-morbidity

Diabetes is a known risk factor for CHD as it magnifies the effect of other known risk factors for CHD, such as raised cholesterol levels, smoking, hypertension and obesity. For men, the presence of diabetes increases the risk of CHD between two to four times while for women this risk is increased by three to five times.

2.5.5 Burden

The cost to the NHS of diagnosed diabetes is around £1.3 billion a year. Most of this cost arises from the long-term complications which result from a lack of proper management. The Wanless Report estimated that it would cost an additional £600 million a year to implement the diabetes NSF[29]. The additional cost of undiagnosed diabetes is not known, but is believed to be significant.

2.6 CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD)

2.6.1 Prevalence

Chronic obstructive pulmonary disease (COPD) is defined to be any disorder that persistently obstructs bronchial airflow. It mainly involves two related diseases - chronic bronchitis and emphysema. Both cause chronic obstruction of air flowing through the airways and in and out of the lungs. Nearly 900,000 people in the UK have been diagnosed with COPD, and about half as many again are thought to be living with undiagnosed COPD. In 2001/2002 there were nearly 100,000 hospital admissions per annum for COPD in the UK, representing almost 1m annual bed days[30].

The exact prevalence of COPD is difficult to determine because of problems with definition and coding. It can also be difficult to differentiate between COPD and chronic severe asthma, and where only mild to moderate disease is present, it may not be identified as COPD.

2.6.2 Risk Factors

Genetic
Although the main risk factor is smoking, genetic factors determine the susceptibility of smokers to COPD. Chinese and Afro-Caribbean ethnic groups have lower susceptibility to COPD. After adjusting for smoking levels and occupational exposure, there is little relative difference in susceptibility between males and females.

Medical
COPD risk is increased for non-smokers if there is a high frequency of respiratory infections in childhood.

Lifestyle
Smoking is the key risk factor, with about 15% of one-pack-per-day smokers, and 25% of two-pack-per-day smokers developing COPD if they continue to smoke[31]. Other risk factors are environmental.

Relative importance of risk factors
This disease is caused mainly by lifestyle, since smoking is so heavily implicated. For non-smoking related COPD, the main risk factors are medical factors and the socio-economic factors determining exposure to air pollution and hazardous working conditions.

2.6.3 Impact on mortality

Depending on severity of COPD, mortality rates of those affected by COPD can be between 50% and 300% times higher than for healthy lives[32].

2.6.4 Co-morbidity

Smoking is a common underlying medical factor for CHD, strokes and COPD and the presence of COPD leads to a higher risk of contracting CHD and/or strokes.

2.6.5 Burden

The total economic cost to the NHS is estimated to be £492m in direct costs, rising to £982m including indirect costs. As well as these costs, it has been estimated that as many as 22 million working days are lost each year due to COPD[33].

3 Scope of analysis

3.1 INTRODUCTION

The key learning from the background material in section 2 of this report and wider information within the public domain are as follows:

- **These diseases continue to present a significant burden on UK healthcare resources.** The five diseases included in this research cost the healthcare service over £8bn per annum, and this excludes the economic cost to the UK economy through sickness absence.
- **The five diseases are heavily inter-related** e.g. diabetes raises risk of CHD in women by up to 8 times[34]
- **Lifestyle** (diet, exercise, smoking, drinking etc) **and socio-economic status** (occupation, location, financial wealth etc) **play a significant part in their incidence**

Hence the working party firmly agrees with the Wanless Report that a study of the various relationships between socio-economic status, lifestyle and the prevalence of these chronic diseases should provide valuable insights on their management, and potentially reduce their incidence and severity over time.

This working party study consists of three phases with each phase building on the work in the previous phase.

3.2 PHASE 1

The first phase is based on an investigation of CHD. This was conducted in conjunction with Islington PCT and involved three local Islington GP practices and an analysis of over 24,000 patient records. This research was funded by Islington PCT and included work on the spatial modelling of CHD risk by Islington neighbourhood.

3.3 PHASE 2

This phase was funded by the Institute of Actuaries and is based on an extension of Phase 1 into four other chronic diseases. The work involved a much more detailed analysis of disease pathways, as well as comparable risk analyses to those undertaken for CHD in Phase 1. Uniquely, it has involved an investigation of co-morbidity based on individual likelihood of contracting more than one chronic disease. It also included a benchmarking of the Islington PCT findings against other published research findings.

Phases 1 and 2 of the analysis involved novel techniques for linking together information from diverse sources to provide a more rounded description of patients, their physical health and living conditions. The data were then used to assess the risk of chronic disease according to various different risk factors. A more detailed description of the methodologies used in these phases and the results is contained in section 4 of this paper.

3.4 PHASE 3

The final phase is based on an analysis of the "THIN" data set (see also section 2.1 and fuller description of the data set in Appendix 7). This data set is national in scope and includes details of GP patient registrations, medical records, prescription drug records and therapeutic values (height, weight, blood pressure and smoking status), visits to surgery, referrals and in-patient spells. Although this data set does not contain any social information about patients, its advantage is its size and the fact that it provides information about 'completed' lives (i.e. patients who have subsequently died) and also the use of health services.

Phase 3 was split into two parts. The first part looked at the use of primary medical services amongst chronically diseased populations. The second part of the analysis uses survival rates to ascertain the probability of survival for a patient diagnosed with a chronic disease and life expectancy based on age of diagnosis. The analysis of the THIN data set is set out in sections 5 and 6 of this paper.

4 Risk ladders and socio-economic factors

4.1 METHODOLOGY

The concept behind the first two phases of the research involved the matching of GP records to administrative data sources, such as housing tenure and council tax bands (a proxy for wealth), to evaluate the prevalence and risk of chronic disease. We were fortunate to be given the necessary permission to link data in this way by Islington PCT, London. The other key partners in the project were three participating GP practices in the PCT, with a combined practice population of 24,401 patients and the local authority. Data protocols were agreed with each data provider for the purposes of the project and, after linking, the data were anonymised.

Data from the different sources (see Table 4.1) were matched to the local property gazetteer, which is simply a current list of all residential properties in the borough. For each record of every database, geographical references were extracted using an address-matching algorithm to link the address on a database to the address on the gazetteer. The extracted x,y co-ordinates were then incorporated within a Geographical Information System (GIS). The value of geographically referenced data is that it can be used to create maps as well as tables. For example, maps can give information that is helpful for locating services or targeting resources.

The proportion of the population in each risk group with a chronic disease is identified by the given risk factors and gives us the prevalence or 'risk' to that sub-group. The influence of each risk factor across all sub-groups is then estimated separately using logistic regression techniques, so that we end up knowing not only the risk factor combinations associated with high risk groups, but also the overall influence of each risk factor. A 'risk factor' does not have to cause a disease to be associated with it. Thus, a factor such as 'housing' is considered to be a proxy for other possible unobserved lifestyle influences, and should not be interpreted as a direct cause of chronic disease in itself[35].

The GP patient records were sorted into 32 mutually exclusive categories according to the chronic diseases diagnosed ranging from 'healthy' (no disease) to 'all' diseases (5 diagnoses). Chronic diseases defined for the purposes of this research were CHD, hypertension, diabetes, stroke, and COPD and were extracted according to the appropriate Read codes[36] (see also section 5). Other data extracted from GP records at the same time included patients' Body Mass Index (BMI) and smoking status (current or lapsed), gender and date of birth. Each disease combination was then further sorted according to the year (age) of diagnosis into time ordered sequences or 'disease pathways'.

Table 4.1: Data sources

Data set	Subject matter	Sample	Source	Main variables extracted
1 Land and local property gazetteer	Housing - approx 105,000 records	All residential properties in Islington	Islington Local Authority	Residential addresses, Unique Property Reference Numbers (UPRNs), grid references
2 Council tax bands by property	Housing -approx 86,900 records	All residential properties in Islington	Islington Local Authority	Council tax band by property in Islington
3 Locations of health providers	Local health services 79 records	All NHS providers	Islington Local Authority	Names and addresses of services
4 GP register	Register of Islington residents registered with GPs 230,000 records	All Islington residents	PCT	Date of birth, gender, address
5 GP practice data	24,000 records	3 Islington GP practices	GP practices	Information on date of diagnosis of any chronic disease, smoking status, date of birth, gender etc.
6 Digitised ward boundaries and major roads	Islington geography	All relevant features	Islington Local Authority	Boundary information, properties etc.

4.1.1 Risk ladders

A risk ladder is an analytical tool to assist in the analysis of the risk or probability of an event (such as being diabetic) and is based on the complete decomposition of a population according to selected risk factors. For each risk factor combination or sub-group the number of patients is established along with the number of patients that have been diagnosed with a given disease. The ratio of those with the disease to the number in the sub-group is defined as the 'risk' exposure, given the particular risk factor combination[37].

Imagine there are five risk factors relating to a particular disease. There are hence 2^5 or 32 risk factor combinations. The general rule is that N, the number of factor combinations, equals 2^n where n is the number of risk factors, or:

$$N = \sum_{r=0}^{r=n} \frac{n!}{r!(n-r)!}.$$

In pathway analysis, it is necessary to establish the time order of the risk factors in each risk factor combination (assuming the risks are based on events e.g. date of diagnosis). The general rule for the number of pathways is given by:

$$N = \sum_{r=0}^{r=n} \frac{n!}{(n-r)!}.$$

If the number of patients with a particular combination of risk factors, i, is k_i and the number in i observed to have a particular disease is the observed risk in that subgroup of finding the disease is defined as:

$r_i = m_i / k_i$ with a standard error given by $\sqrt{r_i(1-r_i)/k_i}$.

In the illustrative risk ladder examples set out in the results section below and in Appendices 1 to 4, up to 6 risk factors have been used. Risk ladders were produced for each disease with typical risk factors such as gender, age, BMI, smoking status, housing type and the co-presence or otherwise of other chronic diseases. In the text we concentrate on CHD; other diseases are considered in the Appendices.

The number of observations in each risk category can range from a few to thousands, whilst risk by definition can range from zero to one. Appendix 5 contains a convenient graph showing the standard error for different values of r and k expressed in percentages, which can be used in conjunction with the risk ladders as a check on the relative accuracy of different risk estimates. So for example, if the observed risk is 27% and the sample size is 40 cases then the standard error of the estimate is ±7% with a 68% probability that the true value lies within these limits (see point P, Figure A5 in Appendix 5).

This graph is based on the assumption of a normal approximation to the binomial distribution. Occasionally very high values of risk are obtained (up to 100%) for small samples (below 5, say). These must be viewed cautiously since the normal approximation is no longer valid and the binomial itself must be used[38]. To some extent small numbers in risk categories can be avoided by limiting the number of risk factors in the analysis to only the most important. An example is the COPD risk ladder (see Appendix 4).

Logistic regression was used to estimate the association of each risk factor within the given diseases and therefore to predict risk outcomes for each factor combination.

The model used has the following general form:

$$\text{logit}(r_i) = f(k_i, X_i, u_i)$$

where:

$$\text{logit}(r_i) = \log\left[\frac{r_i}{1-r_i}\right]$$

X_i is a binary variable indicating the presence or absence of X in the i^{th} risk group (e.g. a BMI>30), k_i, the number of members in each group, is a weighting factor, u_i is the error term, and f is assumed to be linear in X. Model outputs include predicted values of r_i and regression coefficients, β_i which are used to calculate the odds ratios given by e^{β_i} for each independent variable. All model regression coefficients were significant at the 99% level of probability.

A risk 'tree' is an alternative way of presenting risk factors based on the systematic partitioning of risk versus the number of cases at risk. Whereas a risk ladder is tabulation based on 2^n mutually exclusive risk factor combinations, a risk tree is a decomposition of the population into risk subsets according to the presence of different risk factors starting with the whole population. Examples of all these techniques are given below in the results sections.

4.1.2 Maps

Maps used to illustrate aspects of this study were created using ESRI ArcGIS, a standard GIS software package. Population and prevalence rates for chronic diseases were derived from the GP register of Islington residents and the

sample of GP practices medical records (24,401 cases). Risk factor maps are based on the risk factor analysis described in this report using risk factors that contain geographical referencing, which allowed them to be mapped.

4.1.3 Methodological issues arising

The use of GP data in this application is novel and raises several methodological issues, including whether it is reasonable to generalise the results to a larger population. Our extensive literature search (see section 2) established correspondences between our results and those of others in the literature and from surveys. We found that the prevalence rates for each disease based on diagnosis in the Islington practices generally fell within the range expected when compared with the Health Survey for England, with the exception that rates for older ages appeared to be higher than expected for females and lower for males (see Table 4.2). Where similar risk factors had been used in other studies we found a general concordance, but the range of risk widened where we had used risk factors such as housing as general markers for wealth and income.

We also carried out other comparisons with the THIN data set, which is based on a much larger sample of patients (up to 2m). We concluded that the quality of the Islington GP data was good as far as one could tell, although a bias towards under diagnosis of chronic diseases is likely to be present to a degree, as some patients may be unaware that they have a disease (e.g. undiagnosed diabetes).

4.2 RESULTS

4.2.1 Prevalence of chronic diseases

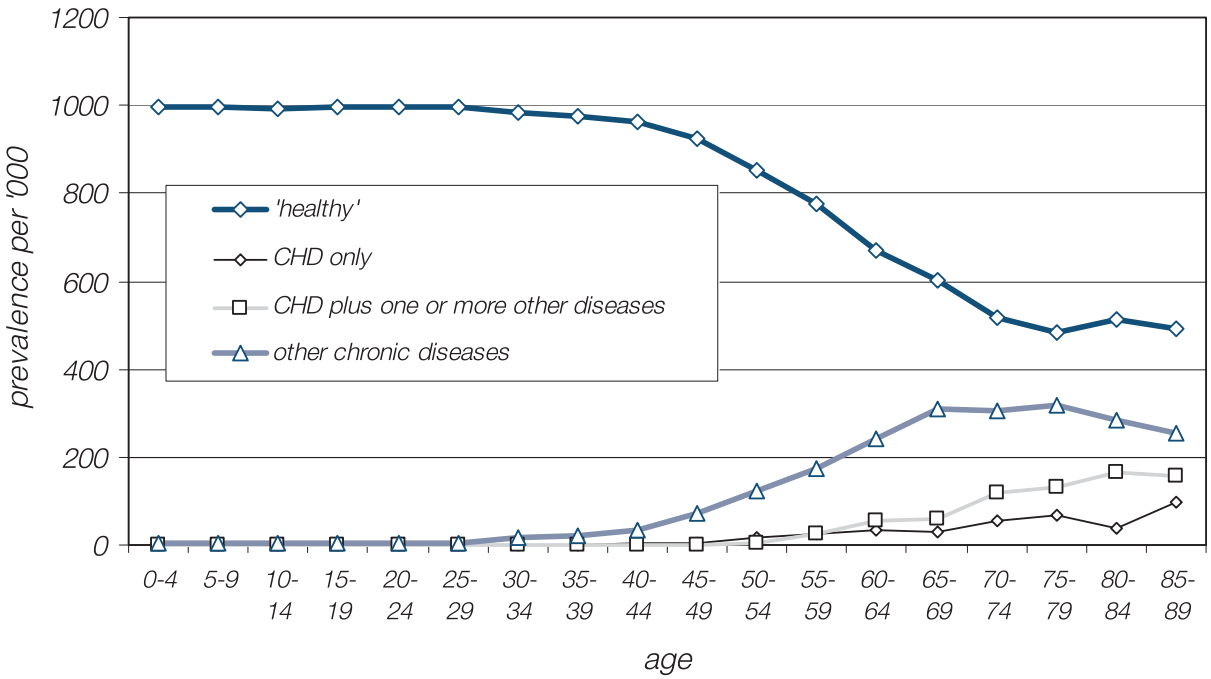
Figure 4.1 based on the GP data shows the prevalence of chronic disease rising with age. It shows that by age 70 the majority of the population have been diagnosed with one or more of the five specified chronic diseases. Figure 4.1 shows that CHD occurs more in conjunction with other diseases than it does on its own, especially after age 65. Chronic diseases other than CHD are more prevalent at all ages due largely to the widespread occurrence of hypertension and diabetes at ages from 40 upwards. Further analysis shows that:

- Females are slightly more likely to suffer from any chronic disease than males at age 70+
- Males are more likely to suffer from CHD than females at age 70+
- Up to age 50 there is a more than 80% chance of being free of any of the given chronic diseases

Table 4.2: Prevalence rates in specified age groups males (M) and females (F) (%)

Disease	M&F 50+	M&F 65+	M 50+	M 65+	F 50+	F 65+
CHD	7.1	13.6	8.4	15.4	5.6	11.9
Stroke	3.8	7.9	3.6	7.2	4.1	8.6
Hypertension	23.1	35.4	20.5	31.5	25.8	39.1
Diabetes	7.0	10.3	7.4	11.3	6.6	9.5
COPD	2.7	5.4	2.4	4.2	3.1	6.5

Figure 4.1: The prevalence by age of chronic disease-free males, males with CHD and no other chronic disease, males with CHD and any other chronic disease, and males with other chronic diseases other than CHD



In this practice population of 24,401 people:

- 61 per 1000 have one chronic disease; 17 per 1000 have two chronic diseases; and 4.4 per 1000 have three or more diseases
- The most common disease is hypertension (54% of diagnoses), followed by diabetes (18%), CHD (14%), stroke (8%), and then COPD (6%)
- Over 30% of hypertension cases were diagnosed before age 50; this compares with less than 12% of CHD and stroke cases and 23 % of COPD cases

4.2.2 Co-prevalence of CHD and other chronic diseases

Normally, prevalence rates for different diseases are only available for one disease at a time. In our research, we were able to measure the co-prevalence of different chronic disease combinations. Measurement of co-prevalence is important in measuring the overall burden of disease and (indirectly) disability; it is also likely to be important from the point of view of health service design and delivery, especially where different specialist clinicians may be involved. Persons with several diseases are also more likely to be more frequent users of health care services, as we show later in this report.

Figure 4.2 concentrates on combinations involving CHD. It shows CHD is prevalent on its own in 39% of cases but in the other 61% of cases it combines with at least one other disease. The most common combinations involve hypertension, diabetes or both. Combinations involving more than two diseases are relatively rare, occurring in 12% of all cases in this example.

4.2.3 Occurrence of chronic disease by age

The occurrence of chronic disease by age shows further distinctive patterns. Figure 4.3 shows the number of diagnoses by age and disease type. The curves have been smoothed for clarity. The frequency of occurrence of diagnoses at older ages declines because as people die there are fewer living cases remaining. Hypertension is most often found in the population followed by diabetes and then CHD. Hypertension and diabetes are more likely to be found at younger ages, the average age of diagnosis for which are 56 and 54 years respectively. The average age of diagnosis for CHD is 63 years, stroke 64 years and COPD 60 years.

Figure 4.2: The co-prevalence of CHD with other chronic diseases including hypertension, stroke, diabetes, and COPD

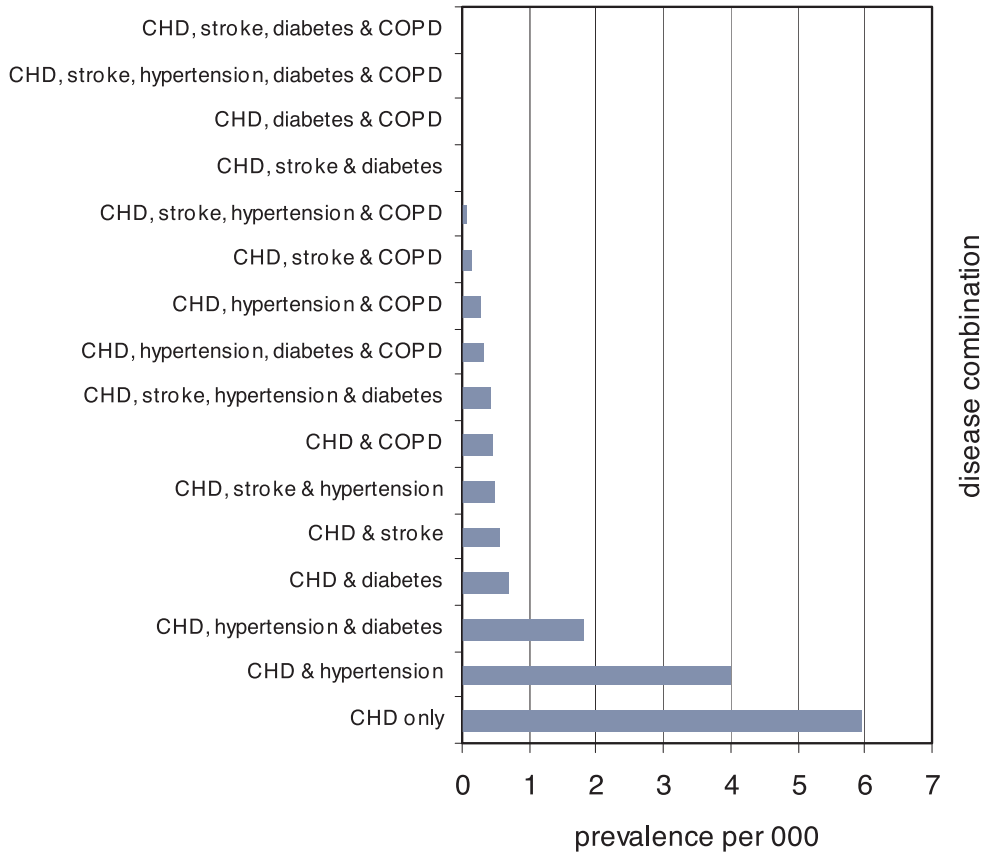
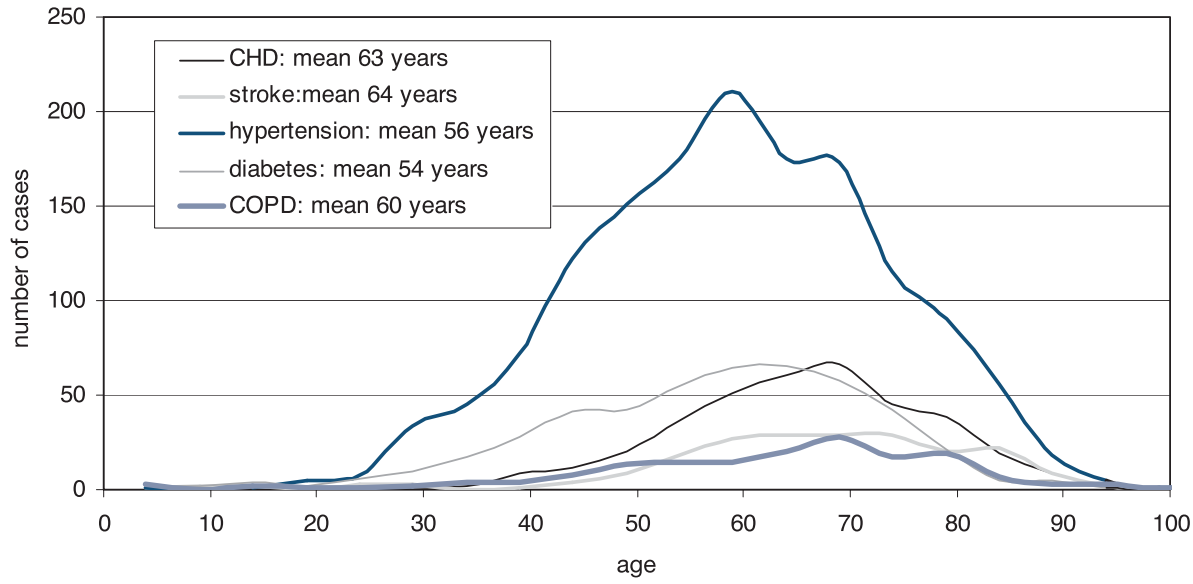


Figure 4.3: The occurrence with age of CHD, hypertension, stroke, diabetes and COPD



4.2.4 Number of chronic diseases by age

The number of people with one or more diseases also tends to rise with age, as might be expected, but then falls away as people die. Figure 4.4 shows that the frequency of disease combinations involving more than three diseases is very rare in this population. The average age of occurrence of any diseases is 55 years. For two diseases it is 64 years, three diseases it is 66 years and four diseases the average age is 70 years. Clearly there is a wide dispersion about the mean ages.

4.2.5 Typical disease pathways

Table 4.3 shows the pathways to CHD in chronological sequence. The letters represent diseases as follows:

A- CHD
B- Stroke
C- Hypertension
D- Diabetes

For example a pathway labelled '3021' records the number of times in our sample that the sequence diabetes, followed by hypertension, followed by CHD was observed. The zero in this case means that diagnosis B, stroke, was not part of this particular pathway.

A separate column indicates the number of patients with a particular combination of chronic diseases observed in the sample. Another column shows the ratio of the observed number of cases to the expected number of cases. A high value of the O/E ratio shows that the number of occurrences in the given sequence is relatively unexpected compared to a randomly generated sequence based on the prevalence of each disease in the population.

Values that are significantly different from '1' might indicate a causal chain of events, with one diagnosis leading to another. Values that are close to 1 could indicate a randomly generated sequence with no apparent causal chain. Note however the small sample sizes; such results would need to be validated on a larger scale before firm conclusions could be drawn.

Table 4.3 shows that the following pathways occur more often than would be suggested by chance:

3 diseases
– Diabetes, hypertension, CHD
– Diabetes, CHD, hypertension
– Hypertension, diabetes, CHD
– CHD, hypertension, diabetes
– CHD, diabetes, hypertension
– Hypertension, CHD, diabetes

2 diseases
Hypertension and CHD occur frequently together, although there is little difference in the chance of hypertension being diagnosed before CHD (59 cases versus 46 cases).

Stroke occurs relatively infrequently in any sequence. About half of all CHD cases occur with other diseases. CHD by itself occurs less frequently than would be suggested by chance (156 cases).

Other
There were just 10 examples of pathways based on 4 diseases, but no common pattern among them except that hypertension or diabetes tended to be the first diagnoses. The sequence 4213, hypertension, stroke, diabetes, CHD for example was observed the most times, 3.

4.2.6 Body Mass Index and CHD

Obesity is a recognised risk factor for several chronic diseases. Obesity and overweight status are measured using the Body Mass Index (see box for definition). Figure 4.5 shows how mean BMI varies with age for males and females combined. It shows that the BMI increases with age from between 15 and 20 at age 10 and climbs steadily levelling out at around 27 between the ages of 60 and 70 before declining. Because of smaller samples at older ages mean BMI tends to exhibit greater variance.

Also indicated in Figure 4.5 is the comparative BMI of people diagnosed with CHD. As the CHD population is smaller the variance is higher. The results show that patients with CHD generally have an elevated BMI at ages

The Body Mass Index (BMI) is an indication of obesity in a population and is also a risk factor implicated in chronic diseases such as CHD

Body Mass index or BMI is defined as:

$$BMI = \frac{\text{weight in kilograms}}{(\text{height in metres})^2}$$

Below 18.5 underweight
18.5 – 24.9 normal
25.0-29.9 overweight
30.0 and above obese

Figure 4.4: The occurrence with age of one, two, three, and four diagnoses

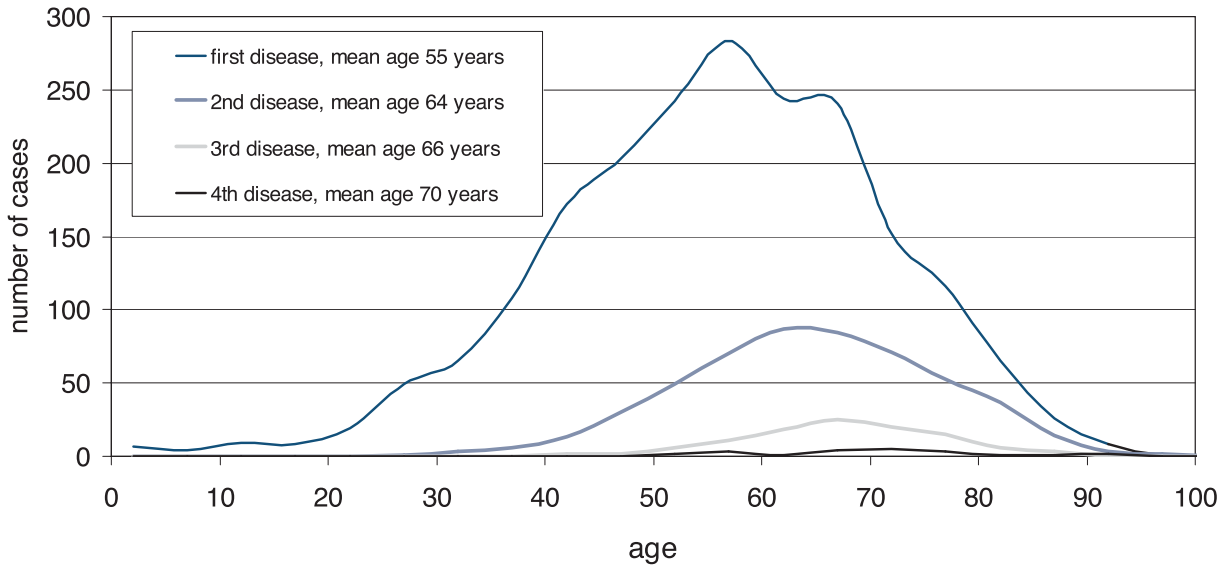
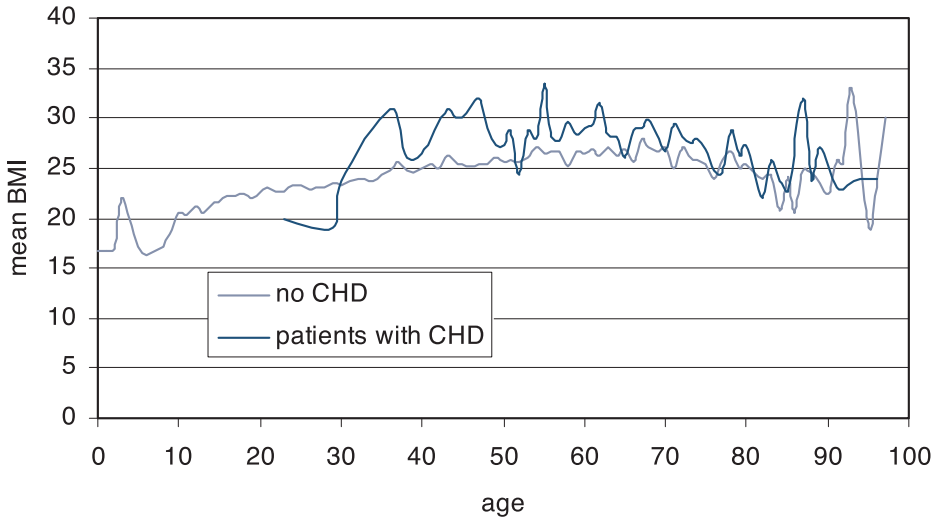


Table 4.3: Frequency of occurrence of chronic disease pathways
(Note: pathways with less than 6 observations are omitted. 0000 indicates disease-free cases.)

ABCD	Observed/expected	Cases observed
3021	168.8	12
2031	154.8	11
3012	126.6	9
1023	112.6	8
1032	84.4	6
2013	84.4	6
1200	7.4	11
2010	5.4	59
1020	4.2	46
2100	4.0	6
1002	3.3	11
2001	2.1	7
0000	1.0	22465
1000	0.5	156

Figure 4.5: Graph showing the relationship between BMI and persons with and without CHD by age, combined male and female



from 30 onwards, but at older ages the BMI tends to reflect that of the general population. The difference in BMI between the CHD population and the rest of the population is highest between ages 30 and 50.

4.3 RISK LADDERS

Risk ladders show the number of persons with any given combination of risk factors, the total exposure in a population to individual risk factors, and the risk of finding a particular diagnosis (such as the presence of diabetes) within any risk factor combination.

Each disease was evaluated using the same approach. Note that the inclusion of a risk factor does not have to cause a particular disease to be associated with it. A full set of risk ladders for each disease is included in the appendices.

Table 4.4 shows a CHD risk ladder. The following factors were selected for inclusion from a larger group as being the most important from our initial set of variables:

- Gender
- Council tax band A-C[39]
- BMI > 30
- Current smoker
- Diabetes
- Hypertension

The second column of the table gives the number of factors included for the given row combination and the third column the number of patients in the sub-group.

Subsequent column entries have either a 'Y', signifying inclusion of the given risk factor, or are blank. The final column shows the percentage of persons in the sub-group that had been diagnosed with CHD. With 6 risk factors we would expect to see a table with 64 rows but, because some risk combinations have no associated observations, 20 rows are omitted.

Various totals are given at the foot of the table. The total in the second column refers to the total number of patients in the three practices; the third column refers to the number of males in the sample; and subsequent columns show the number of times the given factor was present in the population.

Among the findings we note that:

- Female risk of having CHD with no factors present is 0.3% whereas male risk is 0.4% (population sizes 8006, and 7903)
- For a smoker with a high BMI male risk increases to 3.6% and female risk to 1.5% (population size 412, and 528)
- Living in a property banded A-C increases the risk of CHD for males and females with no other factors to 0.7% and 0.5% (population size 1208, and 1274)
- Risk of CHD increases substantially if hypertension or diabetes is present. Thus a male smoker with diabetes has a 7.3% risk which increases to 12.5 % if he has a BMI of over 30. This increases to 15.5% if diabetes is replaced by hypertension.
- Those in the highest risk categories have most risk factors but the sample sizes are small.

Table 4.4: A risk ladder for CHD

Case	Number of factors	Number of patients	Male	CT A-C	BMI>30	Current smoker	Diabetes	Hypertension	Observed risk of CHD %
1	3	14	Y				Y	Y	50.0
2	5	8	Y	Y		Y	Y	Y	50.0
3	4	5		Y		Y	Y	Y	40.0
4	5	20		Y	Y	Y	Y	Y	35.0
5	4	3	Y	Y			Y	Y	33.3
6	5	20	Y	Y	Y	Y		Y	30.0
7	4	42	Y			Y	Y	Y	26.2
8	2	12					Y	Y	25.0
9	5	44	Y		Y	Y	Y	Y	25.0
10	6	16	Y	Y	Y	Y	Y	Y	25.0
11	2	17				Y	Y		23.5
12	4	37	Y	Y		Y		Y	21.6
13	3	11		Y			Y	Y	18.2
14	4	51			Y	Y	Y	Y	15.7
15	4	97	Y		Y	Y		Y	15.5
16	3	23				Y	Y	Y	13.0
17	2	31	Y				Y		12.9
18	3	31	Y	Y				Y	12.9
19	2	78		Y				Y	12.8
20	4	32	Y		Y	Y	Y		12.5
21	2	179	Y					Y	11.7
22	3	176	Y			Y		Y	10.8
23	3	150			Y	Y		Y	9.3
24	4	11		Y	Y	Y	Y		9.1
25	3	26		Y		Y		Y	7.7
26	3	106		Y	Y	Y			7.5
27	3	41	Y			Y	Y		7.3
28	4	16	Y	Y		Y	Y		6.3
29	4	81	Y	Y	Y	Y			6.2
30	2	137				Y		Y	5.8
31	1	246						Y	5.3
32	3	255	Y	Y		Y			5.1
33	1	22					Y		4.5
34	3	412	Y		Y	Y			3.6
35	3	28			Y	Y	Y		3.6
36	4	32		Y	Y	Y		Y	3.1
37	2	196		Y		Y			2.6
38	2	1672	Y			Y			2.0
39	2	528			Y	Y			1.5
40	1	1081				Y			1.1
41	2	1208	Y	Y					0.7
42	1	1274		Y					0.5
43	1	7903	Y						0.4
44	0	8006							0.3
	Total	**24401**	**12330**	**3457**	**1636**	**5372**	**470**	**1458**	**379**

Using a logistic regression model we found that the predicted odds of having CHD in increasing order of ascendancy are:

- 1.3 times if the person has BMI>30
- 1.4 times if the person is male
- 1.7 times if the person lives in housing in council tax bands A-C
- 3.6 times if the person has diabetes
- 3.9 times if the person is a smoker
- 9.2 times if the person has hypertension

Figure 4.6 compares the predicted and observed risk using the model for each risk factor combination.

If we bundle the risk factors together regardless of type we obtain the results shown in Table 4.5. It shows how the risk escalates with the number of risk factors but the number in each risk factor group declines. The risk of CHD with no risk factors is 0.3% (group size 8006) rising to 27.5% with 5 or more factors (group size 116).

Another CHD risk ladder was created that focused more on social factors and included variables such as household

Figure 4.6: Predicted risk of CHD plotted against observed risk for each risk factor combination

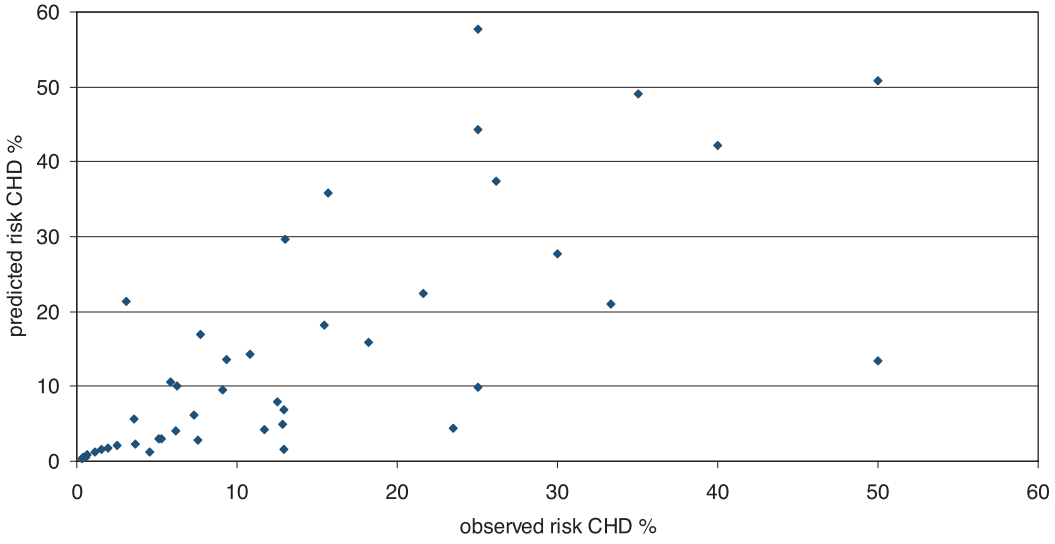
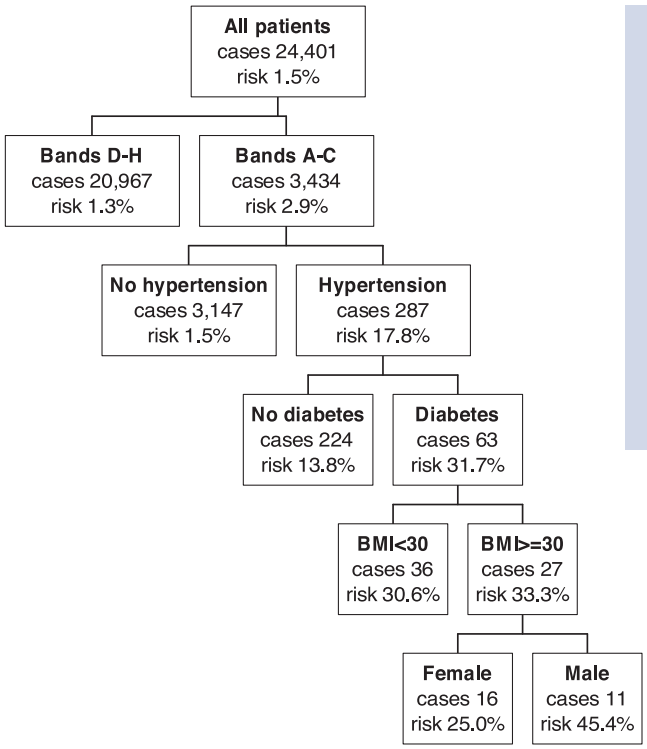


Table 4.5: Grouped risk factors for CHD

CHD factors	Patients in group	Cases	Risk
0	8006	27	0.3
1	10526	68	0.6
2	4065	104	2.6
3	1281	91	7.1
4	407	57	14.0
5	100	28	28.0
6	16	4	25.0

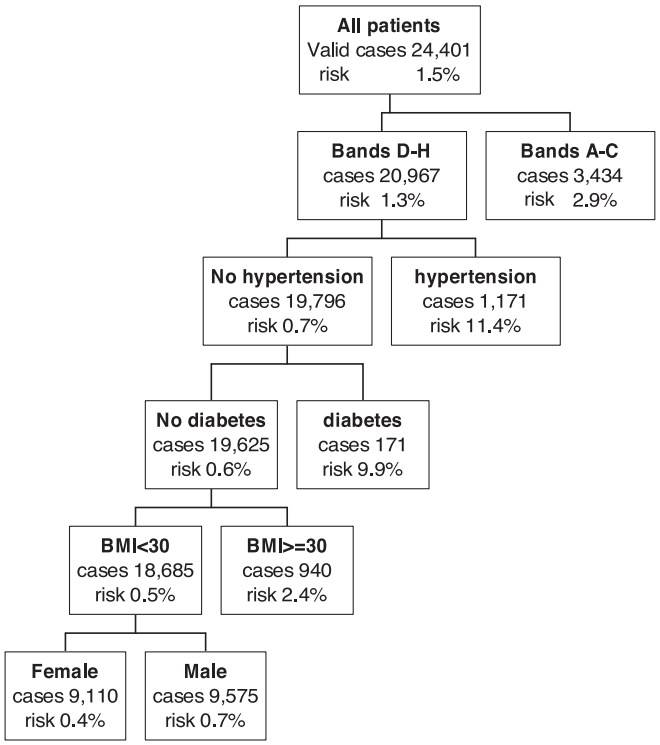
Figure 4.7: A CHD risk tree for the general practice population living in council tax bands A-C indicating the influence of risk factors such as hypertension, diabetes, BMI and gender



In this practice population of 24,401 people:

- The risk of finding CHD in houses in council tax bands A-C is 2.9% and over twice as high as in bands D-H
- A person living in bands A-C with hypertension has a 17.8% risk of CHD, increasing to 31.7% if they also have diabetes, 33.3% if their BMI is over 30, and 45.4% if they are male
- The risk of CHD is lowest in council tax bands D-H among persons who have never been diagnosed with hypertension, diabetes, who are of normal weight and female

Figure 4.8: A CHD risk tree for the general practice population living bands D-H in the absence of specific risk factors



structure including living alone, lone parent, couple with children, and more than five people per household.

The resulting fitted logistic regression model showed that the most influential factors in order of importance were:

- Being over 50
- Being a smoker

An interesting and fairly large group in the higher risk categories were those living alone, who are smokers and over 50.

4.4 RISK TREES

Risk trees show the effects of the progressive addition of risk factors on the risk of chronic disease. Risk trees differ from risk ladders in that they split the population from the top down rather than splitting it into mutually exclusive groups. Risk trees are helpful for quantifying and profiling specific sub-groups of the population for whom, for example, health interventions may be considered appropriate or who might become 'health role models'. Contrast for example the difference in risk of finding CHD between persons living in lower value housing defined as housing council tax bands A-C with persons living in higher value bands, D-H.

The first example shown in Figure 4.7 partitions the practice population living in housing bands A-C according to whether they have hypertension, diabetes, a BMI >30, and by gender. In each box is given the number of persons in the practice population with the characteristics shown, and the observed risk of finding CHD in that group. As is seen the overall risk for males and females is 1.5% but this increases to 2.9% if the patient lives in bands A-C, rises to 17.8% if hypertension is diagnosed, 31.8% if diabetes is also diagnosed, 33.3% if their BMI>30, and 45.4 % if the patient is male. By the time we reach this branch of the tree the population has shrunk from 3434 patients living in bands A-C to just 11 male patients and 16 female patients.

The second example shown in Figure 4.8 is the 'healthy' risk profile. It consists of patients living in higher banded housing (bands D-H) that do not have any of the risk factors of the first group. The overall risk of CHD in this group is 1.3%, or half the risk of those in the first group. This risk declines to 0.7% in the absence of hypertension, 0.6% in the absence of diabetes and hypertension, 0.5% if the person's weight is also normal, and 0.4% if they are female. In reaching this branch of the tree the population has shrunk rather less, from 20,967 living in bands D-H to 9,575 healthy male patients and 9,807 healthy female patients.

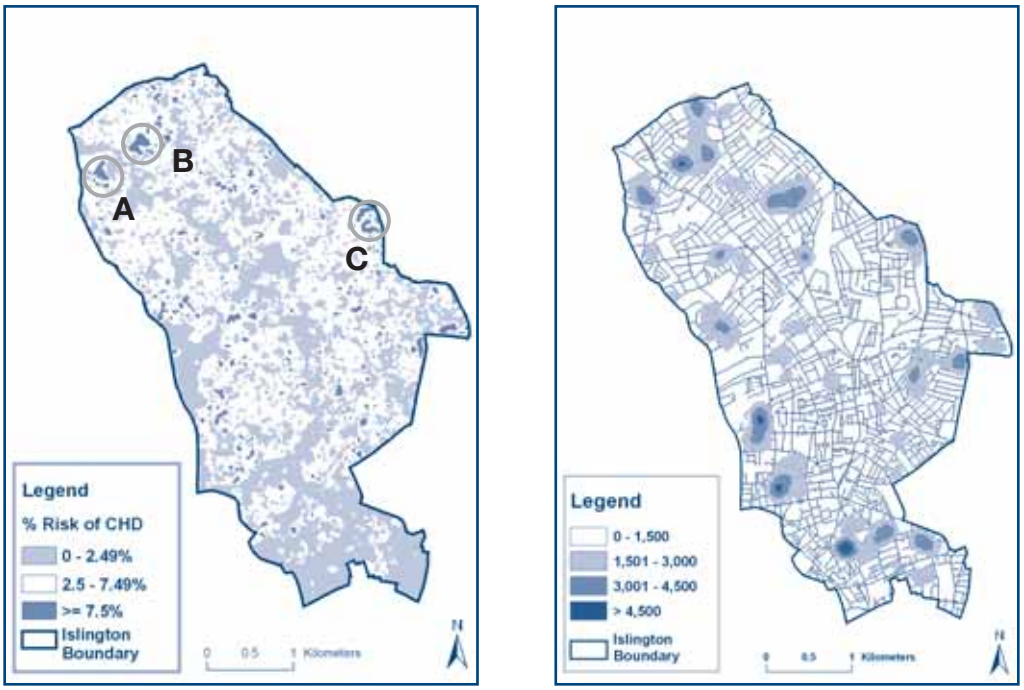
4.5 MAPS OF RISK

In our study all patients were geo-referenced (assigned an x,y co-ordinate) so that the incidence of chronic disease could be mapped. Since we only had detailed information on the prevalence of chronic disease for approximately 15% of the Islington population, we needed to model the risk to the whole population for which we had no information other than age, gender and housing tenure.

Figure 4.9 illustrates how such data can be presented in map form using a Geographical Information System (GIS). The right hand map shows concentrations of either gender aged over 50 living in housing designated in council tax bands A-C. The left hand map is a CHD risk map based on three factors: being male, aged over 50, and living in a house designated in council tax band A-C. Risk is categorised in three bands: 0%-2.49%, 2.6%-7.49%, and >7.5%.

The areas of higher risk tend to be in areas where this sub-group is concentrated as might be expected. Of the circled areas where there are larger risk clusters A and B are in Archway in north Islington and C is in Highbury in east Islington. Each risk concentration has between 200 and 270 males fitting the description of this sub-group. Because of their relative accuracy such maps are likely to be useful for primary care planning or for targeting health messages, for example using advertising hoardings or posters in stations.

Figure 4.9: (i) A map of Islington showing residential areas with males at higher or lower risk of CHD; (ii) a map of Islington showing the density of persons aged 50 and over in housing in council tax bands A-C



5 Chronic disease and health care usage

5.1 METHODOLOGY

The number of chronic disease diagnoses is likely to influence the pattern and intensity of health care usage in terms of visits to the GP, prescriptions, referrals and inpatient admissions. In this section we examine this use of health care resources to produce detailed tables of relative usage that potentially have wide applicability in a number of fields including health care planning and financing.

For this analysis, we used data from The Health Improvement Network (THIN) database, compiled by EPIC (see also section 2.1). A more detailed description of this data set is contained in Appendix 7. Data are contained in four main databases:

- A patient registration database
- A medical records database, containing events such as GP visits with Read codes for diagnoses, signs and symptoms
- A therapy records database, containing drug prescription records
- An Additional Health Data (AHD) records database, containing routine health status records, e.g. height, weight, vaccinations, blood pressure readings, BMI, blood test results etc

Our study population consisted of patients who were permanently registered with a practice as at 1st January 1990 and who were diagnosed with one of the five chronic diseases in our study, either before or after the start of the study period (1st January 1990). Our study population therefore comprised two groups of patients: (i) those who were already suffering from a chronic disease at the start of the study period and (ii) those who were subsequently diagnosed with a chronic disease during our study period (1st January 1990 to 31st December 2004). We examined the records of both sets of patients from the date of their first medical record (which could be prior to 1st January 1990) to the end of the study period (or the date when they left or transferred from the practice if earlier) to see which chronic diseases they had, using the following Read diagnostic codes:

Read Code	Condition
H3.xx	COPD
C10.xx	Diabetes
G66.xx	Stroke
G20.xx	Hypertension
G3.xx	CHD

By examining historical medical records prior to the start of the study period, we could determine which, if any, of the five chronic diseases were suffered by each person in our population from the date of diagnosis. We then collated all the medical record data for our study population and categorised these into the following four groups of services:

1. GP visits (including night and emergency home visits, as well as surgery consults)
2. Prescriptions issued (including repeat prescriptions for which no GP visit occurred)
3. Referrals to secondary care for specialist consultation
4. Inpatient hospital admissions, either elective or emergency

The four groups given do not cover all possible primary care services, but we chose those for which we had the most robust and significant data.

By linking the medical records with the AHD records for each patient, we were able to build a picture of each patient's health status and risk factors over the study period (see Box 5.1).

In order to determine the risk factors and periods of exposure for each patient during the study period, each patient registration record was split into time periods, with start dates and end dates corresponding to certain key events:

1. Being diagnosed with one of the five examined chronic illnesses
2. A recorded change in the risk factor/health status characteristics, i.e. BMI, Blood Pressure, Smoker Status
3. A change in the calendar year
4. A change in the age of the patient

Days of exposure were calculated from the start of the study period to the earlier of the patient's leaving date, date of death, or the end of the study period. See box 5.2 for a timeline example.

Once the exposure was estimated for each patient, the number of medical services used in each separate time period was estimated, using the services identified in the Medical and Therapy records.

We then used a Generalised Linear Modelling technique to build a statistical model for each type of medical service utilisation. This technique is described in Appendix 9. We

Box 5.1

We examined each patient's smoking status, Body Mass Index records, Blood Pressure records, Height and Weight, using the AHD records for our study population. For each record, the value and the date were captured. To assess values of each variable at the start of the study period, the most recent recording prior to January 1, 1990 was taken. Generally, our methodology was to assume the patient continued in the state indicated by the most recent historical record until a more recent record was collected, at which point the patient was assumed to have transferred to the state indicated by the most recent record.

We categorised each record as follows:

Smoking Status

1. Never smoked (i.e. the patient has records which indicate the smoking status has been determined and is coded as "Non Smoker" consistently)
2. Current Smoker (the patient has records which indicate a smoker, or records with a non-zero number of cigarettes entered)
3. Ex-Smoker (the patient has had previous smoking records, but subsequently had two consecutive records indicating "Non-Smoker").

Patients in smoking Status 1 must have remained in this state for the whole of the study period and prior, but patients were allowed to move between States 2 and 3 (and vice versa) as many times as our categorisation methodology allowed.

Body Mass Index

Category	Description	Reading
1	Underweight	< 20
2	Normal	20 to 24
3	Overweight	25 to 29
4	Obese	30 to 34
5	Morbidly Obese 35-39	35 to 39
6	Morbidly Obese 40+	40 +

Blood Pressure

Category	Description	Systolic	Diastolic
1	Low	< 120	<80
2	Normal	120-129	80-84
3	High Normal	130-139	85-89
4	Hypertensive I	140-159	90-99
5	Hypertensive II	160-179	100-10
6	Hypertensive III	180-209	110-119
7	Hypertensive IV	210+	120+

fitted multiplicative Poisson distribution models with a Log Link function, using the following dependent variables:

- Current Age Group (in 5 year age bands between ages 50 and 80)
- Gender
- Calendar Year
- BMI Category
- Smoker Status
- Diabetic (Yes/No)
- COPD (Yes/No)
- Stroke (Yes/No)
- CHD (Yes/No)
- Hypertension (Yes/No)

We initially also used Blood Pressure Category as a dependent variable, but as this variable was highly correlated with Hypertension diagnosis, we excluded this variable from the final models.

We tested the significance of each dependent variable using Chi-Squared tests. All of the dependent variables were found to be highly significant (Probability Chi-Squared < 0.001) in the models.

The model output gives a set of multiplicative factors for each level of each of the dependent variables and a base output for the variable of interest, e.g. GP Visits. A full set of outputs for each model is set out in Appendix 8.

5.1.1 Methodological issues

There were a number of issues that we identified with the data, which may give rise to biases in the results. The most significant is the inability to track patients between different GP practices. Each patient is given a unique identification number at each practice, but they do not transfer this unique identifier across different practices. While patients joining a new practice have some previous medical history transferred (i.e. if they have a chronic disease, the Read code for this chronic disease is likely to be recorded when they join the new practice), this does not give us the exact date of diagnosis. Therefore, our medical records may be incomplete. In order to reduce the significance of incomplete records as far as possible, we chose a sample of patients who had been registered with the same practice since 1990. We then tracked these patients going forward in our study until they: a) transferred out of the practice, b) died, or c) reached the end of our study period. In selecting this patient subset, we implicitly assumed that the patients in our subset would have had similar health care usage profiles to the population as a whole.

When reading in the medical records, some of the "older" records had only month and year of treatment or, in some cases, only the year of treatment. For records where the day was not available the consultation was assumed to be on the 15th of the month. For records where the month was not available it was assumed to be June. We had a small number of medical records with no date coded (approximately 0.5%).

When determining the number of hospital admissions, it was identified that there were significantly more hospital discharges coded than admissions and hence the number of discharges has been used as a proxy for the number of admissions.

The results show for example that:

- A male non-smoker with a normal BMI visits his GP three times a year and is prescribed 18 sets of prescription drugs
- A male with a BMI>30 who has CHD and diabetes visits his GP 14 times a year and is prescribed 28 prescription drugs
- An underweight male (BMI<20) visits the GP as often as an obese male and is just as likely to be admitted to hospital
- A male non-smoker aged 75-79 has a 14% chance of being admitted to hospital compared with a 6% chance for a 50-year old male non-smoker

In general:

- Females visit the GP more often at every age but are less likely than males to be admitted to hospital at older ages
- Current and ex-smokers visit the GP more often, and are more likely to be referred to a consultant or be admitted to hospital

5.2 RESULTS

Table 5.1 below gives the annual number of GP visits[40] in 2003 by age and sex for a population with the following characteristics:

- never smoked
- normal BMI (defined as between 20 and 25)
- have no historical or current diagnoses of one of our five chronic diseases[41]

Multiplicative factors to adjust these visits for chronically diseased populations and other population characteristics, such as different levels of BMI and, where applicable, smoker status, are also given in Tables 5.2 and 5.3. For example, to estimate the average number of GP visits for a Male, aged 64 with a BMI of 29 and a current diagnosis of diabetes and COPD, the base table value of 3.3 visits per person per year should be multiplied by 1.01 and 2.55 to give an average number of GP visits of 8.5 per person per year.

Appendix 8 contains similar tables for other medical services, including the numbers of prescriptions issued, referrals for specialist consultations and inpatient admission. Appendix 8 also contains some suggested calendar year trends in medical services to apply to the 2003 base utilisation.

Table 5.1: Base table showing the number of GP visits per person per year in 2003, for non-smokers, with normal BMI, and none of the following diagnoses: COPD, diabetes, CHD, hypertension, or stroke

Age	Males	Females
0-49	2.6	4.1
50-54	2.9	4.1
55-59	3.1	4.0
60-64	3.3	3.9
65-69	3.4	3.9
70-74	3.6	4.0
75-79	3.9	4.3
80+	4.1	4.3

Table 5.2: Multiplicative factors based on BMI status

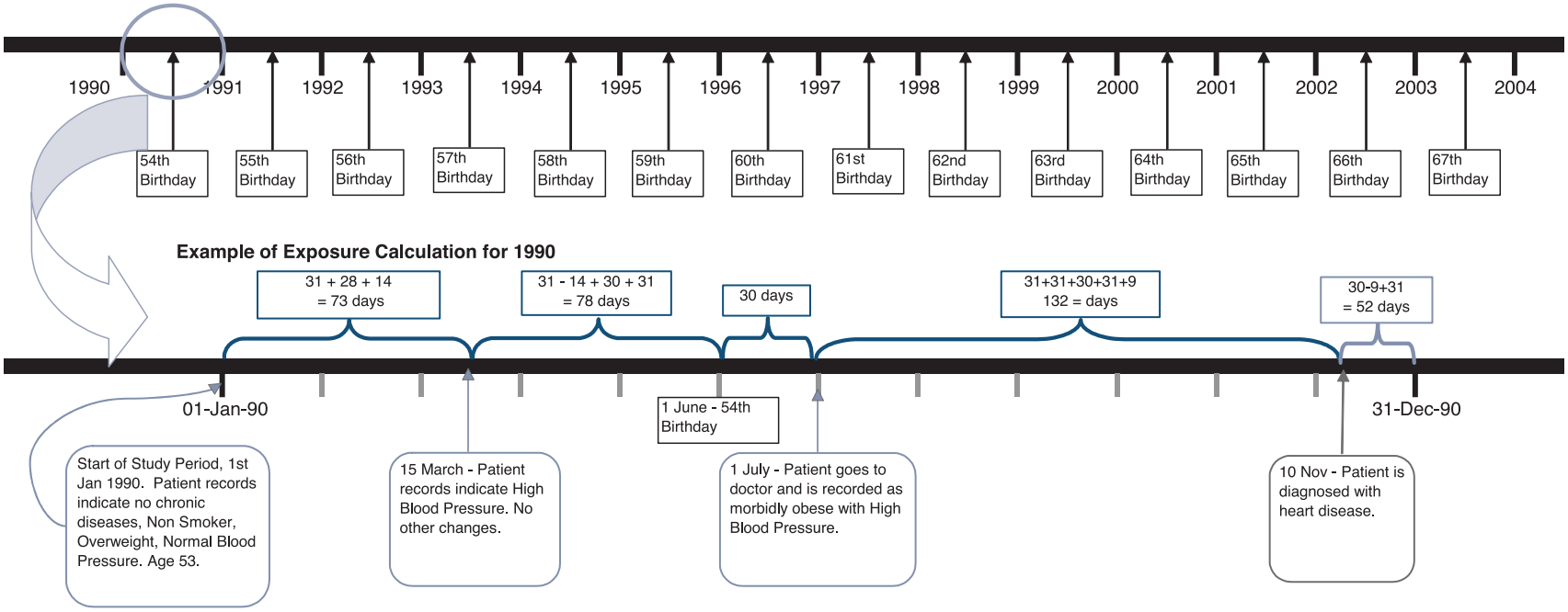
BMI status	Relativity
Underweight (<20)	1.06
Normal (20-25)	1.00
Overweight (25-30)	1.01
Obese (30-35)	1.05
Morbidly Obese (35-40)	1.07
40+	1.10

Table 5.3: Multiplicative factors according to type and number of diagnoses

Sequence	COPD	Stroke	Hypertension	CHD	Diabetes	Relativity
1	Y	Y	Y	Y	Y	3.34
2	Y	Y	Y	Y	N	2.74
3	Y	Y	Y	N	Y	2.94
4	Y	Y	Y	N	N	2.24
5	Y	Y	N	Y	Y	3.46
6	Y	Y	N	Y	N	2.88
7	Y	Y	N	N	Y	2.50
8	Y	Y	N	N	N	2.25
9	Y	N	Y	Y	Y	3.00
10	Y	N	Y	Y	N	2.39
11	Y	N	Y	N	Y	2.63
12	Y	N	Y	N	N	2.03
13	Y	N	N	Y	Y	2.95
14	Y	N	N	Y	N	2.37
15	Y	N	N	N	Y	2.55
16	Y	N	N	N	N	1.91
17	N	Y	Y	Y	Y	2.32
18	N	Y	Y	Y	N	1.83
19	N	Y	Y	N	Y	2.11
20	N	Y	Y	N	N	1.61
21	N	Y	N	Y	Y	2.35
22	N	Y	N	Y	N	1.78
23	N	Y	N	N	Y	1.97
24	N	Y	N	N	N	1.49
25	N	N	Y	Y	Y	2.21
26	N	N	Y	Y	N	1.66
27	N	N	Y	N	Y	1.88
28	N	N	Y	N	N	1.31
29	N	N	N	Y	Y	2.05
30	N	N	N	Y	N	1.51
31	N	N	N	N	Y	1.66
32	N	N	N	N	N	1.00

Box 5.2: Possible Timeline for Exposure Calculation

Timeline for Exposure Calculation



Exposure For 1 Jan 90 to 31 Dec 90		Days
Aged 53	Overweight, Normal Blood Pressure	73
	Overweight, High Blood Pressure	78

		Days
Aged 54	Overweight, High Blood Pressure	30
	Morbidly Obese, High Blood Pressure	132
	Morbidly Obese, High Blood Pressure, Heart disease	52

6 Survival analysis

6.1 INTRODUCTION

The Islington PCT analysis (in section 4) was based entirely on persons that were alive when the data snapshot was taken. Although our patient sample comprising 24,000 records was sufficient to carry out the analyses described, we were aware that much larger samples would be needed to extend the age dimension of our analysis. An age dimension is important because it could potentially inform wider strategies with regard to chronic disease prevention and management, and the financial risks relating to diseased populations for PCTs and insurance companies.

Key questions that come to come mind are:

- What are the survival chances for a person who is diagnosed with a chronic disease at age *x* and how do these compare with the survival chances of the rest of the population?
- How do these survival chances compare with people that have a different chronic disease, and with people that already have one or more chronic diseases who then contract another chronic disease?
- What level of disability should we expect depending on the cumulative number of chronic disease diagnoses at age *x*?

We designate the time from birth of an individual until he or she is diagnosed with any particular chronic disease, or until he or she dies, the 'failure time' due to the corresponding decrement. We denote the random variables representing the failure times from diseases $1,2,\ldots,m-1$ as $T_1,\ldots,T_{m-1}$, and let T_m be the time to death. Further assume that at birth, each individual can be assigned a random vector of failure times $T_1,\ldots,T_m$ corresponding to a subset of diseases that the individual is diagnosed with during the life course, while others remain latent (i.e. unobservable) if death occurs earlier (see Figure 6.1). Potentially we are interested in the joint survival function:

$$S(t_1,\ldots,t_m) = P(T_1 > t_1,\ldots,T_m > t_m)$$

However, to undertake a complete specification of and analysis of all such survival functions would require a considerable amount of research and resources over several years.

We have, however, been able to make a modest start by analysing THIN data for individual, rather than joint disease survival probabilities, based on the simplest case. In what follows we deal with the first of the three questions above – namely the chances of survival contingent on the diagnosis of a chronic disease, disregarding any other condition that a patient may already have or subsequently acquire. Although we have extracted data for five different conditions, we focus our analysis on males and limit it to diagnoses of COPD, hypertension, and CHD.

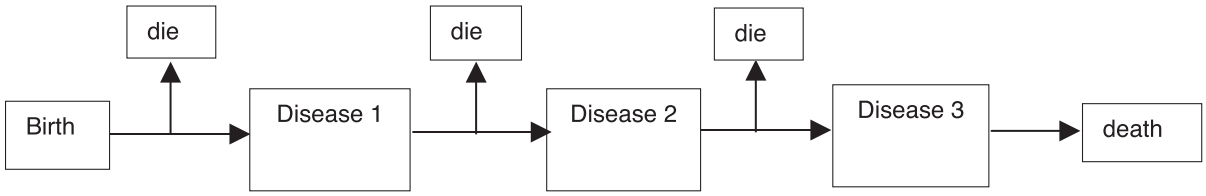
6.2 METHODOLOGY AND RESULTS

The THIN extract comprised data taken from 255 GP surgeries at January 2005. These 255 surgeries collectively represent approximately 3% of the UK population. For our investigation we considered all patients that:

- were registered with a practice at 1st January 1990,
- appeared to be permanently assigned to the practice, and
- had an "acceptable" record i.e. the record did not contain any obviously invalid data.

The medical records for these patients were then tested to see if they had had any of the three conditions: COPD, Hypertension or CHD, at any point in their recorded medical history.

Figure 6.1: Hypothetical progression of a sequence of chronic diseases



The Read codes relating to the three conditions are those shown in the accompanying table:

Read Code	Condition
H3.xx	COPD
G20.xx	Hypertension
G3.xx	CHD

Note that for each Read code shown above there can be several medical codes that relate to the relevant condition. For example, the complete list of medical codes for the condition COPD is: "H3…11" and "H3…00"

Our population was defined as people that had been diagnosed with one of the five conditions between 1st January 1990 and 31st December 1993. To determine if a male patient had one of the conditions prior to 1st January 1990, all previous medical records for each patient were checked. For example, considering CHD, patients would only be included *if they had not previously suffered from any of the conditions including CHD*.

The age at diagnosis was then derived as being the age at the date the condition was reported (the 'event date'). We also assume that the diagnosis was accurate though this may not always have been the case. As exact date of birth is unknown for many of the records, age was defined as reporting year minus year of birth. The number of days survived by each patient was then calculated. All records were then summarised by creating a table showing for each age of diagnosis the number of people that survived for each number of days.

The final task was to identify how many days had elapsed before a certain proportion of the population had died (i.e. 5%, 10%, 20% etc.). The number of years reaching each of these percentage points was then determined as number of days divided by 365. Due to a lack of data at the youngest and oldest ages, we focused our analysis on the age range 55 to 85. The residual matrix was then graphed for different percentiles of the population and curves fitted.

In order, to benchmark our analysis we compared the survival rates of males with different chronic diseases with survival rates extracted from ELT15M, the male life table for England, which comprises deaths from all causes - not just the chronic diseases we examined. Figure 6.2 shows the survival rates based on different percentiles for males of different ages, with 'age' on the vertical axis and 'years survived' on the horizontal axis. A horizontal cross section such as A-B gives the survival curve for a male at any age, in this case 70 years old. A vertical cross section such as P-Q gives the percentage of males at different ages surviving the given number of years, in this case 5 years.

The functional form of the fitted curves is $y = \alpha + \beta \ln x$ where y is a person's age and x is the number years survived. This gave very good fits to ELT15 M survival rates and reasonable fits to the three diseases for percentiles and ages for which we had the most observations. However, in order to avoid nonsensical results we are aware that a more fundamental analysis of suitable functional forms is needed to capture survival behaviour outside the ranges for which we had data, for other diseases and for females as well as males. Our results and predictions must therefore be considered as preliminary in this respect.

Table 6.1 shows a summary of the regression coefficients and R2 values for each disease and for all diseases (ELT15 M) on which graphs like Figures 6.2 were based. Figure 6.3 compares the typical survival characteristics of persons diagnosed with COPD on an identical basis using THIN data.

Consider for example a male who is diagnosed with COPD at age 70. Point P in Figure 6.3 shows that there is a 20% chance that he will die within 3.3 years (80% chance he will live) and point Q shows that there is a 50% chance he will die before 8.1 years (50% chance he will live). Conversely Figure 6.2 shows that any male in the population of the same age with any diagnosis would have a 20% of dying in 4.4 years and a 50% chance of dying in 9.9 years, or 1.1 years and 1.8 years longer than the male diagnosed with COPD.

Based on the data:

- A male diagnosed at age 60 with COPD has a 70% chance of surviving 9.6 years and a 50% chance of surviving 15.5 years
- A male diagnosed COPD at age 70 with has a 80% chance of surviving 3.3 years and a 50% chance of surviving 8.1 years (Points P and Q in Figure 6.3)

Figure 6.2: Survival rates for English males aged 55 to 85 based on ELT15 M

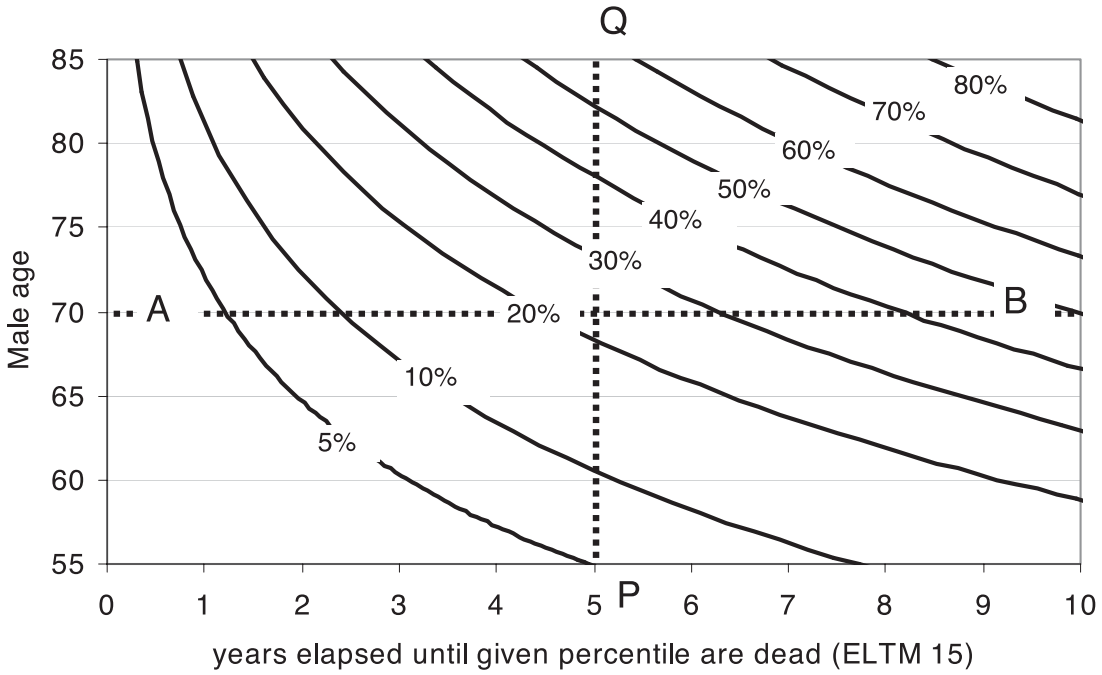
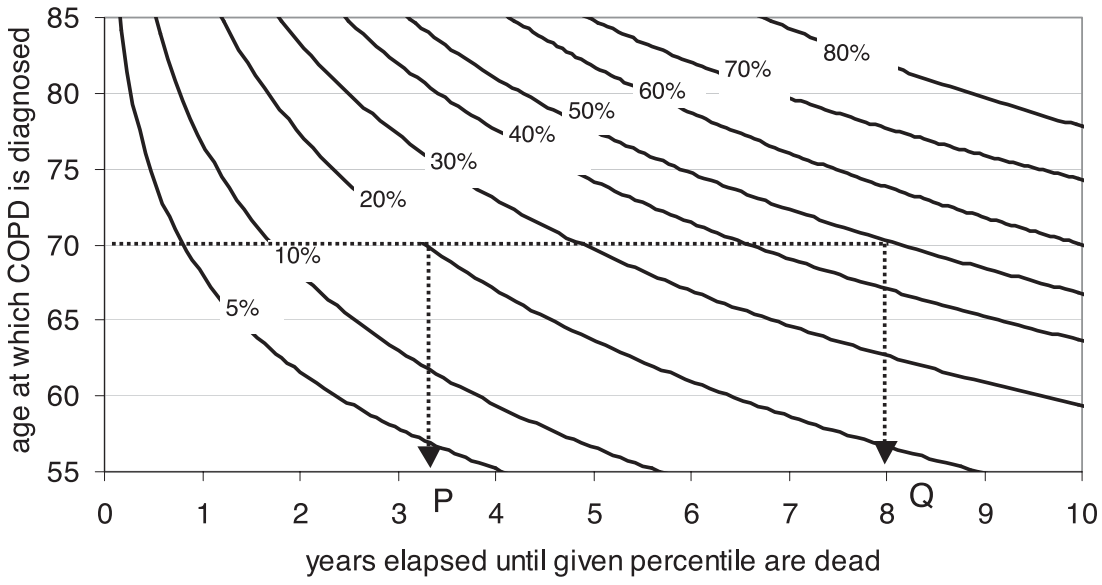


Figure 6.3: Survival curves for males diagnosed with COPD



An important question is whether it makes any difference to life expectancy at birth if one is diagnosed with a chronic disease at any time during life. Take again the example of persons diagnosed with COPD. In Figure 6.4, we plot life expectancy at birth as a function of the age at which COPD is diagnosed. The objective is to compare the difference between the life expectancy of males diagnosed at some age with COPD with the life expectancy of all male lives at the same age.

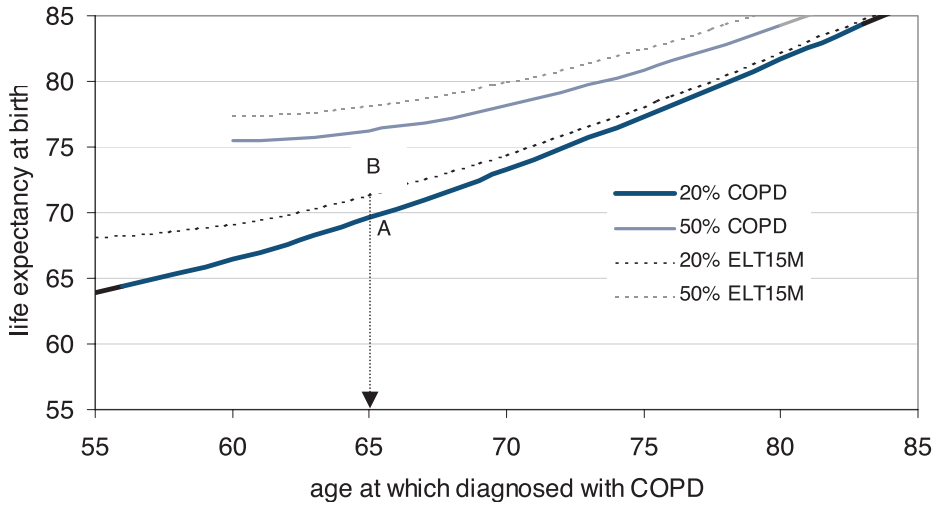
Each line in Figure 6.4 shows the life expectancy at the given ages for:

- 20% of the COPD population
- 50% of the COPD population
- 20 % of all male lives
- 50% of all male lives

Table 6.1: Regression coefficients for survival rates used in three diseases and for ELT15M

Percentile	COPD			Hypertension			CHD			ELT15M		
	α	β	R^2	α	β	R^2	α	β	R^2	α	β	R^2
5%	67.99	-9.23	0.76	78.01	-11.95	0.86	62.1	-4.35	0.74	72.1	-10.72	0.98
10%	76.74	-12.5	0.83	88.03	-13.86	0.94	74.88	-9.73	0.83	81.13	-12.79	0.98
20%	87.64	-14.89	0.92	92.82	-12.18	0.92	83.34	-9.45	0.89	90.51	-13.77	0.98
30%	93.6	-14.87	0.92	95.59	-11.13	0.92	90.88	-10.58	0.91	97.73	-15.09	0.99
40%	98.7	-15.22	0.94	97.86	-10.76	0.88	94.65	-10.7	0.89	104.62	-16.51	0.99
50%	102.59	-15.55	0.9	98.33	-9.68	0.86	99.68	-11.7	0.92	110.92	-17.82	0.99
60%	109.16	-17	0.89	95.7	-7.19	0.68	101.36	-10.95	0.83	117.67	-19.27	0.99

Figure 6.4: Life expectancy as a function of the age at which COPD is diagnosed compared with all male lives



As is seen the graph predicts that all male lives consistently have a higher life expectancy at birth than COPD lives, irrespective of the age at which COPD is diagnosed. Consider for example point A. This tells us that 20% of males diagnosed with COPD at age 65 can expect to live until they are 69.7 years, whereas the corresponding life expectancy for 20% of all male lives at age 65 based on ELT15M is 71.4 years (point B), a difference of 1.8 years. Note that the survival gap narrows with age, so that at older ages being diagnosed with COPD makes less and less difference to life expectancy compared with the 'all lives' group.

It remains to extend this analysis fully to the other diagnoses. For hypertensive lives, our preliminary view is that those diagnosed with hypertension tend to live longer than 'all lives' depending on the age of diagnosis. We consider that there are two potential explanations for this: firstly, the 'all lives' category includes a range of other diseases including cancer, and secondly, treatment for hypertension is relatively effective. As for CHD, our preliminary analysis suggests that life expectancy tends to increase if the disease is diagnosed early compared with the life expectancy of those diagnosed at a later age (approximate range 55-65). This may be related to the effectiveness of early treatment, but also to the more serious effects of heart attacks at older ages. These conclusions need to be validated and then extended in further research.

7 Bringing it all together

7.1 SUMMARY OF RESULTS

Derek Wanless's report in 2004 suggested that further research is needed to help understand effective chronic disease management. The cost of chronic disease both financially and personally is huge and we believe that this paper demonstrates how we can use the data being captured by PCT's and THIN to better understand chronic disease prevalence and progression, and their contributing risk factors.

In this paper we have demonstrated how data currently being collected can be used to:

- Understand the effect of socio-economic status on chronic disease prevalence and progression
- Investigate the variation in chronic disease prevalence within a local area, to help target resources more effectively
- Construct risk ladders to help understand the co-prevalence of diseases, progress by age, most likely sequence of disease, variation with BMI, and effect of socio-economic indicators.
- Calculate the expected survival period for someone diagnosed with a chronic disease
- Understand how the utilisation of medical services such as GP visits, prescriptions, and specialist consultations varies by age, smoker status, BMI, and chronic disease presence. This can also be developed to monitor the trends in such usage.

We now plan to continue our work, in particular to help understand how we can best manage the cost of chronic disease, and help allocate resources and target interventions most effectively. In the meantime we would welcome feedback on the techniques and findings in this report, and the most appropriate way of using this research.

Key applications could include:

- Profiling health care needs of local communities taking into account chronic disease prevalence and demography
- Helping the NHS to evaluate and target health interventions that promote health and delay the onset of chronic disease
- Helping employers and occupational health professionals with sickness management
- Providing the insurance industry with tools that could enable fairer pricing of insurance products

7.2 POSSIBLE USES

7.2.1 National Health Service

We believe there are a range of areas where this work could be of benefit in the NHS. By using the information on the use of health services according to the number and type of chronic diseases, our research could help in:

- The use of marketing techniques for targeting health messages to specific sub-groups by age and risk exposure
- Providing estimates of the future number of referrals to specialists by GP practice
- Estimating the future number of prescriptions by GP practice and total cost
- Modelling the impact on future resources of targeting the treatment or prevention of particular chronic illnesses
- Monitoring and projecting annual trends

At a local level it could be used by GP practices for estimating activity levels by practice. From work such as the analyses of GP visits in Section 5.2, applications could include:

- identifying healthcare services such as new GP practices by local profiling of needs
- estimating the number of future GP visits based on the chronic illness levels in registered patients
- using the notion of risk for segmenting and mapping health needs to ensure appropriate targeting
- improving the knowledge of co-morbidity to ensure preventative measures such as health checks are carried out on the appropriate risk groups
- providing an authoritative guide for GPs on relative risks. What advice should we give to a 35-year old woman who has just been diagnosed with diabetes about the risks of acquiring other chronic diseases? Are people more responsive to lifestyle changes if we can demonstrate their increased chance of coronary heart disease in the presence of hypertension, or a high BMI, or continuation of smoking?
- informing the costs and benefits of different interventions. Is it financially worthwhile to target measures and try to intervene before people visit a GP? Should we target populations at high risk with invitations for health screening or a health visitor check?

7.2.2 Other public and private sector uses

The cost of chronic disease to the country is not just through the health service, but through sickness absence and the long term inability to work. An extension of this work would be to improve estimates of the numbers and cost to individuals, employers, and the economy, of days of sickness resulting from one or more chronic diseases. These estimates could be used as a component in the calculation of the value and effectiveness of different types of health interventions.

7.2.3 Employers

The burden of sickness does not just fall on the Government, and the lost productivity through these illnesses is potentially as great as the medical cost of treatment. We believe that there are applications for this work in helping employers with absence management, for example in evaluating the cost-effectiveness of health-influencing employer intervention e.g. regular medicals for employees, support for giving up smoking, subsidised gym membership, etc.

7.2.4 Insurance

There is a risk with this type of research that it can be seen as simply another way for insurance companies to increase premiums.

We do not feel that this will be the case for the following reason. Much of this research focuses on improving our understanding of the risk of further chronic illnesses in the presence of one or more illnesses. Currently these people will be asked to pay an additional premium for life insurance or will even be declined cover. These decisions are taken by the insurance companies on the basis of available medical information. Where this is limited, the company is forced to take a conservative view in order to ensure that it prices the business profitably. The more knowledge we have of these risks, the more accurately they can be priced. This could lead to people previously unable to obtain life insurance, now being covered.

The ABI give the following guide[42] to interpreting the Disability Discrimination Act 1995 (DDA):

"All your decisions must be based on relevant information or data available at the time which will form the basis of your underwriting manual. This includes:

- actuarial or statistical data
- medical research information
- medical reports about an individual

You should review your underwriting manual periodically to ensure that it is based on reliable, up-to-date information that it is reasonable for you to rely on."

We hope this research will further contribute to providing insurers with the sort of information needed to support this statement.

Table 7.1 below outlines some of our thoughts on the potential uses of this research within the insurance industry.

Table 7.1: Possible Applications to Insurance Products

Insurance Product	Possible Applications of Research
Term or Whole of Life Insurance	• More accurately price business for those with one or more chronic illnesses, in concordance with the DDA. • Council tax banding: better understand the variation in risk based on property value e.g. refine the underwriting of mortgage business. • More accurately price for variation in sums insured: life insurers already make an allowance for variation in risk depending on the sum insured. Premium discounts are awarded for higher sums insured reflecting the typically lower mortality for those in the higher socio-economic groups. Again, there is limited information to support these discounts, and this work may give the opportunity to refine this knowledge.
Critical Illness	• As for life insurance (see above). • Price buyback cover: this is a standard option on critical illness policies, giving the opportunity to purchase further insurance following the occurrence of one critical illness. It is a very difficult option to price and so current costs are likely to be conservative. The lack of data could also lead to the option being withdrawn in future. This data may help to price buyback cover more accurately.
Income Protection	• As for life insurance (see above).
Annuities	• Impaired annuity pricing / pricing by socio-economic group: With the decline of final salary pension schemes, maximising the annuity available at retirement is of key importance. There is a limited market providing enhanced annuities to those who have suffered a chronic illness. However little has been done to offer enhanced annuities to those in lower socio-economic groups. With this type of research we hope both of these markets could be developed.
Private Medical Insurance	• High cost has limited the availability of medical insurance to certain unhealthy lives. Extra information should help to more accurately price for these lives and hence improve availability. • Additional data to help design and price alternative primary care insurance products, to expand the market availability for medical insurance. • Medical insurance providers are increasingly becoming involved in preventative intervention (e.g. medicals) and encouraging health beneficial lifestyle changes (e.g. giving up smoking, health club membership). This research may help to direct their preventative measures more effectively.
Long Term Care	• Use of survival data to price more accurately. • Sales are low at the moment, partly due to cost, and more accurate pricing should help moderate cost in the future.

7.3 FURTHER RESEARCH

There are several areas we see for focusing our future research. We would welcome suggestions from readers of the most productive areas to target.

Before we develop our work into new areas we would like to extend the analysis to the other main chronic diseases (including cancer, asthma and mental illness) looking at both morbidity and co-morbidity, using the techniques detailed in this paper.

We would then be keen to repeat and extend this work with another PCT, possibly from a different type of geographical area, to help verify these findings, and identify new factors. One clear direction is to expand the risk factor analysis to the whole UK population, but some detail may be lost in the process due to data limitations.

Whilst data demands would be heavy if extended to the whole population in the way described for Islington, risk ladders that capture some of the wider determinants of health could be produced from modifications to existing surveys such as the Health Survey for England.

We also believe there is considerable scope to expand on the work in section 4 to other diseases and to disease interactions in order to improve our understanding of 'pathways' and disease progression, so that we can produce better predictive models.

If we go back to our key objectives stated in Section 1.2, the main area we have yet to consider is that of cost. We are looking to further the work done on morbidity and service utilisation to bring in a cost element.

Bringing cost into the models raises interesting questions on intervention. How do we determine the appropriate targeting of medical resources for intervention? Early intervention on hypertension, for example, should lead to a reduction in CHD and the cost of coronary care. However treatment for hypertension will also have a cost, and the person is likely at some point to need treatment for another chronic condition, one that may be longer term in nature and in treatment terms more expensive (e.g. cancer, Alzheimer's Disease). What are the most cost-effective evidence-based strategies for preventing chronic disease?

Finally we would like to expand on the possible applications in different fields of healthcare and insurance, to see how these can be practically implemented.

Appendices

APPENDIX 1: DIABETES RISK LADDER

Case	Number of factors	Number of patients	Male	CT A-C	BMI>30	Current smoker	Over 50	Hypertension	Observed risk of diabetes %
1	6	30	Y	Y	Y	Y	Y	Y	46.7
2	5	42		Y	Y	Y	Y	Y	42.9
3	5	106	Y		Y	Y	Y	Y	36.8
4	5	6	Y	Y	Y	Y		Y	33.3
5	3	41			Y	Y		Y	26.8
6	4	160			Y	Y	Y	Y	25.0
7	5	35	Y	Y		Y	Y	Y	22.9
8	4	181	Y			Y	Y	Y	20.4
9	4	10		Y	Y	Y		Y	20.0
10	5	20	Y	Y	Y	Y	Y		20.0
11	4	118	Y		Y	Y	Y		19.5
12	4	26		Y		Y	Y	Y	19.2
13	4	64	Y	Y		Y	Y		17.2
14	4	37		Y	Y	Y	Y		16.2
15	3	132				Y	Y	Y	15.9
16	4	35	Y		Y	Y		Y	14.3
17	3	74		Y			Y	Y	13.5
18	3	37	Y			Y		Y	13.5
19	3	158			Y	Y	Y		12.7
20	4	26	Y	Y			Y	Y	11.5
21	3	148	Y				Y	Y	8.8
22	2	28				Y		Y	7.1
23	2	15		Y				Y	6.7
24	3	80		Y	Y	Y			6.3
25	2	195					Y	Y	6.2
26	4	69	Y	Y	Y	Y			5.8
27	3	357	Y			Y	Y		5.6
28	2	275				Y	Y		4.0
29	2	157		Y			Y		3.2
30	3	326	Y		Y	Y			2.8
31	3	207	Y	Y		Y			2.4
32	2	166		Y		Y			2.4
33	2	45	Y					Y	2.2
34	2	398			Y	Y			2.0
35	2	1356	Y			Y			1.5
36	2	1229	Y				Y		1.5
37	3	163	Y	Y			Y		1.2
38	1	1078					Y		1.0
39	1	823				Y			0.7
40	1	6705	Y						0.2
41	2	1049	Y	Y					0.2
42	1	1124		Y					0.2
43	0	6950							0.2
	Total	**24401**	**12330**	**3457**	**1636**	**5372**	**4845**	**1450**	**470**

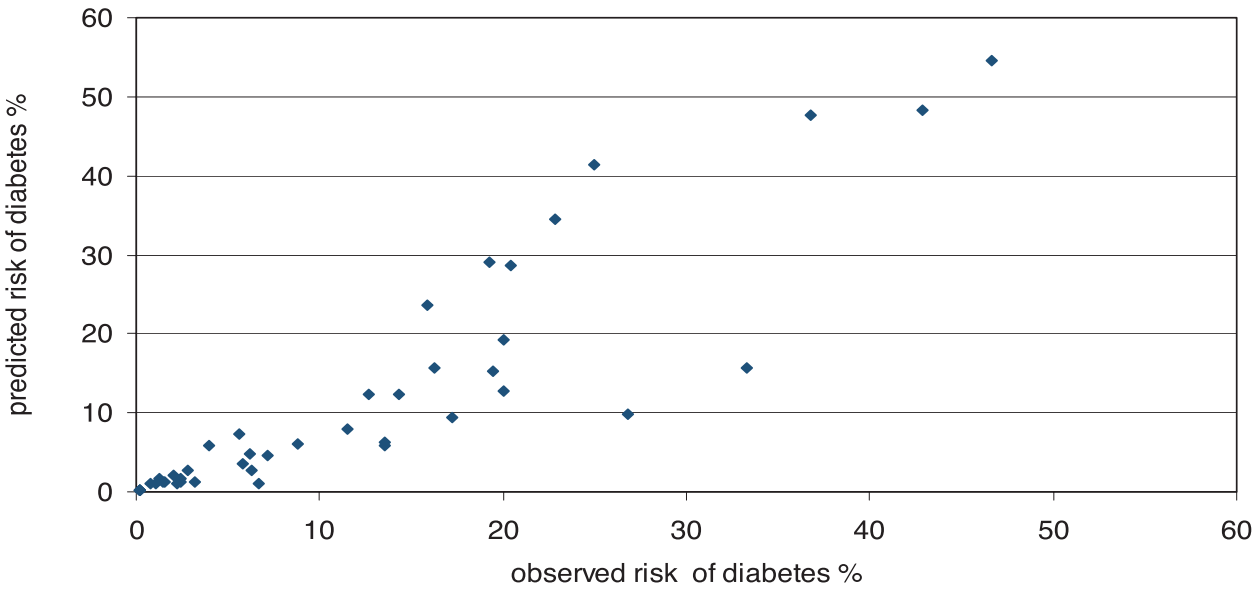
The diabetes risk ladder shows for example that:

- If no factors are present the chances of having diabetes are 0.2%
- For a male smoker with a BMI>30 the risk increases to 2.8%
- If male and over 50 with a BMI>30, a current smoker living in a house in council tax band A-C the risk of diabetes increases to 20%
- If all factors are indicated, including hypertension, the risk increase to 46.7%

Based on the logistic regression model:

- Being male increases the odds of diabetes 1.3 times
- Council tax band A-C 1.3 times
- BMI> 30 2.3 times
- Hypertension 5 times
- Over 50 years old 6.2 times
- Current smoker 6.4 times

Figure A1: Predicted risk of diabetes versus observed risk based on logistic regression model



APPENDIX 2: HYPERTENSION RISK LADDER

Case	Number of factors	Number of patients	Male	CT A-C	BMI>30	Current smoker	CHD	Diabetes	Observed risk of hypertension %
1	3	2		Y			Y	Y	100.0
2	4	2		Y		Y	Y	Y	100.0
3	4	1	Y	Y			Y	Y	100.0
4	6	4	Y	Y	Y	Y	Y	Y	100.0
5	4	9			Y	Y	Y	Y	88.9
6	5	8		Y	Y	Y	Y	Y	87.5
7	5	5	Y	Y		Y	Y	Y	80.0
8	4	14	Y			Y	Y	Y	78.6
9	2	4					Y	Y	75.0
10	5	15	Y		Y	Y	Y	Y	73.3
11	3	22			Y	Y	Y		63.6
12	3	11	Y				Y	Y	63.6
13	3	70			Y	Y		Y	61.4
14	2	33				Y		Y	60.6
15	5	20	Y	Y	Y	Y		Y	60.0
16	2	17		Y			Y		58.8
17	4	23		Y	Y	Y		Y	56.5
18	2	16		Y				Y	56.3
19	5	11	Y	Y	Y	Y	Y		54.5
20	4	61	Y		Y	Y		Y	54.1
21	4	30	Y		Y	Y	Y		50.0
22	3	69	Y			Y		Y	44.9
23	3	7				Y	Y	Y	42.9
24	3	7		Y		Y		Y	42.9
25	2	20				Y	Y		40.0
26	4	21	Y	Y		Y	Y		38.1
27	2	56	Y				Y		37.5
28	3	52	Y			Y	Y		36.5
29	3	6	Y	Y				Y	33.3
30	3	12	Y	Y			Y		33.3
31	1	40					Y		32.5
32	1	30						Y	30.0
33	3	7		Y		Y	Y		28.6
34	3	129		Y	Y	Y			24.0
35	4	19	Y	Y		Y		Y	21.1
36	2	656			Y	Y			20.7
37	2	34	Y					Y	20.6
38	3	479	Y		Y	Y			17.1
39	4	90	Y	Y	Y	Y			15.6
40	2	215		Y		Y			11.2
41	4	9		Y	Y	Y	Y		11.1
42	1	1198				Y			10.8
43	3	271	Y	Y		Y			10.7
44	2	1796	Y			Y			8.7
45	1	1335		Y					5.1
46	0	8212							2.8
47	2	1227	Y	Y					2.2
48	1	8026	Y						2.0
	Total	**24401**	**12330**	**3457**	**1636**	**5372**	**379**	**470**	**1458**

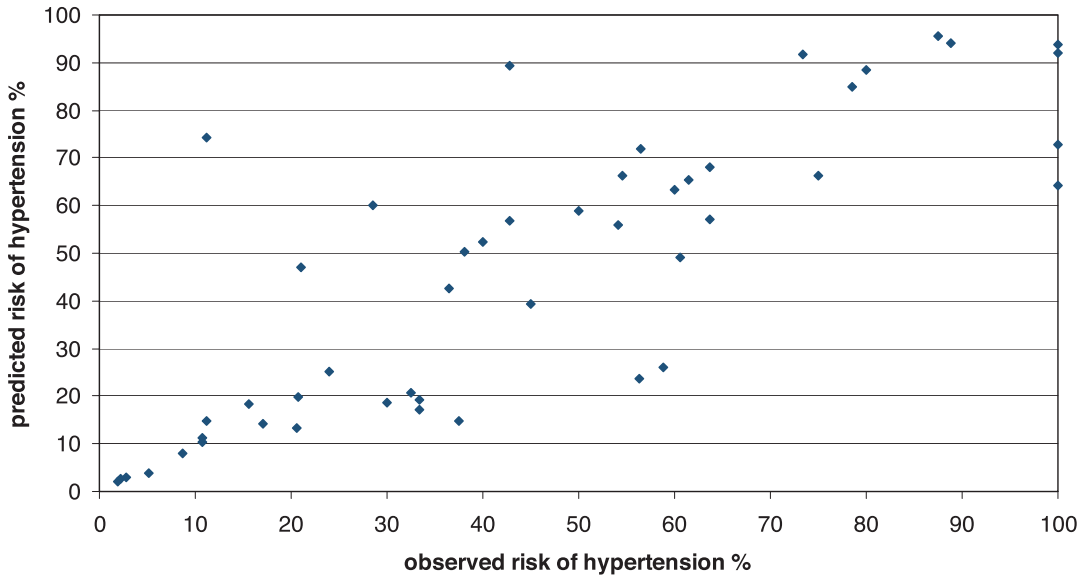
The hypertension risk ladder shows for example that:

- If no factors are present the chances of having hypertension are 2.8%
- Living in a house in council tax band A-C, the lowest bands based on value, raises the risk to 5.1%
- If other risk factors including a BMI of over 30 and being a current smoker are added the risk rises to 24%
- If CHD is also present the risk rises to 54.5%
- The highest risk categories have the most risk factors but the sample sizes in these groups are very small.
- With any four factors the average risk rises to 39.4 % and with any five factors to 67.8 %

According to the logistic regression model:

- Living in a house in council tax band A-C raises it 1.4 times
- A BMI>30 1.4 times
- Current smoker 4.2 times
- Diabetes 7.6 times
- CHD 8.6 times

Figure A2: Predicted risk of hypertension versus observed risk based on logistic regression



APPENDIX 3: STROKE RISK LADDER

Case	Number of factors	Number of patients	Male	CHD	Hypertension	Smoker	Over 50	Observed risk of stroke %
1	0	22		Y	Y		Y	22.7
2	1	20	Y	Y			Y	20.0
3	1	16	Y	Y	Y		Y	18.8
4	0	247			Y		Y	9.3
5	1	98	Y	Y	Y	Y	Y	9.2
6	1	254	Y		Y	Y	Y	8.7
7	0	84		Y	Y	Y	Y	8.3
8	1	158	Y		Y		Y	7.0
9	0	276			Y	Y	Y	6.9
10	0	15		Y	Y	Y		6.7
11	0	16		Y			Y	6.3
12	0	37		Y		Y	Y	5.4
13	1	501	Y			Y	Y	3.2
14	0	467				Y	Y	2.8
15	0	1219					Y	2.6
16	1	1372	Y				Y	2.0
17	1	52	Y		Y			1.9
18	1	58	Y	Y		Y	Y	1.7
19	0	69			Y	Y		1.4
20	1	1919	Y			Y		0.1
21	0	1444				Y		0.1
22	1	7739	Y					0.1
23	0	8061						0.1
24	0	77			Y			0.0
25	0	13		Y				0.0
26	0	23		Y		Y		0.0
27	0	1		Y	Y			0.0
28	1	15	Y	Y				0.0
29	1	76	Y		Y	Y		0.0
30	1	39	Y	Y		Y		0.0
31	1	1	Y	Y	Y			0.0
32	1	12	Y	Y	Y	Y		0.0
Total		**24401**	**12330**	**470**	**1458**	**5372**	**4845**	**212**

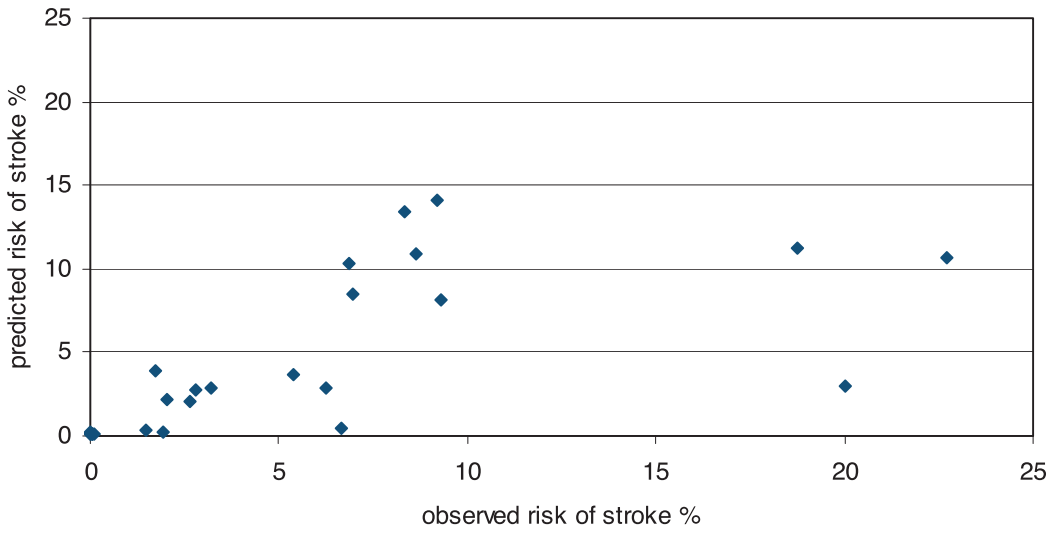
The risk ladder shows for example that:

- If there are no risk factors the risk of stroke in the population is 0.1%
- The risk increases to 8.3% if the person is over 50 is female, a smoker and over 50 with CHD and hypertension and 9.2% if a male

The risk of a stroke increases:

- 1.06 times if a male
- 1.3 times if a smoker
- 1.35 times if diagnosed with CHD
- 4.1 times if diagnosed with hypertension
- 34.1 times if aged over 50

Figure A3: Predicted risk of stroke versus observed risk based on logistic regression



APPENDIX 4: COPD RISK LADDER

Case	Number of factors	Number of patients	Male	Over 50	Current smoker	CT A-C	Hypertension	Observed risk of COPD %
1	4	29	Y	Y		Y	Y	17.2
2	3	75		Y		Y	Y	8.0
3	4	76		Y	Y	Y	Y	7.9
4	2	172		Y		Y		6.4
5	4	91	Y	Y	Y	Y		5.5
6	2	194		Y			Y	4.6
7	5	68	Y	Y	Y	Y	Y	4.4
8	3	145	Y	Y			Y	4.1
9	3	77		Y	Y	Y		3.9
10	3	182	Y	Y		Y		3.8
11	4	284	Y	Y	Y		Y	3.2
12	3	284		Y	Y		Y	2.8
13	2	427		Y	Y			2.8
14	1	1063		Y				2.1
15	2	1210	Y	Y				1.7
16	3	468	Y	Y	Y			1.5
17	3	68	Y		Y		Y	1.5
18	2	1179	Y			Y		0.3
19	3	332	Y		Y	Y		0.3
20	1	1249				Y		0.2
21	2	1626	Y		Y			0.2
22	1	6575	Y					0.1
23	1	1185			Y			0.1
24	0	6825						0.1
	Total	**24401**	**12330**	**4845**	**5372**	**3879**	**1458**	**164**

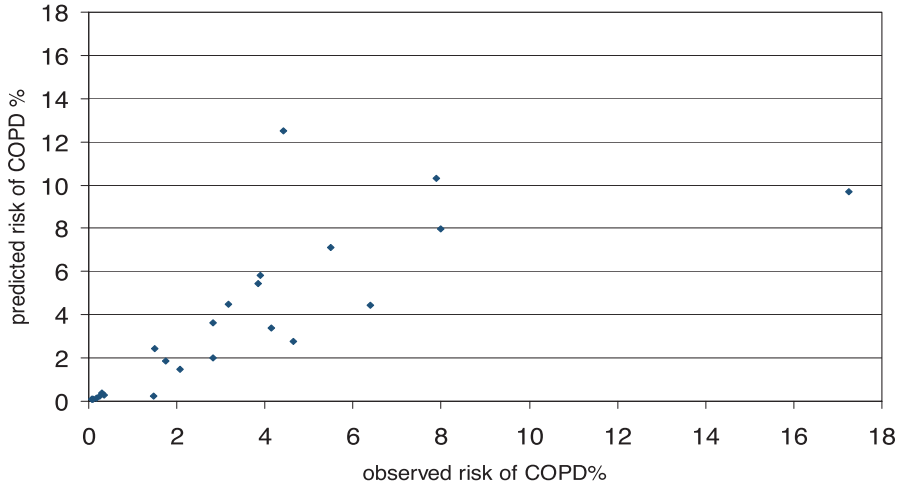
The risk ladder shows for example

- COPD is less prevalent than other chronic diseases in the group with a prevalence of 0.7%
- The main indication is being over 50 years old
- COPD is most frequently associated with males, over 50 years old living in lower value housing
- Smoking is a significant risk factor but there are many that smoke that do not yet have COPD

The risk of COPD increases:

- 1.2 times if male
- 1.3 times if current smoker
- 1.9 times if suffering from hypertension
- 3.0 times if living in house in council tax bands A-C
- 19.7 times if over 50

Figure A4: Predicted Risk of COPD versus observed risk based on logistic regression



APPENDIX 5: NOTE ON THE STANDARD ERROR OF OBSERVED RISK ESTIMATES AND CONSTRUCTING CONFIDENCE INTERVALS

In a typical risk ladder the number of observations range from a few to thousands, whilst the observed risk can range from zero to 100%. In repeated samples of size n with the risk of an occurrence equal to r the expected number of cases at risk will be *nr* with a standard error of $\sqrt{r(1-r)/n}$ assuming a normal approximation to the binomial distribution. Thus, if the sample size is 40 and the observed risk is 27% then the standard error of the risk estimate is approximately ±7% (i.e. 20% to 34%) of the mean based on the graph. See point P in Figure A5.1, which is a plot of sample size up to 100 against risk (%).

It is considered usual practice to use the normal approximation to the binomial distribution only when *nr>5* and *n(1-r)>5*. The boundary condition meeting these criteria is indicated by the dotted line in the domain annotated by A. In this area, confidence intervals based on other specified levels of confidence are generally approximated using:

$$\hat{r} - z_{\alpha/2}\sqrt{\frac{\hat{r}(1-\hat{r})}{n}} < r < \hat{r} + z_{\alpha/2}\sqrt{\frac{\hat{r}(1-\hat{r})}{n}}$$

Where $\hat{r}$ is the observed risk estimate or *x/n* and $z_{\alpha/2}$ is the number of standard deviations using the standard normal distribution corresponding to a chosen level of confidence (1-α)100%. In Figure A5.1 z equals one and the confidence level is 68%, that is the true risk of P in the example above is 27%±7% with a 68% confidence level. For a 95% level of confidence substitute 1.96 for *z*.

The majority of the risk estimates given in this paper for individual factor combinations generally falls with domain A. Of the 182 separate risk estimates in the five risk ladders in the paper, the normal approximation is appropriate for 111 (61%). Figure A5.2 is a plot of sample sizes of 100 or less versus observed risk. The darker points are risk observations that satisfy the normal approximation. Of all such cases 71% have a risk value $\hat{r} - 4\% \le r \le r + 4\%$ with a 68% probability.

However, there are a few classes of exceptions that fall outside A. Where the estimated risks are smaller than say 10%, the sample sizes n need to be much larger than 100 in order to qualify. For very small risks <1%, they need to be larger still, although the usefulness of assigning confidence intervals to such examples may be questionable as the risk is so small anyway. For observed risks in the range 10% to 90% a sample of at least 55 cases is recommended although this can be as few as 10 where the risk is around 50%. For higher risk estimates >70%, say, the sample size conditions are rarely met and so such risk estimates should be treated with caution or as indicative only. There are very few cases in this category as is evident from Figure A5.1. For small samples outside domain A, confidence intervals can still be constructed using special tables. However, the resulting confidence intervals may be so wide as to be of no value. For example for $x = 4$ and $n = 10$, $\hat{r} = 0.25$ with a 95% confidence interval is *12%<r<75%*, which would be considered quite wide.

Figure A5.1: Graph showing the confidence intervals for different sample sizes and observed levels of risk (measured as a percentage)

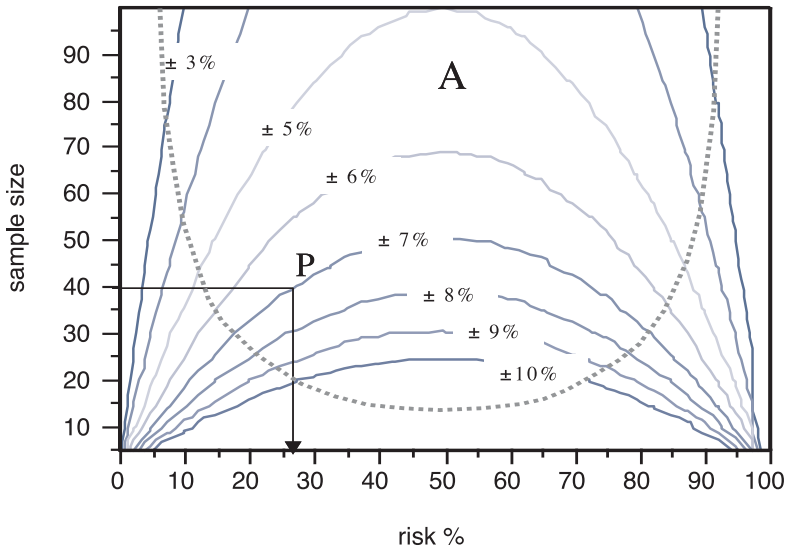
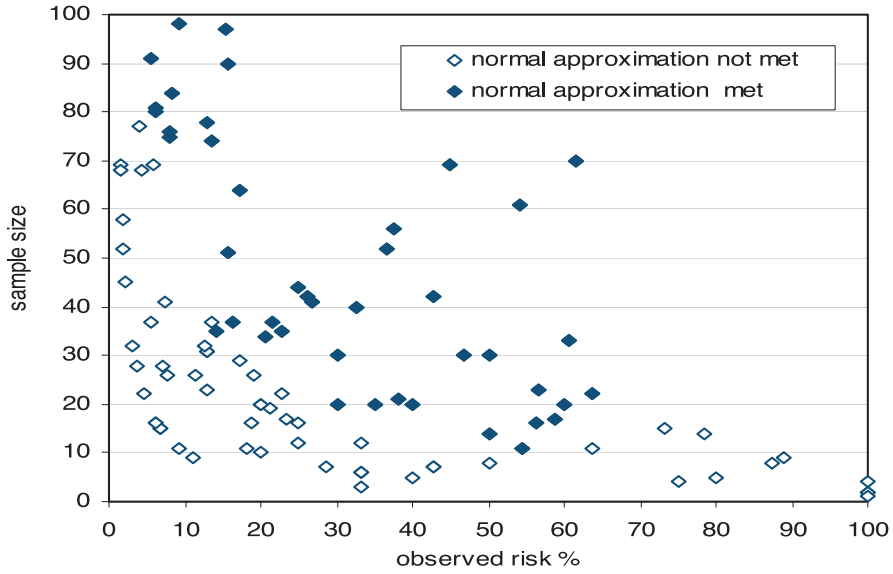


Figure A5.2: Graph showing the standard deviation for different sample sizes and levels of risk (measured as a percentage)



APPENDIX 6: SAMPLE CROSS CHECK OF LITERATURE AGAINST RESULTS

The quantitative statistics from existing literature were checked against the risk ladders deduced from patient data to see if they could be corroborated.

Results reported in existing literature	Islington risk ladders	Comparison against existing literature
Risk of heart attack between ages 55-60 is 2% for men[43]	Risk of CHD is 4.4% for male over 50	Heart Attack resulting from CHD so would expect lower risk for this event. Based on 911 males over 50.
Hypertension raises risk of CHD 2-3 times[44]	Hypertension doubles risk	Risk from 5 factors including hypertension is 25%, without hypertension risk is 12.5%
Obesity raises risk of CHD 3-fold[45]	Obesity reduces risk	Risk with 5 factors including obesity is 25%, without obesity risk is 26.2%. This is counter-intuitive and the anomaly could be explained by low numbers of patients involved.
3% of those over 40 are diabetic[46]	If over 50, chance of diabetes is 1-1.5%	Based on stand-alone risk for males and females

APPENDIX 7: DESCRIPTION OF THIN DATA FROM EPIC

THIN collects data from GP practices automatically using the Vision practice management system. Only those practices that are routinely recording data on their computer system are included in the THIN data collection scheme.

Data collected from the practice system are anonymised at the collection stage; therefore, identifying information is never made available to researchers. The data available to researchers consists of demographic, medical and prescription information at individual patient level. In addition, there is some information on referral to specialists, diagnostics and laboratory results, some lifestyle characteristics and other measurements taken in the GP practice. The data are organised in files by individual practice and provide a longitudinal record for each patient. Practices and patients are assigned computer generated identifiers which are encrypted prior to availability.

The extract comprises data taken from 255 GP surgeries as at January 2005. These 255 surgeries collectively represent approximately 3,975,000 patients that were registered with GPs in the UK between 1989 and 2004.

APPENDIX 8: OTHER RESULTS FROM THIN MODELLING

The Base Tables correspond to 2003 utilisation. We have therefore included some suggested trends, based on calendar trends modelled from the data. There were no significant trends for referrals per 1,000 people, or for admissions.

Type of service	Suggested annual trend
GP Visits per person	3-4% per year
Prescription Drugs per person	6-7% per year

Included below are the rest of the results from the THIN modelling, with relativities for different levels of risk factors.

A) Prescription drugs

This base table corresponds to the number of prescription drugs issued per person per year in 2003 for a non-smoker with no diagnosed chronic diseases and normal Body Mass Index. The tables below show multiplicative relativities, which should be applied to this base table to estimate utilisation for other combinations of risk factors and chronic diseases.

Age	Males	Females
0-49	10.7	13.8
50-54	13.3	17.2
55-59	15.3	19.0
60-64	18.1	21.5
65-69	20.6	23.7
70-74	22.5	25.7
75-79	23.8	27.7
80+	25.3	30.0

Smoker Status Relativities

Smoker status	Relativity
Non-Smoker	1.00
Ever Smoked	1.03
Current Smoker	1.03

BMI Status Relativities

BMI status	Relativity
Underweight (<20)	1.06
Normal (20-25)	1.00
Overweight (25-30)	1.04
Obese (30-35)	1.13
Morbidly Obese (35-40)	1.22
Morbidly Obese 40+	1.31

Diagnosis Relativities

Sequence	COPD	Hyper-tension	Stroke	CHD	Diabetes	Relativity
1	Y	Y	Y	Y	Y	5.21
2	Y	Y	Y	Y	N	3.77
3	Y	Y	Y	N	Y	3.88
4	Y	Y	Y	N	N	2.99
5	Y	Y	N	Y	Y	6.04
6	Y	Y	N	Y	N	4.02
7	Y	Y	N	N	Y	4.03
8	Y	Y	N	N	N	2.96
9	Y	N	Y	Y	Y	3.84
10	Y	N	Y	Y	N	3.18
11	Y	N	Y	N	Y	3.56
12	Y	N	Y	N	N	2.49
13	Y	N	N	Y	Y	4.57
14	Y	N	N	Y	N	3.54
15	Y	N	N	N	Y	3.61
16	Y	N	N	N	N	2.35
17	N	Y	Y	Y	Y	3.70
18	N	Y	Y	Y	N	2.65
19	N	Y	Y	N	Y	3.19
20	N	Y	Y	N	N	2.16
21	N	Y	N	Y	Y	3.83
22	N	Y	N	Y	N	2.80
23	N	Y	N	N	Y	3.17
24	N	Y	N	N	N	2.09
25	N	N	Y	Y	Y	3.46
26	N	N	Y	Y	N	2.28
27	N	N	Y	N	Y	2.61
28	N	N	Y	N	N	1.42
28	N	N	N	Y	Y	3.58
30	N	N	N	Y	N	2.23
31	N	N	N	N	Y	2.32
32	N	N	N	N	N	1.00

B) Referrals per 100 people per year

This base table corresponds to the number of referrals to consultants per 100 people per year in 2003 for a non-smoker with no diagnosed chronic diseases and normal Body Mass Index. The tables below show multiplicative relativities, which should be applied to this base table to estimate utilisation for other combinations of risk factors and chronic diseases.

Age	Males	Females
0-49	2.3	3.7
50-54	2.7	3.7
55-59	3.0	3.6
60-64	3.2	3.6
65-69	3.3	3.6
70-74	3.5	3.6
75-79	3.8	3.7
80+	3.6	3.4

Smoker Status Relativities

Smoker status	Relativity
Non-Smoker	1.00
Ever Smoked	1.03
Current Smoker	1.03

BMI Status Relativities

BMI status	Relativity
Underweight (<20)	1.04
Normal (20-25)	1.00
Overweight (25-30	0.99
Obese (30-35)	0.98
Morbidly Obese (35-40)	0.98
Morbidly Obese 40+	0.94

Diagnosis Relativities

No of chronic diseases	Relativity
0	1.00
1	1.38
2	1.79
3	2.25
4/5	3.06

C) Admissions per 100 people per year

This base table corresponds to the number of inpatient hospital admissions per 100 people per year in 2003 for a non-smoker with no diagnosed chronic diseases and normal Body Mass Index. The tables below show multiplicative relativities, which should be applied to this base table to estimate utilisation for other combinations of risk factors and chronic diseases.

Age	Males	Females
0-49	6.3	7.4
50-54	6.1	7.0
55-59	7.2	6.9
60-64	8.3	7.2
65-69	9.2	7.9
70-74	11.9	9.5
75-79	14.0	11.4
80+	16.4	14.8

Smoker status relativities

Smoker status	Relativity
Non-Smoker	1.00
Ever Smoked	1.17
Current Smoker	1.03

BMI status relativities

BMI status	Relativity
Underweight (<20)	1.32
Normal (20-25)	1.00
Overweight (25-30)	0.97
Obese (30-35)	1.01
Morbidly Obese (35-40)	1.02
Morbidly Obese 40+	1.15

Diagnosis relativities

No of chronic diseases	Relativity
0	1.00
1	1.67
2	2.70
3	4.16
4/5	5.84

APPENDIX 9: DESCRIPTION OF GLIM TECHNIQUES[47]

A traditional linear model is of the form

$$y_i = \mathrm{x}_i^{\prime}\beta + \varepsilon_i$$

where *yi* is the response variable for the *i*th observation. The quantity x_i is a column vector of covariates, or explanatory variables, for observation i that is known from the experimental setting and is considered to be fixed, or non-random. The vector of unknown coefficients β is estimated by a least squares fit to the data y. The ε_i are assumed to be independent, normal random variables with zero mean and constant variance. The expected value of y_i, denoted by, μ_i is

$$\mu_i = \mathrm{x}_i^{\prime}\beta$$

While traditional linear models are used extensively in statistical data analysis, there are types of problems for which they are not appropriate.

- It may not be reasonable to assume that data are normally distributed. For example, the normal distribution (which is continuous) may not be adequate for modelling counts or measured proportions that are considered to be discrete.
- If the mean of the data is naturally restricted to a range of values, the traditional linear model may not be appropriate, since the linear predictor $\mathrm{x}_i^{\prime}\beta$ can take on any value. For example, the mean of a measured proportion is between 0 and 1, but the linear predictor of the mean in a traditional linear model is not restricted to this range.
- It may not be realistic to assume that the variance of the data is constant for all observations. For example, it is not unusual to observe data where the variance increases with the mean of the data.

A generalised linear model extends the traditional linear model and is, therefore, applicable to a wider range of data analysis problems. A generalised linear model consists of the following components:

- The linear component is defined just as it is for traditional linear models: $\eta_i = \mathrm{x}_i^{\prime}\beta$
- A monotonic differentiable link function g describes how the expected value of yi is related to the linear predictor η_i:
  $g(\mu_i) = \mathrm{x}_i^{\prime}\beta$
- The response variables y_i are independent for i = 1, 2,...and have a probability distribution from an exponential family. This implies that the variance of the response depends on the mean through a *variance function V*:

$$var(y_i) = \frac{\phi V(\mu_i)}{\omega_i}$$

where ϕ is a constant and w_i is a known weight for each observation. The *dispersion parameter* ϕ is either known (for example, for the binomial or Poisson distribution, ϕ = 1) or it must be estimated.

As in the case of traditional linear models, fitted generalized linear models can be summarised through statistics such as parameter estimates, their standard errors, and goodness-of-fit statistics. One can also make statistical inference about the parameters using confidence intervals and hypothesis tests. However, specific inference procedures are usually based on asymptotic considerations, since exact distribution theory is not available or is not practical for all generalised linear models.

The model fitting procedure

We fit generalised linear models to the data by maximum likelihood estimation of the parameter vector β. There is, in general, no closed form solution for the maximum likelihood estimates of the parameters. Therefore the parameters of the model are estimated numerically through an iterative fitting process. The dispersion parameter ϕ is also estimated by maximum likelihood or, optionally, by the residual deviance or by Pearson's chi-square divided by the degrees of freedom. Covariances and standard errors are computed for the estimated parameters based on the asymptotic normality of maximum likelihood estimators.

We generally use a log link function, of the form:

- log: $g(\mu) = \log(\mu)$

with either a Poisson or gamma distribution, with a variance function of:

- Poisson: $V(\mu) = \mu$
- gamma: $V(\mu) = \mu^2$

An important aspect of generalised linear modelling is the selection of explanatory variables in the model. Changes in goodness-of-fit statistics are often used to evaluate the contribution of subsets of explanatory variables to a particular model. The deviance, defined to be twice the

difference between the maximum attainable log likelihood and the log likelihood of the model under consideration, is often used as a measure of goodness of fit. The maximum attainable log likelihood is achieved with a model that has a parameter for every observation.

We fit a sequence of models, beginning with a simple model with only an intercept term, and then include one additional explanatory variable in each successive model. One can measure the importance of the additional explanatory variable by the difference in deviances or fitted log likelihoods between successive models. We use asymptotic tests to assess the statistical significance of the additional term.

We compute:

- Wald statistics and likelihood ratio statistics for each term in the model and *p*-values based on their asymptotic chi-square distributions
- estimated values, standard errors, and confidence limits for user-defined contrasts and least-squares means
- confidence intervals for model parameters based on either the profile likelihood function or asymptotic normality

Overview of the statistical methods used

The GENMOD procedure we use in SAS© fits generalised linear models, as defined by Nelder and Wedderburn (1972)[48]. The class of generalised linear models is an extension of traditional linear models that allows the mean of a population to depend on a *linear predictor* through a nonlinear *link function* and allows the response probability distribution to be any member of an exponential family of distributions. Many widely used statistical models are generalised linear models. These include classical linear models with normal errors, logistic and probit models for binary data, and log-linear models for multinomial data. Many other useful statistical models can be formulated as generalised linear models by the selection of an appropriate link function and response probability distribution. McCullagh and Nelder (1989)[49] include a discussion of statistical modelling using generalised linear models.

The analysis of correlated data arising from repeated measurements when the measurements are assumed to be multivariate normal has been studied extensively. However, the normality assumption may not always be reasonable; for example, different methodology must be used in the data analysis when the responses are discrete and correlated. Generalised Estimating Equations (GEEs) provide a practical method with reasonable statistical efficiency to analyse such data.

Liang and Zeger (1986)[50] introduced GEEs as a method of dealing with correlated data when, except for the correlation among responses, the data can be modelled as a generalized linear model. For example, correlated binary and count data in many cases can be modelled in this way.

The GENMOD procedure can fit models to correlated responses by the GEE method. One can use PROC GENMOD to fit models with most of the correlation structures from Liang and Zeger (1986) using GEEs. Refer to Liang and Zeger (1986), Diggle, Liang, and Zeger (1994)[51], and Lipsitz, Fitzmaurice, Orav, and Laird (1994)[52] for more details on GEEs.

References

1 Mayhew (2000) Health and Elderly Care Expenditure in an Aging World. RR-00-21, International Institute for Applied Systems Analysis, Laxenburg, Austria.

2 Wanless; Securing our future health: Taking a long term view, April 2002. Crown copyright

3 Department of Health (2000), The NHS Plan – A plan for investment, a plan for reform, Cmd 4818-1 The Stationery Office, London. Crown copyright

4 Saving Lives: Our Healthier Nation, HSC 1999/152, HMSO. Crown copyright

5 Choosing Health; Making Healthier Choices Easier; Public Health White Paper (2004). Department of Health. Crown copyright

6 Wanless D., Securing Good Health for the Whole Population, February 2004. Crown Copyright

7 EPIC is a research data service provider. The THIN longitudinal database takes data from Vision software, which is provided to GPs by InPractice Systems. These data are based on frequent downloads from over 300 practices comprising nearly 5.0m old and currently patients and available 3 times a year, with data going back as early as 1985. At the time of our analysis the sample was drawn from 255 practices and nearly 4 million patients.

8 British Heart Foundation, Coronary Heart Disease Statistics 2005; based on 1999 and 2003 Health Surveys for England. www.bhf.org.uk

9 World Heart Federation Fact-Sheet, 2002. www.worldheart.org

10 British Heart Foundation Fact-Sheet. www.bhf.org.uk

11 British Heart Foundation, Coronary Heart Disease Statistics 2005. www.bhf.org.uk; based on 1999 and 2003 Health Surveys for England

12 Continuous Mortality Investigation Report 20, The Mortality of Impaired Assured Lives

13 British Heart Foundation Factsheet 2005. www.bhf.org.uk; based on UK deaths 2003 (ONS, General Register Office Edinburgh, General Register Office NI)

14 Wanless D., Securing our future health: Taking a long term view, April 2002. Crown Copyright.

15 Incidence from Office of Health Economics, 1989, London; Health Survey for England, 1991, Population from Office for National Statistics (ONS)

16 Health Survey for England, 1998. Department of Health

17 The Stroke Association, Q&A Fact-Sheet 2001. www.stroke.org.uk

18 ONS, General Register Office Edinburgh, General Register Office NI

19 Continuous Mortality Investigation Report 20, The Mortality of Impaired Assured Lives

20 Burdens of Disease: A discussion document, 1996. Department of Health

21 Stroke: Pathways to future therapy, Scrip Reports PJP Publications 1998

22 The Blood Pressure Association. www.bpassoc.org.uk

23 British Heart Foundation, Coronary Heart Disease Statistics 2005. www.bhf.org.uk; based on 1999 and 2003 Health Surveys for England. Department of Health

24 Continuous Mortality Investigation Report 20, The Mortality of Impaired Assured Lives

25 National Institute for Health and Clinical Excellence, press release 2004/036

26 1999 and 2003 Health Surveys for England

27 British Heart Foundation Fact-Sheet. www.bhf.org.uk

28 Continuous Mortality Investigation Report 20, The Mortality of Impaired Assured Lives

29 Wanless D., Securing our future health: Taking a long term view, April 2002. Crown copyright

30 www.doh.gov.uk

31 British Lung Foundation. www.lunguk.org

32 Continuous Mortality Investigation Report 20, The Mortality of Impaired Assured Lives

33 National Institute for Health and Clinical Excellence, Management of Adults with COPD in Primary and Secondary Care, Working Draft, August 2003

34 World Heart Federation Factsheet 2002. www.worldheart.org

35 For more details and description of this methodology see: http://www.cass.city.ac.uk/ri/RI002.pdf

36 Read codes form the basis for the Quality and Outcomes Framework in the new GP contracts. For description of Read codes see www.equip.ac.uk/readCodes/docs/index.html

37 A systematic approach for the analysis of health and social risks at neighbourhood level, L. Mayhew (2005). The Risk Institute, Cass, City University. http://www.cass.city.ac.uk/ri/RI002.pdf

38 e.g. see Table VI, p.483, Freund 1973 Freund, J.E. (1973) Modern Elementary Statistics (Prentice-Hall International, London) or p.177-120 Armitage P. and Berry G. (1987) Statistical Methods in Medical Research (Blackwell, Oxford).

39 Under the current council tax system properties are banded from A, the lowest value properties, to H, the highest value properties. Any property in bands A-C is therefore in a low value category and is assumed to be an indicator of relatively low wealth.

40 GP Visits include night and emergency visits to patient's homes, as well as in-surgery consultations.

41 Note that this population will develop one of the five studied chronic diseases in the future, but they currently have no diagnosis. The population we examined does not included people who never develop one of the five chronic diseases in our study period.

42 An Insurer's Guide to the Disability Discrimination Act 1995 – ABI, January 2003.

43 British Heart Foundation: "11-2002 Understanding Risk Part II" (www.bhf.org.uk/professionals)

44 British Heart Foundation: "11-2002 Understanding Risk Part II" (www.bhf.org.uk/professionals)

45 British Heart Foundation: "11-2002 Understanding Risk Part II" (www.bhf.org.uk/professionals)

46 1999 and 2003 Health Surveys for England

47 This Appendix draws heavily from SAS online help. See http//support.sas.com/onlinedoc/912 for details.

48 Nelder, J.A. and Wedderburn, R.W.M. (1972), "Generalized Linear Models," Journal of the Royal Statistical Society A, 135, 370 - 384.

49 McCullagh, P. and Nelder, J.A. (1989), Generalized Linear Models, Second Edition, London: Chapman and Hall.

50 Liang, K.Y. and Zeger, S.L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," Biometrika, 73, 13 - 22.

51 Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994), Analysis of Longitudinal Data, Oxford: Clarendon Press.

52 Lipsitz, S.H., Fitzmaurice, G.M., Orav, E.J., and Laird, N.M. (1994), "Performance of Generalized Estimating Equations in Practical Situations,"