Environmental and behavioural determinants of geographic variation in coronary heart disease in England: an ecological study

Submitted for the degree of doctor of philosophy (medical sciences), October 2009

Peter Scarborough, Linacre College

Contents

Environmental and behavioural determinants of geographic variation in coronary heart disease in England: an ecological study

Peter Scarborough, Linacre College, submitted for the degree of doctor of philosophy, Trinity term 2009

ABSTRACT

Coronary heart disease rates show substantial geographic variation in England, which could be due to environmental variables (e.g. climate, air quality) or behavioural risk factors for coronary heart disease within populations. Previous work investigating this geographic variation has either used ecological analysis (i.e. areas as units of observation) or individual-level analysis. Ecological studies have been unable to account adequately for differences in behavioural risk factors within populations; individual-level studies have been under-powered at the area-level to include all potentially explanatory environmental variables. This thesis reports on ecological multi-level and spatial error regression analyses of coronary heart disease mortality and hospitalisation rates for all wards in England using environmental variables and synthetic estimates of the prevalence of behavioural risk factors as explanatory variables. Existing sets of synthetic estimates were subjected to studies of their validity. Validated synthetic estimates of the prevalence of smoking, low fruit and vegetable consumption, raised blood pressure, obesity and raised cholesterol were combined into a single index of unhealthy lifestyle to take account of collinearity between them. Final models successfully explained around 80% of large scale geographic variation (i.e. variation between wards in different areas of the country) in mortality rates for coronary heart disease and 60% in hospitalisation rates, and around 20% of the small scale geographic variation (i.e. variation between wards in close proximity) in mortality rates, and 30% in hospitalisation rates. The climate explained around 15% of large scale geographic variation in coronary heart disease rates after adjustment for the index of unhealthy lifestyle and socioeconomic deprivation. Urbanicity and air pollution explained a small amount of small scale geographic variation in coronary heart disease rates. The majority of explained geographic variation was due to the index of unhealthy lifestyle and deprivation. The results of this thesis confirm and extend findings from the British Regional Heart Study, report on the validity of synthetic estimates currently used to guide healthcare resource allocation, and introduce an index of unhealthy lifestyle that could be used in future ecological studies of chronic disease.

Thesis word count (excluding acknowledgements, bibliography, appendices, diagrams and tables): 41,500

Acknowledgments

Your honour, with all due respect, past and present, and without further to do…

I would like to thank Prof Michael Goldacre, Dr Mike Rayner and Dr Steven Allender for the insightful comments that they have provided since the work on this thesis began in September 2005. I would not have been able to complete the thesis without their gracious support at all stages of the development of this work.

The other members of the British Heart Foundation Health Promotion Research Group have also provided valuable support, either in the form of helpful advice about the subject area or what a DPhil thesis should look like, or just putting up with me tearing me hair out and complaining. Thank you Charlie, Gill, Asha, Anne, Anna, Anu, Viv, Sophie, Ca, Paul, Prachi, Nick, Justin and Dushy.

Various people have helpfully supplied me with data, advice or comments on early drafts of the thesis or individual chapters. Particular thanks go to Pat Yudkin, Andrew Neil, Ray Fitzpatrick, Liz Twigg, Chris Dibben, Graham Moon, Sarah Lewington, David Merrick, Gillian Bryant, Stuart Simms and Shaun Scholes.

I know you heard this rap before. Your honour, I mean it. This is the truth…

Most importantly of all, I am eternally grateful for the wonderful support of my family. Mum and Dad, thank you for giving me all the support I needed to get to Oxford in the first place, I will always appreciate it. Thank you Oscar and Aobh, you never allowed my mind to stay too focussed on spatial epidemiology. Instead, the four years it took to complete this thesis were filled with the joys of snakes and ladders, trampolines, dressing up, My Neighbour Totoro, Countdown and Kerwhizz.

I couldn't possibly thank Aoife enough. There certainly would be no thesis without her love and care. Thank you for making me a better person.

And I want to thank almighty God, without whom no case gets tossed.

This thesis has been written with support from a supervisory panel consisting of Prof Michael Goldacre (primary supervisor), Dr Mike Rayner and Dr Steven Allender. I devised the initial idea for the research. The important themes underlying the research area, the structure of the thesis, the general analytical techniques and the themes of the literature review were crystalised with support from both the supervisory panel and Prof Andrew Neil and Dr Pat Yudkin who conducted the 'transfer of status' interview. The literature review was conducted by me initially in February 2007 and then updated in December 2008. The identification of relevant datasets, their acquisition, manipulation and cleaning of data were all conducted by me. This included a study of synthetic estimates of the prevalence of risk factors for coronary heart disease for wards in England and the application of the published synthetic estimation models to 2001 census data – this application was a painstaking task that took many months of constructing intricate 'do' files for Stata and performing internal validity checks. I designed and executed the validity exercises for the identified synthetic estimates, which again included identification and manipulation of a number of large external datasets. The ecological analyses reported in this thesis were designed and conducted by me. The interpretations of the results generated by the analyses reported in this thesis and the conclusions that are drawn are my own.

The problem

Geographical inequalities in health within the United Kingdom have been systematically measured for over one hundred and fifty years, since the conception of the General Register Office by William Farr in 1837. During the nineteenth century, when communicable diseases were the most common cause of mortality in the United Kingdom, these geographic inequalities proved to be powerful political drivers to improve living conditions in poor inner city areas (Chadwick, 1842; Drever and Whitehead, 1997), and they provided vital epidemiological evidence for the study of the aetiology of disease - most notably the identification of contaminated water supplies as a causal factor for cholera by John Snow in 1854 (Lewes, 1983).

During the twentieth century the burden of disease in the United Kingdom fundamentally changed from high infant mortality and high death rates from communicable diseases to low infant mortality and a majority of deaths from chronic disease such as cardiovascular disease and cancer (Drever and Whitehead, 1997). This is illustrated by the pattern of coronary heart disease (CHD) over the first half of the twentieth century, which showed a slow but steady increase in mortality from the start of the century until the late 1960s (Griffiths and Brock, 2003). With this epidemiological transition from communicable to chronic disease there came a change in focus from improving the environment in which

people lived to identifying and reducing behavioural risk factors for chronic disease. Since the late 1960s cardiovascular disease mortality rates have been in decline (Griffiths and Brock, 2003), due in part to reductions in the prevalence of behavioural risk factors for CHD, particularly smoking which had been identified as a causal factor for lung cancer and other conditions in the 1950s (Doll and Hill, 1954). However, CHD is still the single biggest killer in the United Kingdom, responsible for around 95,000 deaths every year, around 30,000 of which occur before the age of 75 (Allender et al., 2008). CHD costs the UK economy around £9 billion annually, around £3.2 billion of which is due to health care costs and the remainder due to production losses resulting from early mortality, morbidity, or informal care of people suffering from CHD (Allender et al., 2008). The disease is a massive drain on hospital resources: in 2006 there were nearly 430,000 admissions for CHD in English NHS hospitals (Department of Health, 2008).

Geographical inequalities in CHD rates in England are substantial and persistent. Since the late 1970s, male CHD mortality rates have been at least 30% higher in the North of England than in the South East, and the relative differences between North and South for female rates have been even larger (Scarborough et al., 2008). Mirroring this large scale geographic variation in CHD rates are substantial small scale geographic variations, where the highest female mortality rates for CHD in local authorities in the South East of England are more than double those of the lowest in the same region, and neighbouring wards within local authorities can experience considerably different CHD mortality rates (Scarborough et al., 2008). Tackling geographical inequalities in health has long been a priority to both the UK Government and health practitioners for both humanitarian and

pragmatic reasons. The humanitarian argument for tackling inequalities is clear - the desire for equity of health experience was a driving force behind the establishment of the NHS and it remains one of its core values. In the foreword of the National Service Framework for CHD Alan Milburn (the health secretary at the time of publication) described the 'postcode lottery' of health care as unacceptable (Department of Health, 2000a). The pragmatic argument lies with another core value of the NHS, that of efficiency. If all local authorities shared the same CHD mortality rate as Kensington & Chelsea then there would be over 32,000 fewer deaths from CHD in England every year (Scarborough et al., 2008) – the fact that low mortality rates are attained in some areas implies that they are an achievable target with modern standards of prevention and treatment.

Historically, geographic variations in CHD mortality have been used to generate hypotheses about its aetiology, just as was the case for many communicable diseases in the nineteenth century. The development of epidemiology as a research discipline, and the discovery that CHD mortality affected different societies with differing strength, provided compelling evidence that CHD is a chronic disease with modifiable risk factors, rather than an inevitable consequence of ageing (Stamler, 2005). The prospective studies conducted after the second world war such as the Framingham study helped to establish the major risk factors (smoking, high blood pressure and high blood cholesterol) (Kannel et al., 1961), and later huge prospective trials such as the Multiple Risk Factor Intervention Trial confirmed their causal association with the development of CHD (MRFIT Research Group, 1990; Epstein, 2005). Today, it is generally agreed that there

are four established non-modifiable risk factors - age, sex, ethnicity and family history of CHD. There are nine established modifiable risk factors for CHD - smoking, poor diet, high alcohol intake, physical inactivity, high stress levels, obesity, raised blood pressure, raised blood cholesterol and diabetes. The authors of the INTERHEART study (a case-control study that investigated myocardial infarction incidence in 15,152 cases and 14,820 controls in 52 countries worldwide) suggest that these nine modifiable risk factors are responsible for 94% of all heart attacks in Western Europe (Yusuf et al., 2004).

The massive predictive power of the nine modifiable risk factors for CHD would seem to suggest that virtually all of the geographic variation in CHD rates must be due to differences in the lifestyle of the populations in different areas. However, this interpretation does not sit easily with evidence that environmental variables – such as climate or air pollution - can have a big influence on CHD. The most striking of this evidence is excess winter mortality. It has been observed that CHD mortality rates are seasonal, with more deaths in winter than in summer. This phenomenon is quite substantial – in 2004/05 the winter excess accounted for an additional 6,000 deaths in the UK (Allender et al., 2008). This, and other similar observations of raised CHD mortality in particular environmental conditions, has led researchers to investigate the role of environmental variables in the development of CHD.

It is unclear how much of the geographic variation in CHD in England is a result of differences in the prevalence of established behavioural risk factors for CHD, and how much is due to differences in environmental variables such as climate and air pollution.

Previous studies that have addressed this issue have either used data on individuals collected from different sites but have been under-powered at the area-level to consider more than one environmental variable simultaneously (Morris et al., 2001; Lawlor et al., 2003), or have used area-level data and have been unable to adjust analyses adequately for behavioural risk factors for CHD (Pocock et al., 1980; Maheswaran et al., 1999). It is the aim of this thesis to estimate the amount of geographic variation in CHD rates in England that is a result of environmental variables and how much is a result of differences in behavioural risk factors between populations.

Why is this important?

The ground breaking work that identified the risk factors for CHD was based on epidemiological principles: identifying subgroups of the population that were more likely to suffer CHD and analysing the common factors among these subgroups. The identification of the risk factors led to a greater understanding of the physiological aetiology of heart disease, which in turn allowed for the development of improved methods of treatment and prevention. A greater understanding of the impact of environmental variables could lead to similar advances in the understanding of the physiological aetiology of CHD.

Investigating the impact of environmental variables on CHD could lead to new interventions aimed at reducing CHD in the population, such as initiatives to improve the insulation of housing, or to reduce air pollution in built up urban areas. Although these

are generally seen as important for reasons other than cardiovascular health (such as the link between air pollution and asthma (Sandstrom and Kelly, 2009)), evidence of the influence of environmental variables on CHD would increase the demand for the implementation of such interventions and could therefore have a wide ranging impact on health.

Both mortality and hospitalisation rates are used as outcome variables in the analyses in this thesis, and this is important for a number of reasons. It has been observed that the geographic pattern of CHD mortality rates and hospitalisation rates around England differ in many important regards. This is illustrated in figures 1.1a and 1.1b, taken from a statistical compendium exploring geographic and social inequalities in CHD in the United Kingdom (Scarborough et al., 2008). Figure 1.1a shows CHD mortality rates by local authority, and figure 1.1b shows hospitalisation (finished consultant episodes, or FCEs) rates for CHD by local authority. The differences in the patterns are evident. For example, high hospitalisation rates for CHD appear to be clustered in urban centres (particularly in the North West, and West Yorkshire), and are also found in the East and South West of England, in contrast to the geographic pattern in mortality rates.

Figure 1.1a Age-standardised male mortality rates for CHD, 2001-06 (local authorities, n = 354). Taken from *Regional and Social Differences in Coronary Heart Disease* (Scarborough et al., 2008)

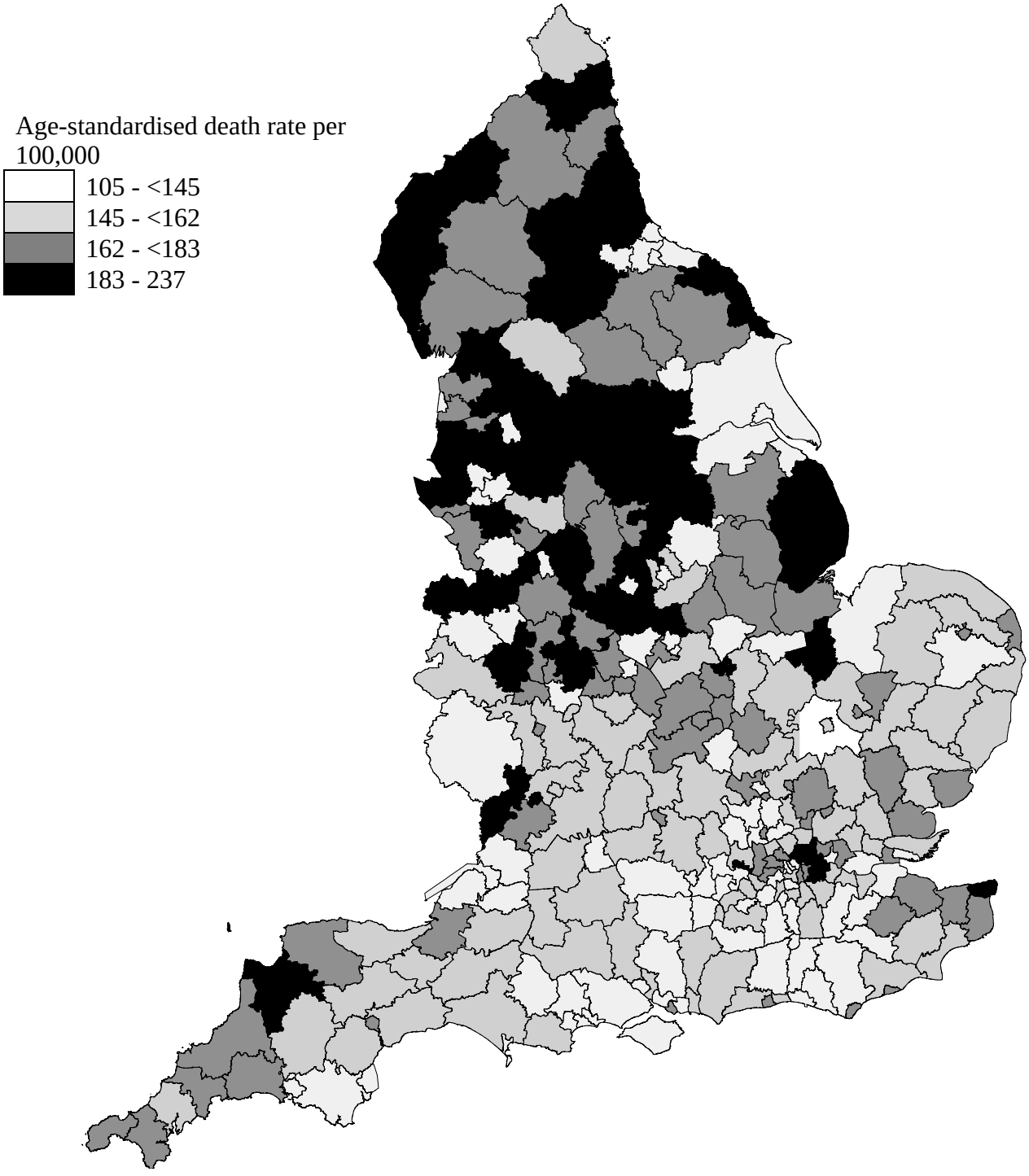
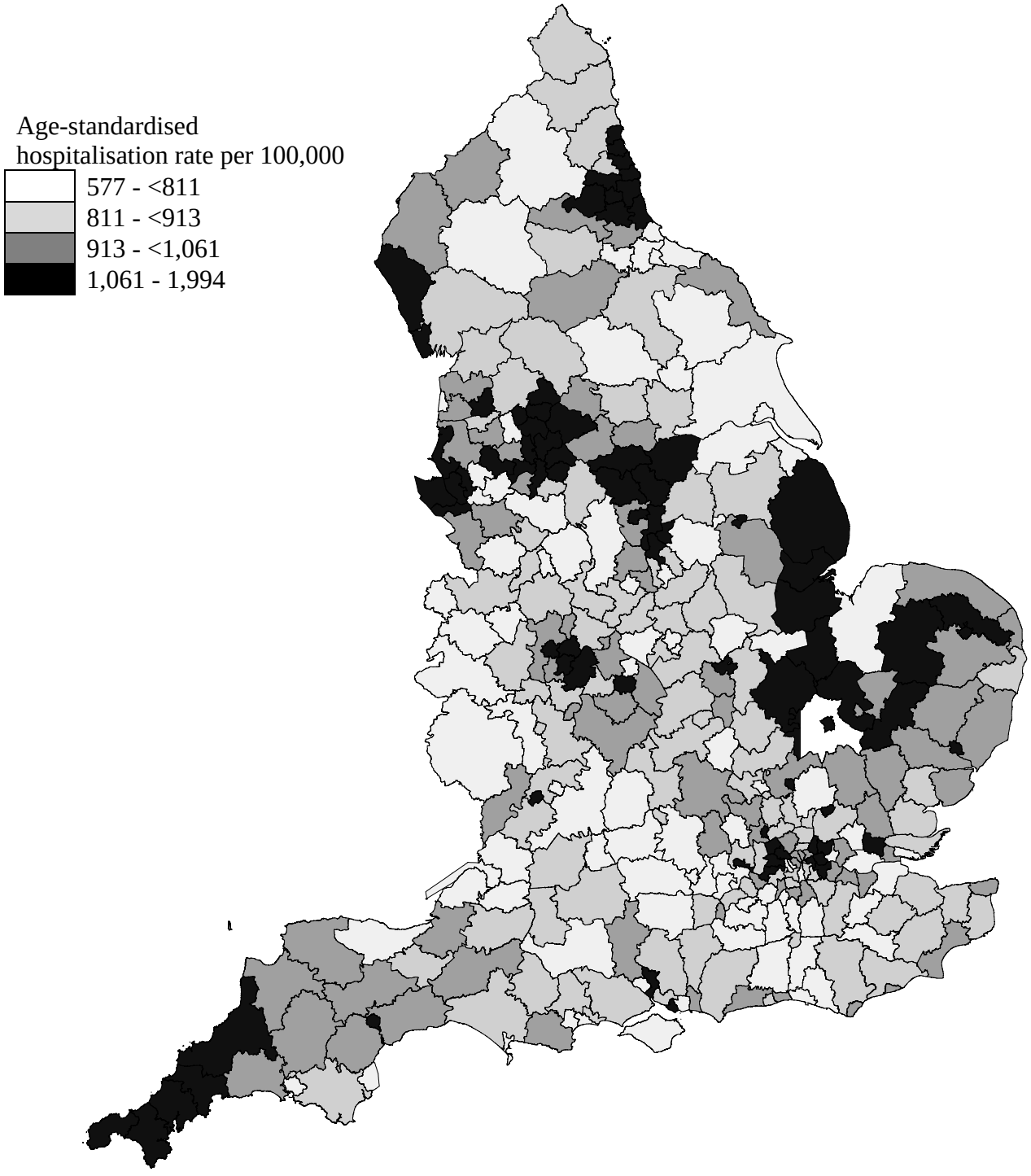


Figure 1.1b Age-standardised male hospitalisation rates for CHD, 2001-06 (local authorities, n = 354). Taken from *Regional and Social Differences in Coronary Heart Disease* (Scarborough et al., 2008)



Another reason for including both mortality and hospitalisation rates in the analyses reported here is that they measure different aspects of CHD. The hospitalisation admissions data can be viewed in some senses as a proxy for heart disease morbidity, but they are also a measure of the use of health care resources within an area. Whilst the human toll of mortality and morbidity are important factors to investigate in their own right, outcomes that track health care resources are necessary as the current resource use related to CHD is vast and is likely to grow over the course of the next ten years as a result of an ageing population (Capewell et al., 2009). Finally, inclusion of the two measures could aid greater insights into the aetiology of CHD. For example, a hypothesis that an environmental variable is likely to produce severe myocardial infarction that is more likely to result in sudden death would be supported by the finding that the environmental variable is more strongly associated with increased mortality rates than with increased hospital admission rates.

The structure of the thesis

The thesis is presented in nine chapters, including this introduction. Chapters two, three and four are, respectively, a literature review, a description of the study design and a description of the statistical methods used. The literature review aims to establish the current understanding of geographic variations in CHD in the UK and separately evaluates estimates of the impact of *contextual factors* (i.e. environmental variables) and *compositional factors* (i.e. differences between areas in the behavioural risk factor profile of populations). The results of the literature review inform the study design for the thesis,

which is comprehensively detailed in chapter three. The analyses in the thesis use a variety of statistical techniques – the methods, required assumptions and interpretations of the output of the techniques are explained in chapter four. The four subsequent chapters describe the analyses that have been conducted for the thesis. Chapter five reports an analysis of the amount of geographic variation in CHD that is explained purely by environmental variables. Chapter six reports the findings of a validation exercise that assessed the accuracy of sets of model-based estimates of the prevalence of behavioural risk factors for wards in England. The sets of model-based estimates that are shown to be valid and accurate in this chapter are then included in the analysis for chapter seven, which estimates the amount of geographic variation in CHD that is explained by differences in the behavioural risk factor profiles of populations. Chapter eight reports an assessment of the impact of confounding of the relationship between environmental variables and CHD by both the behavioural risk factor profiles of populations and socio-economic deprivation. The final chapter of the thesis presents the conclusions that are drawn from the analyses presented in chapters five through eight. There are four appendices to the thesis, the first three of which provide the complete set of results of the analyses that are reported in the thesis, and the fourth is an article published in the peer-reviewed journal *Health & Place* based on the validation exercise reported in chapter six.

Terminology

Throughout this thesis, the risk factors for CHD are considered in three separate groups: non-modifiable risk factors (e.g. sex, age); modifiable individual-level risk factors (e.g.

smoking, poor diet, raised cholesterol); modifiable[1] environmental risk factors (e.g. climate, air pollution). In general in this thesis, the second group of risk factors are referred to simply as *behavioural risk factors*. This is somewhat counter-intuitive since the medical risk factors (obesity, raised blood pressure, raised cholesterol and diabetes) are clearly not measures of behaviour, although they are all strongly dependent upon behaviour. The term 'behavioural' is used rather than 'individual-level' to avoid the implication that individual-level data have been used in the ecological analyses that are presented here. Similarly, the term 'behavioural risk factor profile of the population' is used as a group term for the prevalence of behavioural risk factors for CHD (e.g. the prevalence of smoking, the prevalence of obesity).

The terms 'large scale' and 'small scale' are used throughout this thesis to describe the geographic variation in CHD rates in England. Here, 'large scale' refers to the general difference in CHD rates that is found between large geographical regions of England, such as the North and South of England, and 'small scale' refers to the difference between two small areas in close proximity to each other, such as the difference in CHD mortality rate between two wards within a single city. Within the multilevel models that are the basis of the analysis in chapters five to eight of the thesis, large scale geographic variations are modelled by the variance in CHD rates for the local authorities around England, and small scale geographic variations are modelled by the average variance in CHD rates for all wards within a local authority.

[1] Although environmental variables such as climate may not be directly modifiable, the human experience of these variables is modifiable since individuals can modify their behaviour to better cope with the environmental factors (e.g. moving to a different area).

INTRODUCTION

This chapter aims to review the evidence regarding causal factors for geographic variation in coronary heart disease (CHD) rates in the UK. Estimates of the impact of *compositional* and *contextual* factors on geographic variation in CHD are discussed separately. Initially, the physiological plausibility of the effect of both behavioural and environmental risk factors on CHD is explored, and then the evidence of the impact on geographic variation in CHD is reported, and a brief commentary on the reviewed studies is provided.

The impact of deprivation on geographic variation in CHD is explored separately from both contextual and compositional factors. This is because the concept of an area-level deprivation score incorporates both compositional factors (the social class of the individuals living in the area) and also contextual factors (the material deprivation of an area: quality of services; community cohesion etc.). The different theoretical concepts regarding the causal chain linking deprivation, individual-level risk factors and CHD are discussed, and the evidence of the impact of deprivation on geographic variation in CHD is reported.

METHODS

The review was conducted by searching the AMED, British Nursing Index, CINAHL, EMBASE, Global Health, MEDLINE and PsycINFO databases using Medical Subject Headings (MeSH) terms that relate to coronary heart disease, cardiovascular disease, geographic inequalities and the United Kingdom, initially in April 2006 and then updated in April 2008. These terms were as follows: ‘cardiovascular diseases’, ‘myocardial ischaemia’, ‘coronary disease’, ‘environment and public health’, ‘small-area analysis’, ‘Great Britain’, ‘England’, ‘Scotland’, ‘Northern Ireland’, ‘Wales’.

Journal articles were ordered on the basis of their titles and abstracts, and relevant articles or reports that were identified in the reference list of the ordered articles were also reviewed. The initial search was restricted to studies published since 1990, but this restriction was relaxed for studies that made a substantial contribution to the field. For some of the factors (e.g. the association between air pollution and CHD rates) there were few studies based in the UK of their association with CHD rates, and therefore some relevant non-UK studies were also included in the review. Whilst not claiming to be either comprehensive or systematic, the aim of the review was to provide a good understanding of the work that has previously been reported surrounding geographic variations in CHD rates.

RESULTS

Compositional factors

There are nine established behavioural risk factors for CHD - smoking, poor diet, high alcohol intake, physical inactivity, high stress levels, obesity, raised blood pressure, raised blood cholesterol and diabetes. The physiological mechanism that links the behavioural risk factors and CHD has been well documented (Stamler, 2005), and is briefly summarised as follows. Raised levels of cholesterol in the blood results in a build up of plaque inside the arteries which can lead to atherosclerosis and thrombosis; raised blood pressure results in increased pressure on the walls of the arteries which can stiffen and produce blockages; diabetes puts extra pressure on all organs of the body, including the heart. Obesity can cause raised blood pressure and diabetes, but also independently impacts on CHD risk by putting extra pressure on the heart. Smoking increases blood pressure levels, and independently puts pressure on the heart through inhaled toxins; poor diets tend to increase blood pressure, raise cholesterol levels and increase the risk of obesity, as does a sedentary lifestyle; high alcohol intake increases blood pressure levels, as does raised stress levels. In addition, stress releases adrenaline and fibrinogen into the blood stream, which increase blood coagulability and hence the risk of thrombosis (Bartley, 2004).

It is not the purpose of this chapter to review the evidence linking these established risk factors with risk of CHD in individuals - there have been many large cohort studies designed to measure the strength of the association and estimate the increased risk that the risk factors confer, and some of the major studies are mentioned in the previous

chapter. Instead, analyses of the amount of geographic variation in CHD that can be explained by differences in the risk factor profile of the population of different areas are reviewed here. Such analyses can produce surprising results: despite the convincing evidence of a strong association between cholesterol levels and CHD incidence in individuals (Yusuf et al., 2004), differences in the prevalence of raised cholesterol at area-level may explain very little of the variation in CHD incidence rates if prevalence rates of the risk factor are reasonably uniform between different areas. Table 2.1 shows the results taken from a number of studies that estimated the amount of geographic variation in CHD rates that was due to differences in compositional factors.

Table 2.1 Results from the literature regarding the explanation of geographic variation in CHD outcomes by differences in the behavioural risk factor profile of populations

Outcome variable	Behavioural factors as explanatory variables	Further explanatory variables	Population, setting	Study design, n	Reported results	Reference
CHD prevalence rate	Blood pressure, cholesterol, fruit consumption, obesity, physical inactivity, smoking	Age, social class	Women aged 60-79, 23 small towns, Britain, 1999-2001	Cross-sectional (baseline of prospective cohort), n = 4,286	Odds ratio (base South East England), age adjusted only: Midlands and Wales 1.77 (1.40, 2.24); North England 1.27 (1.04, 1.55); Scotland 1.88 (1.47, 2.41) Odds ratio (base South East England), adjusted for all explanatory variables: Midlands and Wales 1.22 (0.91, 1.65); North England 1.14 (0.88, 1.46); Scotland 1.85 (1.37, 2.50)	(Lawlor et al., 2003)
CHD incidence rate	Blood pressure, physical inactivity, smoking	Age, height, social class	Men aged 40-59 at baseline (1978-80), 24 small towns, Britain, 1978-2000	Prospective cohort analysis (22yrs follow-up), n = 7,609	Hazard ratio for Rest of Britain versus Southern England reduced by 42% after adjustment for all independent variables. Adjustment for smoking, blood pressure and age alone reduced the hazard ratio by 29%.	(Morris et al., 2003)
CHD incidence rate	Blood pressure, cholesterol, heavy drinking, obesity, physical inactivity, smoking	Age, height, social class	Men aged 40-59 at baseline (1978-80), 24 small towns, Britain, 1978-1996	Prospective cohort analysis (18yrs follow-up), n = 7,735	Percentage of between towns variance explained by age, smoking, social class, blood pressure, physical inactivity, height and obesity: 79% Percentage of between towns variance explained by age, smoking, social class, blood pressure, physical inactivity, height, obesity, heavy drinking and cholesterol: 50%	(Morris et al., 2001)
CHD mortality rate	Alcohol consumption, smoking	Age, cold climate, H Pylori infection rate, proportion of ethnic minorities, sex	All local authorities in England and Wales, 1992	Cross-sectional ecological study, n = 403	Relative risk for most Northern LAs compared to most Southern LAs, before adjustment for independent variables: 1.46 Relative risk for most Northern LAs compared to most Southern LAs, after adjustment for independent variables: 1.25 ($p<0.05$)	(Law and Morris, 1998)

Table 2.1 (cont.)

Outcome variable	Behavioural factors as explanatory variables	Further explanatory variables	Population, setting	Study design, n	Reported results	Reference
CHD mortality rate	Average serum cholesterol rates	-	Countries where adequate estimates of CHD mortality and cholesterol were available.	Cross-sectional ecological analysis, n = 17	Variation in cholesterol rates explains four fifths of the geographical variation in CHD mortality rates.	(Law and Wald, 1994)
CHD incidence rate	Blood pressure, cholesterol diabetes, obesity, smoking	Age, educational level	White males aged 55-74, USA, 1982-87	Cross-sectional study, n = 1,838	Age-adjusted relative risk for living in non-Western states: 1.39 (1.16, 1.64). Fully adjusted relative risk for living in non-Western states: 1.38 (1.16, 1.64).	(Garg et al., 1992)
Cardiac event rate (CHD mortality or first MI)	Smoking	Age, deprivation	Women aged 45-73, Malmo, 1989-1997	Cross-sectional ecological study (wards as units), n = 17	Percentage of variance in cardiac events rates explained by smoking: 56%.	(Janzon et al., 2007)

The literature reveals a great range in the amount of geographic variation in CHD rates that can be explained by differences in the behavioural risk factor profile of populations, because of differing analytical techniques that have been employed (different study designs, different risk factors included, different confounding factors adjusted for etc.), and differences in the populations that were studied. Some international results have been included in table 2.1 as they reveal the extent to which both large scale and small scale geographic variation in CHD can be explained by compositional factors. For example, Law et al. estimated that four fifths of the variance in CHD mortality rates between countries can be explained by differences in cholesterol levels alone (Law and Wald, 1994). This result was based on a small non-random sample of countries and was not adjusted for potentially confounding factors, but it illustrates the potential power of compositional factors to explain geographic variations in CHD rates. With analysis using much smaller areas (wards rather than countries), nearly six tenths of the variance in female cardiac event rates for wards in Malmo was attributed to differences in smoking rates, even after adjustment for deprivation (Janzon et al., 2007). Yet in contrast to these result, an American study showed that adjustment for smoking, cholesterol, blood pressure, obesity and diabetes had virtually no influence on large scale variation in male CHD incidence rates (Garg et al., 1992). These results suggest that behavioural risk factors could potentially have a large impact on geographic variations in CHD, but that their impact is dependent upon the specific populations under investigation.

There have been two large prospective British cohort studies that have been designed to investigate the causes of geographic variations in CHD. The British Regional Heart Study

(BRHS) collected baseline data in 1978-1980 on men aged between 40 and 59 from 24 similar small towns from around Britain, and the men have been followed since then for incidence of coronary events (detected at regular screenings or confirmed by medical records) and CHD mortality (Walker et al., 2004). The British Women's Heart Health Study (BWHHS) used a similar design to the BRHS, and collected baseline data in 1999-2001 on women aged between 60 and 79 from 23 similar small towns in Britain, and the women have been followed since for similar coronary events (Lawlor et al., 2003). Currently, only cross-sectional analyses from the baseline data collection regarding geographical variation in heart disease have been reported using the BWHHS dataset.

The results from the BRHS show that differences in the prevalence of the behavioural risk factors explain a large amount of the variation in coronary event rates between the 24 towns included in the study. Smoking alone explained 45% of the between-towns variance, and blood pressure levels, physical activity, obesity and alcohol intake each explained between 6% and 27% of the between-towns variance (paradoxically, introducing cholesterol levels into the model increased the between-towns variance by over 30%). When smoking, blood pressure, physical activity and obesity were included in the model (also with social class and height), nearly 80% of the between-towns variance was explained (Morris et al., 2001). A later analysis of the BRHS dataset showed that smoking and blood pressure alone reduced the size of the hazard ratio for coronary events in the North of Britain compared to the South of England by nearly 30%, and further inclusion of physical activity, social class and height reduced the hazard ratio by over 40% (Morris et al., 2003). The BWHHS baseline dataset showed that adjustment for

blood pressure, cholesterol, obesity, physical activity, fruit consumption, smoking and social class explained 50% of the difference in coronary event rates between the South East and North of England, but only 3% of the difference between the South East of England and Scotland.

Most of the studies that considered the amount of geographical variation in CHD that is due to compositional factors used individuals as the units of analysis. However, such studies do not include individuals drawn from all areas in Britain. Comprehensive coverage of the study population can be achieved by ecological studies. One such study on all local authorities in England and Wales used ecological data collected from a variety of data sources including national health surveys and standardised indices, and estimated the impact of smoking, alcohol consumption, *H pylori* infection and socioeconomic deprivation on variation between mortality rates for various causes between the North of England and the South of England (Law and Morris, 1998). For CHD, the increased risk in the North of England was reduced by nearly 50% when the risk factors were adjusted for. However, only smoking and alcohol consumption were included as potentially explanatory behavioural risk factors, and the prevalence estimates for these were drawn from national survey data for regions of England much larger than the local authorities that were used as units of analysis.

Contextual factors

This section reports on the influence of contextual factors on geographic variation in CHD – that is, factors that are a property of the area in which people live, rather than a property of the people who live in the area (also referred to as 'environmental variables' in this thesis). The associations between contextual factors and CHD are less well understood than those between behavioural risk factors and CHD. This is because the risk associated with contextual factors tends to be small (although this small risk impacts upon a large number of people) and is hard to measure because of the large amount of potentially confounding factors, both at the area-level and the individual-level, that exist in studies of different areas. The difficulty in accounting for all the potentially confounding factors is discussed in more detail later.

The review identified estimates of the amount of geographic variation in CHD that can be explained by the climate (specifically sunlight exposure and ambient temperature), air pollution, water hardness, urbanicity (i.e. categorical measures of whether an area is rural, urban, metropolitan etc.), and a range of socio-cultural variables such as perceived safety and neighbourhood aesthetics. The theoretical physiological mechanisms that convey the risk of these contextual factors are summarised below.

Climate

Several possible physiological mechanisms of the effect of cold weather on CHD mortality have been suggested. It is thought that cold weather could increase blood viscosity, thereby increasing the risk of thrombosis, or it could induce a mild

inflammatory response thereby increasing blood coagulability (Toledano et al., 2005). It is known that cold weather tends to increase both blood pressure and blood cholesterol levels, resulting in an indirect increase in cardiovascular risk (Toledano et al., 2005). The increase in blood cholesterol levels has been hypothesised to be a result of low exposure to sunlight, since laboratory studies have shown that sunlight is a catalyst for the synthesis of a precursor for cholesterol (squalene) into vitamin D (Grimes et al., 1996).

The impact of the climate on CHD rates is most clearly demonstrated by the phenomenon of excess winter mortality (Keatinge et al., 1997), where CHD mortality levels have been shown to be nearly 20% higher during the winter months in England and Wales than in the summer months (Allender et al., 2008). But excess winter mortality is a temporal rather than a geographic phenomenon and is therefore not a focus of this review. By its nature this review is concerned with the association of the general climatic conditions of geographical areas with CHD – because the difference in climate between summer and winter months in the UK is more substantial than the difference between regions of the UK this impact is likely to be smaller than is seen for excess winter mortality. Since excess winter mortality does not seem to affect regions of England differentially (Allender et al., 2008), the impact of climate on geographic variations in CHD is likely to be mediated by cumulative exposure over time producing increased CHD risk in individuals living in areas with worse climatic conditions.

Air pollution

Physiological mechanisms that link air pollution and CHD are still under debate. It has been suggested that exposure to air pollution can provoke an inflammatory response, which increases blood coagulability (and hence risk of thrombosis), or that the association between air pollution and lung disease also affects CHD via hypoxia, or possibly that air pollution may affect the autonomic nervous system leading to heart rate variability (Pope, 2005).

Water hardness

It is now generally agreed that if water hardness is protective of CHD then it is because of increased magnesium levels in harder water (Monarca et al., 2004). There are a number of potential mechanisms including suppression of arrhythmias, cardiac irritability and free radical tissue damage (Marx and Neutra, 1997).

Urbanicity

The relationship between urbanicity and health is complicated by other factors that may either be interpreted as confounders of the relationship or to be on the causal pathway. These factors include urban stress, access to health-related resources, air pollution, deprivation and socio-cultural factors. 'Urban stress' refers to the additional pressure on health that is a result of living in close proximity to other people, long working hours and hyper-competitiveness, which can add social pressure resulting in raised stress levels and poor health outcomes (Godfrey and Julien, 2005, Macintyre et al., 2002). Access to health-related resources can be different in urban and rural areas, for example access to

green space - associated with greater physical activity levels - which is more abundant in less urban areas. Other health-related resources with urban / rural gradients include access to healthcare resources (e.g. GP surgeries, hospitals) which tends to be higher in urban areas (Farmer et al., 2006), and access to adequate nutrition which can be low in some inner-city areas, although it has been suggested that this may only have an impact on a small number of urban residents (White et al., 2004). The difference between air pollution levels in urban and rural areas is primarily due to transport in urban areas, but can also be a result of close proximity between residential and industrial areas. Urban areas tend to be more deprived than rural areas (Office for the Deputy Prime Minister, 2004), and large populations living in close proximity in urban areas can result in complex social networks, which can have an additional impact on health (see section on socio-cultural factors below).

Throughout this thesis, the influence of urbanicity on health is interpreted as the influence of urban stress and access to health-related resources, and not the associated influence of air pollution, deprivation and socio-cultural factors.

Socio-cultural factors

A framework of five elements of environmental features that may influence health has been suggested, which includes a domain of *socio-cultural features of a neighbourhood* (Macintyre et al., 2002). A similar framework identifies *individual interaction with specific local cultures* as a potential environmental influence on health (Jones and Duncan, 1995). These domains include the political, ethnic and religious history of the

area, the perceived safety of the area, and perceived levels of community support. These socio-cultural factors arise from complex social cohesion networks, which can influence an individual's behaviour on the basis of the social environment in which they live (Stafford et al., 2003). This influence can be via physical manifestations of perceived behavioural norms, such as the higher than average number of cigarettes smoked by individuals in areas with a high prevalence of smoking (Duncan et al., 1996). It can be via physical manifestations of perceived social insecurity, such as reduced rates of regular exercise in areas perceived to be unsafe (Foster et al., 2004). It can also be via a physical manifestation (e.g. raised stress levels) of a lack of social capital, such as residence in a social environment that is not conducive to social support (Stafford et al., 2003).

Because of the indeterminate nature of these socio-cultural factors, it has proved difficult to measure their association with health outcomes over large geographical regions. One study looked at the association between ward-level all cause mortality rates and a number of socio-cultural factors generated from the 1991 census, including separate measures of affluence, deprivation, rurality, social class and ethnicity (Congdon et al., 1997). The results showed that the effects of the social environment on health cannot be explained purely by a measure of deprivation. For example, both the deprivation of an area and the affluence of an area were found to be independently associated with mortality, suggesting that affluence and deprivation have different associations with health, and are not simply inverse measures of each other. Furthermore, the results suggested that the effect of urbanicity on health is independent of measures of deprivation, and that this association

may be non-linear. These findings, along with the theoretical work outlining the potential influences of the socio-cultural environment on health described above, outlines the importance of considering urbanicity, deprivation and socio-cultural factors as independent yet associated explanatory variables in analyses of health outcomes.

Table 2.2 Results from literature review regarding the explanation of geographic variation in CHD outcomes by contextual factors.

Outcome variable	Environmental factors as explanatory variables	Other explanatory variables	Population, setting	Study design, n	Reported results	Reference
CHD mortality rate	Annual hours of sunshine	Age, sex	Health districts, Britain, 1991	Cross-sectional ecological study, n = 200	Correlation between hours of sunshine and CHD mortality rate: r = -0.59 ($p<0.001$)	(Grimes et al., 1996)
CHD mortality and major non-fatal CHD	Annual hours of sunshine; Mean daily maximum temperature; Mean daily minimum temperature; Total annual rainfall; Water hardness (mineral content mmol/l)	Age, alcohol consumption, blood pressure, BMI, cholesterol, height, physical inactivity, smoking, social class	Men aged 40-59 at baseline (1978-80), 24 small towns, Britain, 1978-1996	Prospective cohort analysis (18yrs follow-up), n = 7,735	Each analysis contains one environmental factor only: Mean max temp of towns explains 30% of between towns variance after individual-level variables have been accounted for. Mean min temp explains 19% of remaining between towns variance. Total sunshine explains 21% of remaining between towns variance. Total rainfall explains 11% of remaining between towns variance. Water hardness explains 9% of remaining between towns variance.	(Morris et al., 2001)
CVD mortality rate	Mean daily maximum temperature; Number of rainy days per year; Water hardness (mineral content mmol/l)	Age, deprivation, sex	Towns, England and Wales, 1969-73	Cross-sectional ecological study, n = 253	All environmental factors mutually adjusted for each other: Increase in number of rainy days by 5% resulted in increase of CVD mortality rate by 4%. Increase in mean max temperature by 0.9°C resulted in decrease of CVD mortality rate by 3%. Increase in water hardness by 1 mmol/l resulted in decrease of CVD mortality rate by 7%.	(Pocock et al., 1980)
CHD mortality rate	Mean daily rainfall Mean daily temperature;	Age, deprivation, sex	County and London boroughs, England and Wales, 1969-71	Cross-sectional ecological study, n = 115	All environmental factors mutually adjusted for each other: Correlation between temperature and CHD mortality rate: r = -0.42 ($p<0.001$). Correlation between rainfall and CHD mortality rate: r = 0.27 ($p<0.01$).	(West and Lowe, 1976)

Table 2.2 (cont.)

Outcome variable	Environmental factors as explanatory variables	Other explanatory variables	Population, setting	Study design, n	Reported results	Reference
CHD mortality rate	Water hardness (magnesium content mmol/l)	Age, CHD mortality gradient, deprivation, sex	Enumeration districts, North West England, 1990-92	Cross-sectional ecological study, n = 13,794	Mortality rate ratio, quadrupling of Mg: 1.01 (0.98, 1.03)	(Maheswaran et al., 1999)
CHD mortality and major non-fatal CHD	Water hardness (mineral content mmol/l)	Age, alcohol consumption, BMI, cholesterol, height, physical activity, smoking, social class,	Men aged 40-59 at baseline (1978-80), 24 small towns, Britain, 1978-2004	Prospective cohort analysis (26yrs follow-up), n = 5,796	Hazard ratio for two fold increase of water hardness: 0.99 (0.94, 1.04)	(Morris et al., 2008)
CHD mortality rates; CHD hospital admission rates	Air pollution (nitrogen oxide; carbon monoxide; particulates)	Age, deprivation, sex, smoking prevalence rate	Enumeration districts, Sheffield, 1994-99	Cross-sectional ecological study, n = 1,030	Nitrogen oxides: Mortality rate ratio, quintile 5 vs. 1 : 1.15 (1.04, 1.27) Hospital admissions rate ratio, quintile 5 vs. 1 : 0.94 (0.85, 1.05) Carbon monoxide: Mortality rate ratio, quintile 5 vs. 1 : 1.07 (0.96, 1.18) Hospital admissions rate ratio, quintile 5 vs. 1 : 0.90 (0.81, 1.00) Particulates: Mortality rate ratio, quintile 5 vs. 1 : 1.07 (0.96, 1.21) Hospital admissions rate ratio, quintile 5 vs. 1 : 1.07 (0.95, 1.20)	(Maheswaran et al., 2005)
CHD mortality	Air pollution (particulates)	Age, alcohol consumption, BMI, diet, education, ethnicity, marital status, occupational exposures, sex, smoking	Adults aged 30+ at baseline (1982), USA, 1988-2000	Prospective cohort analysis (12yrs follow-up), n ~ 320,000	Mortality rate ratio for 10μg/m^3 increment in particulates: 1.18 (1.14, 1.23)	(Pope et al., 2004)

Table 2.2 (cont.)

Outcome variable	Environmental factors as explanatory variables	Other explanatory variables	Population, setting	Study design, n	Reported results	Reference
CHD mortality rate	Urbanicity	Age, ethnicity, sex	Local authorities, England and Wales, 1992	Cross-sectional ecological study, n = 403	Mortality rate ratio, metropolitan versus rural areas: 1.05 (1.03, 1.07)	(Law and Morris, 1998)
CVD mortality	Urbanicity (defined at lower super output area level)	Age, deprivation, sex	All residents, England and Wales, 2002-2004	Cross-sectional study, n ~ 55,000,000	England: Odds ratio, rural vs. urban, men: 0.99 (0.97, 1.01) Odds ratio, rural vs. urban, women: 1.02 (1.00, 1.04) Wales: Odds ratio, rural vs. urban, men: 0.98 (0.92, 1.03) Odds ratio, rural vs. urban, women: 1.00 (0.95, 1.06)	(Gartner et al., 2008)
Coronary artery bypass graft rate (under 75s)	Urbanicity (proximity to cardiothoracic unit)	Age, deprivation	Electoral wards, North East Thames region, 1981-85	Cross-sectional ecological study, n = 576	Men: CABG rate ratio, deprivation quartile 4 vs.1, unadjusted for proximity to cardiothoracic unit: 1.42 (1.16, 1.73) CABG rate ratio, deprivation quartile 4 vs. 1, adjusted for proximity to cardiothoracic unit: 1.15 (0.91, 1.44) Women: CABG rate ratio, deprivation quartile 4 vs.1, unadjusted for proximity to cardiothoracic unit: 2.36 (1.46, 3.81) CABG rate ratio, deprivation quartile 4 vs. 1, adjusted for proximity to cardiothoracic unit: 1.75 (1.03, 2.96)	(Ben-Shlomo and Chaturvedi, 1995)

Table 2.2 shows the results taken from a number of studies that estimated the amount of geographic variation in CHD rates that was due to differences in contextual factors. A common method for analysing the impact of environmental variables on health is to use a time series analysis. Here, an area is selected and the environmental factor and health outcome of interest are measured regularly over a certain period of time (measured every day over the course of a year, for example), and the analysis aims to show whether there is an association between the daily measures of the environmental variable and the health outcome after an appropriate lag period. Such analyses provide evidence of the degree of association between the environmental variable and the health outcome and provide evidence of a cause-effect relationship, but they are unable to produce estimates of the amount of geographic variation in the health outcome that is explained by the environmental variable, since they are restricted to a single area. Because of this, time series analyses have not been included in this review.

The evidence presented in table 2.2 suggests that the climate has a significant impact on geographic variation in CHD rates in Britain. The two ecological studies conducted by Pocock et al. and West and Lowe showed significant correlations between different aspects of the climate and CHD mortality rates even after adjustment for each other and deprivation (Pocock et al., 1980; West and Lowe, 1976), and a stronger correlation between mortality rates and hours of sunshine was shown more recently, although this analysis did not adjust for deprivation (an important confounder given the North-South gradient in both climate and deprivation) (Grimes et al., 1996). The ecological studies were not able to show whether the effects of climate were independent of compositional

effects (other than the compositional differences that are associated with different levels of deprivation), but this is suggested by the results taken from the BRHS where, after adjustment for individual-level risk factors for CHD, the climate variables still explained a small but significant amount of the variation in CHD incidence (Morris et al., 2001).

The impact of air pollution on small scale geographic variations in CHD mortality and hospitalisation rates in a British setting (Sheffield) was estimated by Maheswaran et al., using census enumeration districts as fine analytical units and sophisticated spatial modelling techniques to estimate the underlying level of nitrogen oxide, carbon monoxide and particulates for each small area (Maheswaran et al., 2005). The authors found that levels of nitrogen oxides were significantly associated with increased mortality rates but not hospitalisation rates, and concluded that this may suggest that pollution-related CHD is more likely to result in sudden death, which is supported by the physiological evidence which suggests that air pollution causes thrombosis. Their results are supported by a large American cohort study which found that the underlying particulate air pollution of an area was significantly associated with increased CHD mortality of the residents, after adjustment for individual-level risk factors (Pope et al., 2004).

The evidence regarding the impact of water hardness on geographic variations in CHD is mixed. Both the ecological cross-sectional studies that considered water hardness and were identified by the review estimated the amount of geographic variation in CHD mortality rates that is explained by water hardness after deprivation has been taken into

account, but whereas the earlier of the two studies found that water hardness is significantly negatively associated with CHD mortality rates (Pocock et al., 1980), the later study found no association (Maheswaran et al., 1999). The differences in study design could explain these differing results: the later of the two studies was restricted to the North West of England and used smaller geographical analytical units than the earlier study. Also, the later study adjusted for a geographic gradient in CHD mortality rates, which was designed to remove any underlying geographic trends from the small areas under investigation – this de-trending may have contributed to the non-significant results. However, two analyses of the BRHS cohort that used the same analytical methods also showed mixed evidence of an effect of water hardness on CHD incidence rates. After 18 years of follow-up, water hardness explained a small but significant amount of the geographic variation in CHD incidence that remained after behavioural risk factors had been accounted for (Morris et al., 2001), but this association was no longer apparent after 26 years of follow-up (Morris et al., 2008).

An ecological study comparing CHD mortality rates in England and Wales at different levels of urbanicity found that mortality rates were approximately 5% higher in urban areas than in rural areas (Law et al., 1998), but a more recent study that used much smaller areas to define urbanicity found that any differences in cardiovascular mortality were removed after adjustment for deprivation (Gartner et al., 2008). Meanwhile, studies have suggested that healthcare provision is greater in urban areas than rural areas since they tend to be in closer proximity to healthcare resources (GP services, hospitals etc.). Ben-Shlomo and Chaturvedi showed that the deprivation gradient in coronary artery

bypass graft rates for wards in North East London was substantially attenuated when the proximity of the ward to a cardiothoracic unit was taken into account (Ben-Shlomo and Chaturvedi, 1995). A further examination of revascularisation rates in seven health regions in Britain (covering a population of 11.6 million) found that rates were higher in districts that were closer to cardiology specialist centres, although this analysis was not adjusted for deprivation (Black et al., 1995: not included in table 2.2 as full results are not reported in the original article).

The review did not identify any estimates of the amount of geographic variation in CHD rates that can be explained by socio-cultural factors.

Deprivation

The role of socio-economic status in geographic variation in CHD is discussed in this section. The terminology supporting socio-economic status can be confusing. For the purpose of this thesis, the term 'social class' refers to any measure of socio-economic status that acts at the level of the individual (e.g. occupational social class, income, etc.), and the term 'deprivation' refers to any measure of socio-economic status that acts at the level of an area or population (e.g. aggregated deprivation indices, direct measures of material deprivation of areas). The term 'socio-economic status' is used here as the general term that incorporates both area-level and individual-level notions of wealth, social position, access to resources, etc.

The association between social class and CHD (amongst many other health outcomes) has long been established (Townsend et al., 1986), with higher CHD incidence and mortality in lower social groups – the so-called social gradient. This social gradient has been shown in many studies to be fine-grained: that is, the social gradient is apparent if any subset of the social spectrum is under study. For example, the Whitehall study followed a cohort of men and found that CHD mortality levels were lower for men working in higher employment grades within the civil service (Marmot, 1989), even though the employment grades studied would be classified as relatively high social class when compared to the population of Britain. Findings such as these have been interpreted as evidence that social class has an indirect impact on CHD that is mediated by the behavioural risk factors, since the alternative hypothesis that social class has a direct impact on CHD via material deprivation (e.g. income levels too low to allow access to healthy food and shelter) would produce a distribution of health outcomes with a spike for the very low social classes. Such a distribution has not been demonstrated (Bartley, 2004).

Further evidence of the indirect nature of the association between social class and CHD has been provided by studies that have measured other risk factors in the study population accurately enough to 'explain away' the social class-health association. For example, a study of the association between income and mortality in a cohort of over 2,000 Finnish men found that the social gradient in mortality (the lowest income quintile had a mortality rate three times higher than the highest quintile) was entirely explained by differences in behavioural risk factors (Lynch et al., 1996). Results such as these present

the role of social class in a new light; that of an individual-level component that accounts for the residual variance (i.e. the variance not explained by other explanatory risk factors) in the health outcome because of either missing explanatory variables or measurement error.

Deprivation is usually measured using an index of aggregated social variables for the population within an area. For example, the Carstairs measure of deprivation is an index of the percentage of unemployed men, overcrowded accommodation, low social classes, and individuals with no access to a car within an area (Carstairs and Morris, 1990). There is some debate about whether the deprivation of an area has an impact on health that is independent of the social classes of the population resident in the area. Sloggett and Joshi argued that deprivation has no such independent impact, and used a cohort of nearly 300,000 people followed for nine years to show that area deprivation had no impact on mortality rates after social class had been accounted for (Sloggett and Joshi, 1994). Other researchers have argued that the deprivation of an area reflects both the social classes of the population *and* the material deprivation of the area, even though the measure of deprivation may be derived from purely compositional factors. This material deprivation can have an impact on health via a number of different contextual factors, such as air pollution, access to services or the socio-cultural factors that were discussed earlier (Macintyre et al., 1993). Evidence of this material deprivation effect has been shown by many studies that found both social class and deprivation are independently associated with cardiovascular health outcomes (Lawlor et al., 2005; Davey Smith et al., 1998; Shohaimi et al., 2003).

The relationship of deprivation with CHD is therefore a complex one. The mechanisms for the effect on CHD are indirect, either via increased prevalence of behavioural risk factors (e.g. smoking, obesity) or via environmental variables (e.g. air pollution, access to services, socio-cultural factors). Because deprivation can be considered to be both an aggregated measure of the social classes of the population and also a measure of material deprivation of an area it contains elements that are both compositional and contextual. Because of the indirect nature of the relationship between deprivation and CHD, any associations that are found can be attributed to explanation of residual variance due to missing variables or measurement error. These three roles of deprivation – as a compositional element, contextual element, and indicator of missing variables or measurement error – should be considered when interpreting analyses of the geographic variation of CHD.

It has also been argued that, since socioeconomic status is a relative measure that compares the social position of either individuals or areas, the size of the inequality between the most advantaged and most disadvantaged within a society should also be associated with health outcomes, independently of the other effects of socioeconomic status discussed above. This theory provides one possible explanation of why CHD mortality rates are lower in Cuba than in the USA (Wennemo, 1993) – Cuba has far lower total wealth than USA, but the wealth is distributed more evenly around the country. The theory has been tested by comparing CHD mortality rates amongst industrialised countries. A clear relationship was found between income inequality within the countries

and CHD mortality rates for women (McIsaac and Wilkinson, 1997). If this theory held for areas within the UK then it would be expected that relative deprivation inequalities within regions should be associated with CHD rates after adjustment for the 'absolute' deprivation (measured using a deprivation index) of the areas under investigation. Such a hypothesis has previously been tested using CHD mortality rates for all wards in England as the outcome variable, and two explanatory variables: the deprivation of the ward (measured using an aggregated social deprivation index (Carstairs and Morris, 1990)) and the deprivation inequality of the local authority in which the ward is located (measured using the variance of the deprivation index for all wards within the local authority). The analysis found that relative inequality was only associated very weakly with CHD mortality rates after adjustment for 'absolute' deprivation (Allender et al., forthcoming), suggesting that there is little evidence that relative deprivation inequalities are strongly associated with CHD rates in England.

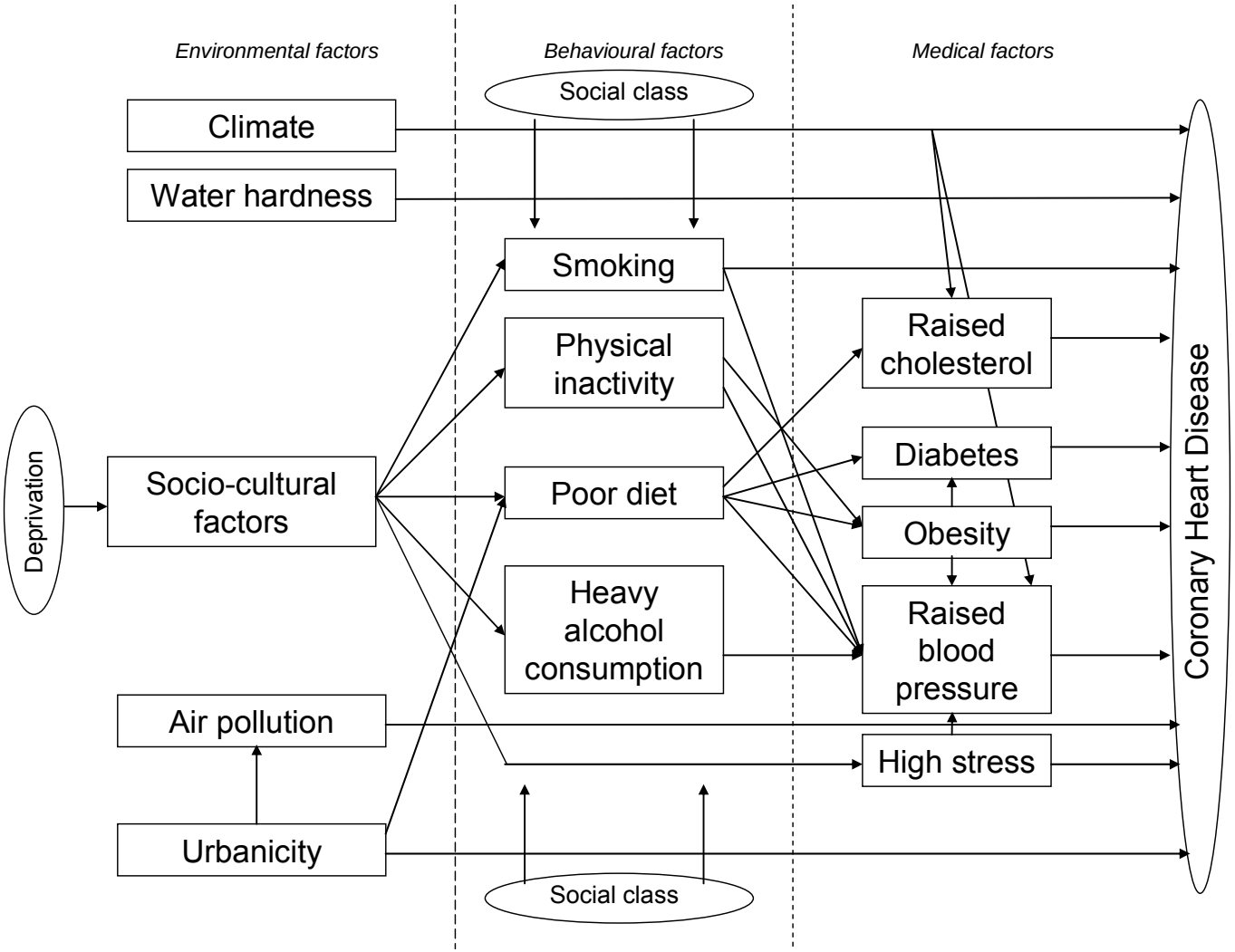
DISCUSSION

The geographic variation of CHD in the UK is a well-studied area that has been explored by a number of different study designs. Most commonly, ecological cross-sectional studies have investigated the association between environmental variables and CHD mortality rates after adjustment for deprivation, but two prospective cohort studies - the British Regional Heart Study and the British Women's Heart Health Study - have been set up in the UK specifically for the study of geographic variation in CHD. These cohort studies allow for investigation of the association between the prevalence of behavioural

risk factors and geographic variation in CHD mortality, incidence and prevalence rates. Less commonly, geographic variations in CHD service utilisation have been investigated (e.g. hospitalisation rates, revascularisation rates) using ecological studies and investigating the influence of environmental variables (urbanicity, air pollution).

A conceptual framework linking the direct and indirect relationships between environmental variables, behavioural risk factors and CHD that have been identified in the literature review is shown in figure 2.1. Here, only modifiable risk factors are included and the arrows in the figure suggest a causal relationship (though do not indicate whether the causal association is positive or negative). For example, the conceptual framework suggests that urbanicity has a direct impact on CHD (since access to health services is greater in urban areas), and an indirect impact via increased air pollution levels and worse provision of shops providing access to healthy food. The conceptual framework is a model of CHD incidence, and as such is concerned with the factors that affect the first presentation of CHD within an individual. For this reason, treatment effects have not been included in the model, although it is acknowledged that some treatments can be applied before CHD has presented within an individual (e.g. prescription of lipid-lowering drugs in those at high risk of CHD; referral to NHS Quit Smoking clinics etc.)

Figure 2.1 Conceptual framework connecting environmental, behavioural and medical risk factors with CHD



Under this conceptual framework, geographic variations in each of the factors included in the model could produce geographic variations in CHD. There is strong evidence provided by prospective cohort studies that variations in the behavioural risk factor profile of populations are responsible for some of the geographic variation in CHD incidence and mortality, and the ecological studies included in the review provide some evidence that the environmental variables are responsible for some geographic variation in CHD mortality and service utilisation rates.

Ecological studies that have investigated the amount of geographic variation in CHD that is explained by contextual factors have relied on measures of deprivation to represent a proxy of the compositional factors that could also influence rates (if compositional factors have been adjusted for at all). This is problematic for two reasons: firstly, although deprivation is associated with prevalence rates of behavioural risk factors it is not a direct measure of such prevalence rates and hence cannot be expected to accurately account for the geographic variation in CHD that is a result of the compositional factors. Secondly, as has been discussed above and is represented in the conceptual framework, deprivation is also associated with environmental risk factors for CHD, therefore adjusting for deprivation may also remove some of the geographic variation that is due to contextual factors.

For different reasons, the prospective cohort studies that have investigated geographic variation in CHD (the British Regional Heart Study and the British Women's Heart and Health Study) are not ideally suited to estimate the amount of geographic variation that is

due to contextual factors. The cohort studies are well-designed to estimate the amount of large scale geographic variation in urban CHD rates that is due to the behavioural risk factor profiles of the populations. However, the powerful multi-level structure of the datasets (7,609 men nested in 24 towns for the BRHS, 7,173 women nested in 23 towns for the BWHHS) is necessarily restricted to only a small number of level-2 units (towns) since increasing the number of areas that were included in the analyses would greatly increase the number of respondents required, and hence the resources required for the study. This restriction of the level-2 units has three consequences: firstly, there is little scope to include more than one potentially explanatory environmental variable in any single analysis, as there is not sufficient statistical power at the area-level. Secondly, because the level-2 units are drawn from disparate areas around Britain there is no opportunity to study the impact of environmental variables on small scale geographic variation in CHD rates. Thirdly, in order to control for potentially explanatory environmental factors each of the level-2 units were selected to be reasonably similar small towns with similar populations, thereby limiting the potential to extrapolate the results to the entire UK population.

Ideally, a study aimed at estimating the amount of geographic variation in CHD that can be explained by contextual factors must clearly state whether this includes indirect causes of CHD (such as all of the impact of socio-cultural factors and the impact of urbanicity via inadequate access to shops providing healthy foods) or is restricted to direct contextual causes (climate, water hardness, air pollution and access to health services) as this distinction will inform the design of the analysis. A comprehensive analysis of the

impact of contextual factors should include all of the variables included in the conceptual framework. The number of potentially explanatory environmental variables implies that a study must include a large number of area-level units in order for the analyses to be adequately statistically powered. Ecological analyses struggle to find data sources for the compositional factors for small areas, whereas individual-level studies struggle to provide adequate power for area-level analyses. Because of these problems, a comprehensive study of the amount of geographic variation in CHD in the UK that can be explained by environmental variables has yet to be conducted. Accordingly, the analyses reported in this thesis attempt to fill this knowledge gap by examining geographic variations in CHD using ecological data modelled for all wards in England that include both environmental data and the behavioural risk factor profiles of populations. The data sources and modelling techniques utilised for these ecological analyses are explained in detail in the following chapter.

INTRODUCTION

This chapter describes the design of the analyses conducted for the thesis. The research questions are stated initially, and then the analytical methods, setting, population and units of analysis are described. A comprehensive consideration of the data used for this thesis is then reported. A description of the statistical methods used in this thesis is provided in the following chapter.

RESEARCH QUESTIONS

RQ1: How much of the large scale and small scale geographic variation in coronary heart disease (CHD) *mortality* rates in England is explained by the direct influence of environmental variables, and how much by the behavioural risk factor profile of populations?

RQ2: Similarly, how much of the large scale and small scale geographic variation in CHD *hospitalisation* rates in England is explained by the direct influence of environmental variables, and how much by the behavioural risk factor profile of populations?

For the purpose of this thesis, the *direct influence of environmental variables* refers to the impact of climate, water hardness, air pollution and urbanicity on CHD that is not

mediated by behavioural risk factors (see figure 2.1, chapter two). *Behavioural risk factor profile of populations* refers to the prevalence of both behavioural and medical risk factors for CHD that are present in a population (e.g. the prevalence of smoking, high salt consumption, but also raised cholesterol, obesity etc.). The conceptual framework shown in figure 2.1 is used as a basis for interpretation of the results, and as such only environmental variables that are displayed in the model were included in the analyses. These include all environmental variables with an association with CHD that were identified in a review of the literature (reported in the previous chapter).

STUDY DESIGN

The findings of this thesis are based on a series of ecological cross-sectional analyses of geographic variation in CHD mortality and hospitalisation rates. Multi-level models were built to explore both small scale and large scale variation simultaneously, and spatial error regression models were built to assess whether the analyses were prone to spatial auto-correlation bias. An outline of the issues that informed the choice of study design is provided below.

Population and setting

The population of interest for this study is the entire adult population of England, and the setting for the study is the calendar year of 2001. This choice of population and setting was dependent upon the available data for analysis. The setting of 2001 is because of the

comprehensive small area demographic data that were required for the analyses that are only available from the UK census. The restriction of the setting to England is because the data on prevalence rates for behavioural risk factors were derived from the Health Survey for England series (Department of Health, 2003). Comparable data sources for Scotland, Wales and Northern Ireland for 2001 were not available.

Units of analysis

The units of analysis used in this thesis were standard table wards. These are administrative units that are used by various providers of statistics to produce data on population, mortality and hospitalisation rates, amongst others. Unlike electoral wards which are subject to periodic boundary changes, the boundaries of standard table wards have been fixed – they are based on the electoral ward boundaries at 1 January 2003. In general in this thesis, standard table wards are referred to simply as 'wards'.

Wards are one part of an administrative structure in England that has been designed so that the lower-level areas can be used to build all of the higher-level areas. This structure consists of (from smallest to largest populations) output areas, super output areas, wards, local authorities and government office regions. The lowest building blocks of this structure – output areas – were introduced for 2001 census data in England and Wales and were designed to have similar population sizes and to be as socially homogeneous as possible. These advantages were retained somewhat by the super output area structure, which was designed for the provision of health, economic and social statistics using a

structure that is not dependent upon boundaries drawn for administrative purposes. The boundaries for wards were designed for electoral purposes. Although the intention of the electoral boundary system is to identify areas of reasonably the same size and population and to identify areas which residents can identify with, the heterogeneity of population sizes at the ward level is evident. The smallest wards in England contain approximately 1,000 residents, whereas the largest wards can have populations larger than 30,000 (Office for National Statistics, 2009). Details about this structure are provided in table 3.1.

Table 3.1. Population details by administrative structure, England, c.2001

Structure	Number of units	Population Range	Population Mean	Population Median
Government office regions	9	2.55M – 8.11M	5.57M	5.33M
Local authorities	355	2,200 – 992,400	141,100	116,000
Standard table wards	7,932	1,000 – 36,000	6,000	4,800
Super output areas	~32,600	750 – 7,000	1,500	1,500
Output areas	~175,400	~100 – ~1,000	300	300

Ideally, ecological analyses should use the smallest area measure available, since the smaller the area, the more homogeneous the population living in the area is likely to be in terms of social class, ethnicity and other factors that could potentially affect the associations under analysis. Further, using smaller areas provides a greater number of units of analysis and hence greater statistical power to detect associations, provided that spatial auto-correlation is adequately accounted for (Elliott and Wakefield, 2000).

However, neither output areas nor super output areas have been used as the analytical units for this thesis because data for many of the explanatory variables are not available for areas at this level, and the populations are so small that few outcome events are recorded each year when further stratified by age and sex.

The nesting of wards in local authorities means that the spatial structure of the data used in this thesis is well-defined and lends itself to multi-level analysis. This allows for identification of small scale geographic variation in CHD rates (modelled by the average variance between wards in a local authority) and large scale variance in CHD rates (modelled by the variance between all local authorities in England).

Controlling for confounding

CHD incidence and mortality increase dramatically with age (Allender et al., 2008), therefore the age structure of an area must be taken into account when calculating CHD mortality and hospitalisation rates. All analyses in this thesis are adjusted for age using the direct method by applying age-specific rates in each study population to the age structure of the European Standard Population (West Midlands Public Health Observatory, 2009). Similarly, men and women experience CHD differently with higher age-specific incidence and mortality rates in men (Allender et al., 2008). To account for these differences, all analyses were stratified by sex, and the results are presented for men and women separately. Further non-modifiable individual-level risk factors for CHD include ethnicity and genetic disposition. These have not been controlled for since

comprehensive mortality and hospitalisation data stratified by genetic disposition are not available, and stratification by ethnicity is only partially possible with routinely collected data sources – mortality data can be provided by 'country of birth' which is increasingly less related to ethnicity, and although *Hospital Episode Statistics* collect data on ethnicity, there is missing data for around 35% of all hospital admissions (Hospital Episode Statistics, 2004).

The conceptual framework displayed in figure 2.1 was used as a basis for the analyses. Under this interpretation, some behavioural risk factors are on the causal chain between environmental variables and CHD, such as raised blood pressure on the causal chain between cold temperature and CHD. Some behavioural risk factors potentially confound the relationship between environmental variables and CHD, such as heavy alcohol consumption, which is associated with both climate - binge drinking is more prevalent in the North of England (Allender et al., 2008) - and CHD. Also, socioeconomic status is a potential confounding factor to the relationship between CHD and environmental variables *and* between CHD and the behavioural risk factor profile of populations. In order to properly account for these relationships and to be able to assess the independent role of environmental and behavioural variables in the geographic variation of CHD, the following analytical structure was followed. Initially, only environmental variables were included as potentially explanatory variables in the analysis of CHD rates, and the environmental variables that showed independent associations with CHD were retained. Then only behavioural risk factor profiles of populations (prevalence rates of smoking or raised cholesterol, for example) were included in the models, and only variables that

showed independent associations were retained. Finally a series of models were built that included all possible combinations of the three following sets of variables: environmental variables, behavioural risk factor profiles of populations, and deprivation. The results from these series of models allowed for an estimation of the effect of including or excluding any one set of variables, and hence provided evidence of the role of these variables on geographic variation in CHD that is independent of all other explanatory variables.

Data

Mortality

Mortality data were provided by the Office for National Statistics for the years 1999 to 2004 (inclusive) stratified by sex, ward and five year age group. The mortality data included all deaths in England where CHD was recorded as the primary cause of death (for 1999 and 2000, ICD codes 410-414; for 2001-2004, ICD codes I20-25). Rates were constructed using mid-2001 population data stratified by sex, ward and five year age group, collected for the 2001 UK census. Change in ICD coding over the data collection period is thought to have had little impact on reporting of CHD mortalities (Griffiths et al., 2004).

The mortality data collection period was greater than the calendar year of 2001 because the annual number of CHD deaths in each ward is low. In 2001, 240 wards did not register any male CHD deaths and 559 did not register any female CHD deaths, and

estimates of mortality rates based on one year of data are therefore unlikely to reflect the underlying mortality rate in each ward. An exercise was conducted to estimate the data collection period that would be needed to generate reasonable mortality rate estimates. The purpose of this exercise was to determine the accuracy of ward-level CHD mortality rate estimates when the data period covers first one year, then two years, and so on up to ten years. 'Accuracy' here is measured by the distance of the calculated CHD mortality rate from the (prescribed and hence known) underlying rate for the ward. The prescribed underlying rate used for the wards was set as the 2001 CHD mortality rates for England and Wales. The age and sex stratified rates from this prescribed underlying rate were applied to the population of all wards within England and Wales (n = 8,839), and the resultant number of expected deaths in each age, sex and ward stratum was rounded to the nearest whole number. These expected deaths were then used to calculate the age-standardised mortality rate for each ward, and the difference between this rate and the prescribed underlying rate was calculated. The results are shown in table 3.2.

Table 3.2 Estimated percentage of wards with CHD mortality rate estimates within 5%, 10% and 25% of the 'underlying' CHD mortality rate by years of data collection (n = 7,929)

Years of data:	*1*	*2*	*3*	*4*	*5*	***6***	*7*	*8*	*9*	*10*
Male mortality rate										
Within 25%	65	90	97	98	99	**100**	100	100	100	100
Within 10%	34	64	80	88	92	**95**	97	97	98	98
Within 5%	18	40	56	65	73	**79**	84	86	88	90
Female mortality rate										
Within 25%	50	78	89	94	97	**98**	99	99	99	100
Within 10%	15	37	54	65	75	**81**	85	89	92	93
Within 5%	6	18	28	37	46	**52**	58	64	71	73

A six year data collection period was selected on the basis of these results, since at least 80% of the calculated ward-level mortality rates for both men and women should be within 10% of the underlying rate for the ward. A longer data collection period was rejected in order that the data would closely reflect the setting of the analysis. A similar investigation showed that premature (under 75) mortality rates for wards were poor indicators of the underlying rate even with a six year data collection period, and therefore premature mortality was rejected as a potential outcome for this thesis.

Hospital admissions

Hospitalisation data were obtained from Hospital Episode Statistics for the financial years 1998/99 to 2003/04 inclusive stratified by sex, ward and broad age band (0-4, 5-14, 15-44, 45-64, 65-74, 75-84, 85+). A 'hospitalisation' was defined as a finished consultant episode (a period of admitted patient care under one consultant within one healthcare provider) where the principal diagnosis was CHD (for 1999 and 2000, ICD codes 410-414; for 2001-2004, ICD codes I20-25). A six year data collection period was used for comparison with the mortality data; since there were a greater number of hospitalisation events than mortalities, this should allow for reasonable estimates of the underlying ward-level hospitalisation rates.

Climate

Climate data were provided by the Meteorological Office for 37 English weather stations that measured temperature, sunlight and rainfall between 2000 and 2002. For each calendar month in this time period data were provided on the following variables: mean

daily maximum temperature, mean daily minimum temperature, mean daily hours of sunlight, and mean daily rainfall.

The 37 weather stations were distributed around England so that 97% of the wards were within 50km of a weather station. The data were used to generate model-based ward-level estimates of the four different climate variables for each month between 2000 and 2002 using second order trend surface modelling (Cressie, 2000). The trend surface modelling methods are described in the following chapter. Similar versions of trend surface modelling have previously been used to produce climate estimates using a small number of data points, including estimates of the amount of UV radiation in different regions of the United States of America (Schwartz and Hanchette, 2006). The modelled monthly estimates were then combined to produce aggregated estimates of the four climate variables for the period 2000-2002. Two examples (mean daily maximum temperature and annual rainfall) are shown in figures 3.1a and 3.1b, where all wards in England have been split into quartiles by the climate variables. These figures show how the trend surface modelling of the climate variables results in smooth climate gradients across England, and that the different climate variables follow unique geographical patterns.

Figure 3.1a Mean daily maximum temperature, 2000-2002, England (wards, n = 7,929).

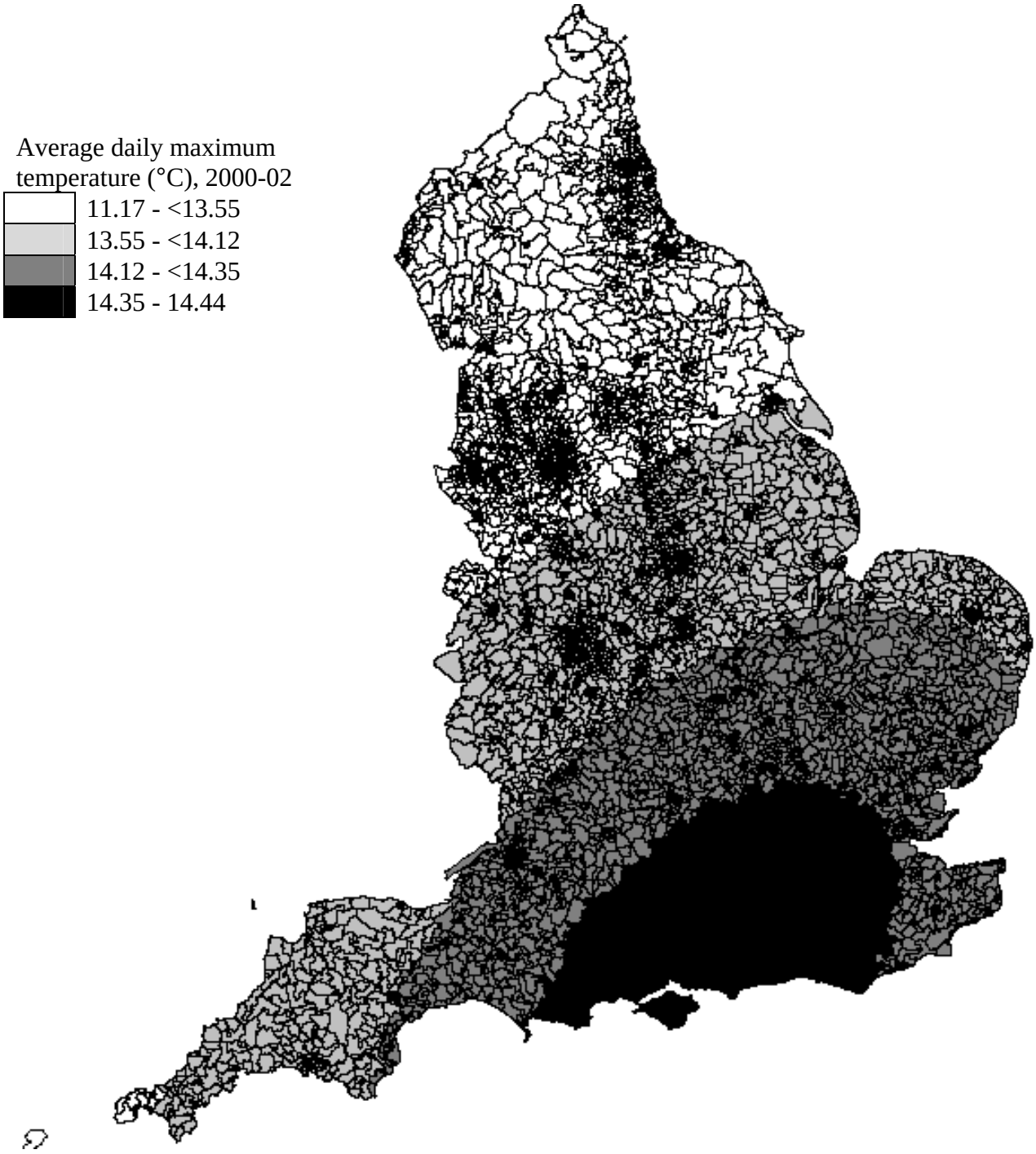
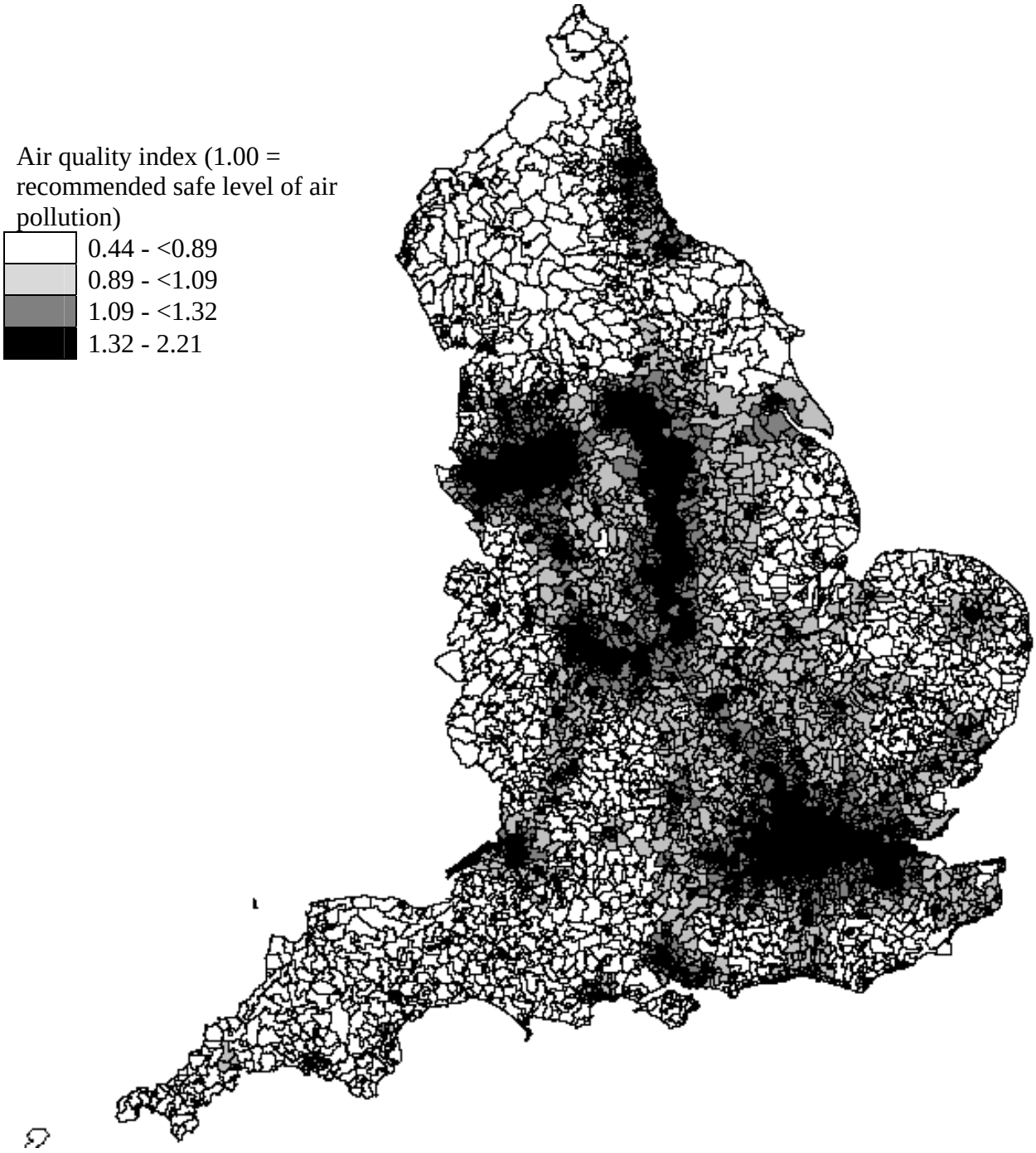


Figure 3.1b Total annual rainfall, 2000-2002, England (wards, n = 7,929).

Air pollution

Air pollution data were collected in 2001 for the development of the physical environment domain of the Index of Multiple Deprivation 2004 (Office of the Deputy Prime Minister, 2004). The data were drawn from the National Atmospheric Emissions Inventory which estimated annual mean concentrations of benzene, nitrogen dioxide, sulphur dioxide and particulates for all 1km grid scores within the United Kingdom, using data on location of roads, housing, agriculture and point sources of emissions (e.g. power stations) and estimated emissions for these sources (Bush et al., 2008). These data were used to model estimated annual mean concentrations for each super output area in England. In addition, a single measure – the air quality index – was constructed that is a standardised index of levels of the four pollutants with comparison to recognised safe levels (Neighbourhood Statistics, 2007). The air quality index was the variable included in the Index of Multiple Deprivation 2004, and was aggregated to ward level by producing averages of the super output area estimates, weighted by population. The geographic pattern of the air pollution data is shown in figure 3.2 where all wards have been split into quartiles. Figure 3.2 demonstrates how the larger number of data points used for the trend surface modelling (in comparison to the modelling of the climate variables) has allowed for estimates of air pollution to vary drastically over short geographical distances.

Figure 3.2 Air quality index, 2001, England (wards, n = 7,929).



Urbanicity

The urbanicity variable used in this thesis was a categorisation of all wards into one of three groups: coastal and countryside, urban, and metropolitan. This categorisation was based on the Office for National Statistics area classification variable, which categorises all wards in the United Kingdom into nine supergroups, 17 groups and 27 subgroups, based on a cluster analysis on demographic structure, household composition, housing, socio-economic status, employment, and industry (Office for National Statistics, 2001). The categorisation of English wards into the nine supergroups is displayed in table 3.3.

Table 3.3 Categorisation of English wards (n = 7,932) by the Office for National Statistics (ONS) area classification variable

ONS area classification	Wards (%)	Population (%)
Industrial hinterlands	1,211 (15)	9.46M (19)
Traditional manufacturing	524 (7)	4.69M (9)
Built up areas	163 (2)	0.95M (2)
Prospering metropolitan	169 (2)	1.86M (4)
Student communities	306 (4)	2.64M (5)
Multicultural metropolitan	318 (4)	4.01M (8)
Suburbs and small towns	2,504 (32)	14.90M (30)
Coastal and countryside	1,838 (23)	8.14M (16)
Accessible countryside	899 (11)	2.79M (6)

The three categories used in the thesis consisted of the following supergroups: coastal & country (comprising 'accessible countryside' and 'coastal and countryside'); urban (comprising 'industrial hinterlands', 'traditional manufacturing', 'built up areas', 'student communities' and 'suburbs and small towns'); and metropolitan (comprising 'prospering metropolitan' and 'multicultural metropolitan'). The supergroups were collapsed into

these three categories in order to ease interpretation of the analyses, and to increase the statistical power to detect differences by urbanicity.

Water hardness

Water hardness was not included as an explanatory variable in this thesis as it proved impossible to locate a source of water hardness data that could produce estimates of local water hardness for all wards in England. The possibility of using trend surface modelling to produce modelled estimates of water hardness from a sample of data points was rejected, as water hardness is dependent upon water supply sources and therefore does not vary smoothly around England – a prerequisite for trend surface modelling (see chapter four).

Synthetic estimates of the prevalence of behavioural risk factors for CHD

The ecological study design of this thesis requires estimates of the prevalence of behavioural risk factors for CHD for each ward in England. Classically, such estimates would be made by means of a survey of a representative sample of the population of interest. However, collecting enough survey data to produce reliable prevalence estimates for all wards in England would be prohibitively expensive. Most national surveys are only designed to produce prevalence estimates at the government office region level (e.g. Health Survey for England (Department of Health, 2003), General Household Survey (Office for National Statistics, 2006a)), and those that can be used to produce estimates at ward level tend to be based on questionnaire mail outs or telephone surveys that produce poor response rates and hence dubious estimates (e.g. the Active People Survey 2005/06

which achieved a response rate of 21% (Purslow et al., 2007)). Because of these limitations, the individual-level variables used in this thesis are derived from a model-based estimation process known as synthetic estimation.

Synthetic estimation involves using data from national surveys to develop a multi-level logistic regression model that estimates an individual's probability of having a risk factor on the basis of potential explanatory variables taken both from the national survey and from census data attached to the ward of residence of the individual. This model is then applied to census data to estimate the prevalence rate of the risk factor for every ward in England. For example, each respondent to the Health Survey for England provided their postcode, which can be used to derive the ward of residence. Census data regarding the characteristics of this ward (e.g. percentage of residents living in rented accommodation, percentage of non-white residents etc.) are then added to the Health Survey for England dataset, and a multi-level logistic model is built that estimates the probability of an individual smoking on the basis of individual-level variables such as age, sex and ethnicity and also ward-level variables taken from the census. The resultant model is applied to the number of people in each age, sex and ethnicity group for all wards in England to estimate the number of people who smoke (and hence the prevalence of smoking) for all wards.

The synthetic estimates that were used in this thesis are shown in table 3.4. All of the sets of synthetic estimates used in this thesis were previously developed by researchers with an interest in the field of health geography (citations for each set of synthetic estimates

are given in table 3.4). The actual synthetic estimates used for the analyses reported in this thesis were generated using the parameter estimates for the models that were published in the supporting literature. These models were applied to data from the 2001 census, downloaded from the MIMAS centre at the University of Manchester (MIMAS, 2006). This process involved isolating the exact census variables of interest (the independent variables in the published synthetic estimation models) and developing complicated syntax to apply the individual synthetic estimation models to all wards in England. The manipulation of data and the application of the models to the data to generate the synthetic estimates were subject to numerous reliability and validity checks to ensure accuracy.

The logistic models for each of these synthetic estimates include both age and sex as individual-level explanatory variables, which allows for the development of age-standardised gender-stratified synthetic estimates of the prevalence of each risk factor. This is essential since the outcome variables are also age-standardised and sex-stratified. It was not possible to include the synthetic estimates developed by the National Centre for Social Research (Pickering et al., 2004) and utilised by the Neighbourhood Statistics website because the accompanying logistic regression models do not include age and sex as explanatory variables. Synthetic estimates of obesity that were recently developed (Moon et al., 2007) were also not included as it was not possible to obtain full details of the accompanying logistic regression model.

Table 3.4 Sets of synthetic estimates used in thesis

Synthetic estimate of percentage of population…	Reference
eating less than 5 portions of fruit and vegetables per day, 2001 model	(Dibben et al., 2004)
eating less than 5 portions of fruit and vegetables per day, 2003 model	(Dibben et al., 2004)
doing under five hours of physical activity in a week, 2001 model	(Dibben et al., 2004)
doing under five hours of physical activity in a week, 2003 model	(Dibben et al., 2004)
consuming greater than weekly recommended average intake of alcohol	(Twigg et al., 2000)
who are current smokers, 2000 model	(Twigg et al., 2000)
who are current smokers, 2004 model	(Twigg et al., 2004)
who are current smokers, 2001 model	(Dibben et al., 2004)
who are current smokers, 2003 model	(Dibben et al., 2004)
with BMI>=30kg/m^2, 2001 model	(Dibben et al., 2004)
with BMI>=30kg/m^2, 2003 model	(Dibben et al., 2004)
with SBP>=160mmHg or DBP>=95mmHg, 2001 model	(Dibben et al., 2004)
with SBP>=160mmHg or DBP>=95mmHg, 2003 model	(Dibben et al., 2004)
with total cholesterol >=6.5mmol/l, 2001 model	(Dibben et al., 2004)
with total cholesterol >=6.5mmol/l, 2003 model	(Dibben et al., 2004)
with diagnosed and undiagnosed type 1 and type 2 diabetes	(YHPHO, 2005)

BMI = Body Mass Index, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure. The Health Poverty Index synthetic estimates (Dibben et al., 2004) were originally developed in 2001 and then updated in 2003 (hence the two models for each risk factor).

Deprivation

The deprivation variable used for this thesis was the Carstairs index (Carstairs and Morris, 1990), generated using data from the 2001 census at ward level (Morgan and Baker, 2006). The index is a sum of the z scores[1] of the following census variables:

- Unemployed males aged 16 and over as a proportion of all economically active males aged 16 and over *(unemployment)*;
- Persons in households with one or more persons per room as a proportion of all residents in households *(overcrowding)*;
- Residents in households with no car as a proportion of all residents in households *(car ownership)*;

[1] (Measure – mean) / standard deviation

- Residents in households with an economically active head of household in social class IV or V, approximated from the National Statistics socio-economic classification (NS-SEC) as a proportion of all residents in households *(low social class)*. (Morgan and Baker, 2006).

The choice of which deprivation index to use for this study was largely pragmatic. The Carstairs index is applicable to any administrative structure used to disseminate census outputs, which includes wards. This is not the case for the Index of Multiple Deprivation developed in 2000 (Department of the Environment, Transport and the Regions, 2000). The Index of Multiple Deprivation developed in 2004 (Office of the Deputy Prime Minister, 2004) can be applied to wards, but it is unsuitable for these analyses because elements of the index are either included as potentially explanatory variables for this thesis (the air pollution data) or are strongly correlated with the outcome data (years of potential life lost). The Carstairs index has been shown to be strongly positively associated with cardiovascular disease rates in England (Romeri, 2006). Therefore, it was felt that the index should be an appropriate measure of the residual variance unexplained by the environmental and behavioural risk factors because of missing variables or measurement error, as discussed in the previous chapter.

Data comparability issues

The setting for this thesis is England in the calendar year 2001, but the data collection periods for the different data sources vary from variable to variable. For example, the

CHD outcome data collection period spans six years from 1999 to 2004, whereas the climate data collection period is three years from 2000 to 2002. The synthetic estimates used in this thesis are set in 2001 since they are based on models that use the 2001 census data for input, but the national surveys used to derive the various accompanying logistic models are drawn from the period 1995 to 2002. The different data collection periods introduce a data comparability problem, but this problem is unlikely to introduce any substantial bias since the phenomena under investigation were reasonably stable in the years surrounding 2001. CHD mortality rates are falling, but there was not a fundamental shift in the trend in CHD mortality rates between 1999 and 2004, for example. Similar observations could be made about the environmental phenomena and the individual-level risk factors under investigation.

The explanatory variables that were used in this thesis are all model-based estimates that rely on trend surface analysis, cluster analysis or synthetic estimation (a full explanation of these techniques is given in chapter four). As with all model-based systems, the resultant estimates are accompanied with some error. The result of inaccuracies in explanatory variables in regression analyses is a bias towards the null hypothesis of no association with the outcome variable, provided that the inaccuracies are randomly distributed in the explanatory variable (as is assumed to be the case with model-based error). Therefore the analyses reported in this thesis are likely to be biased towards the null hypothesis. However, the degree of this bias is dependent upon the size of the errors in the explanatory variables, which is not equal across all variables. For example, the synthetic estimation process tends to produce estimates with large confidence intervals at

ward-level (Heady et al., 2003), whereas trend surface regression tends to produce reasonably accurate estimates provided that the phenomenon being modelled varies reasonably smoothly over the trend surface, as is the case for the climate variables (e.g. a trend surface model that predicted UV radiation accounted for 96% of the variance in measures from the UV detectors (Scwartz and Hanchette, 2006)). This differential in model-based error between the explanatory variables should be considered upon interpretation of the analyses reported in this thesis.

INTRODUCTION

This chapter explains the statistical techniques that were used for the analytical chapters of the thesis. The chapter is intended to give a brief overview of the techniques, with reference to the datasets described in the previous chapter. Appendix one provides the histograms and details about the outcome variables that were used to determine the analytical techniques. The chapter begins with a discussion of the distribution of the outcome variables, and then five techniques are described: multi-level regression analysis, spatial error regression analysis, synthetic estimation, trend surface modelling and principal components analysis. The first two of these techniques were used for the analyses reported in chapters five, seven and eight to estimate the impact of environmental variables and behavioural risk factor profiles of populations on CHD rates, and a brief guide to interpretation of the results tables in those chapters is provided here. Synthetic estimation is the process that has been used to estimate the prevalence of behavioural risk factors for CHD at ward-level. Trend surface modelling has been applied to both the climate and air pollution data to produce the explanatory variables used in this thesis. Principal components analysis was used for the analyses reported in chapter seven to account for collinearities between the sets of synthetic estimates.

Distribution of outcome variables

The main analyses that are described in this thesis are multi-level regression analyses and spatial error regression analyses, both of which are derived from the theory, assumptions and estimation processes of standard linear regression. The model supporting standard linear regression is shown in equation 4.1 below, where y is the outcome variable, x_1, …, x_n are the explanatory variables, α, β_1, …, β_n are regression parameters and ε is the random error term.

$$(4.1)\quad y = \alpha + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon$$

Providing that y is a continuous normally distributed variable there are various algorithms that will estimate values for the regression parameters on the basis of the values of the outcome and explanatory variables, so that the sum of the squared error terms is as small as possible (i.e. the parameters describe a 'best fit line' through the available data). If the assumptions that the estimation process is based on are largely sound, then the error terms will be normally distributed with a mean of zero (Altman, 1991).

Standard linear regression analysis relies on a continuous normally distributed outcome variable. The outcome variables used in this thesis are counts of CHD events (either deaths or hospitalisations) within wards. These CHD events are rare (within the population as a whole) and occurrences are independent of each other (i.e. an event in one individual is not associated with events in other people): therefore they are likely to follow the Poisson distribution (Altman, 1991). If this were the case, then the most appropriate analytical technique would be Poisson regression analysis, where the outcome is transformed to follow a normal distribution and regression analysis is

conducted that accounts for the differing population size of the wards. Figure 4.1 in appendix one displays the distribution of the CHD events amongst wards in England, and clearly shows that the outcome variables do *not* follow a Poisson distribution. This is most clearly seen by referring to the mean and variance of the outcome variables - a Poisson distribution has equal mean and variance, whereas the variances of the four outcome variables (counts of male and female mortalities and hospitalisations) are far higher than the means. In such cases the variables are described as *overdispersed;* Poisson regression analysis is not appropriate in such cases since the variables can not be transformed to follow a normal distribution. This overdispersion was probably because the expected number of events for each ward was proportional to the population size of the ward, which was not constant.

Instead of using Poisson regression analysis, the count data were converted into rates for each ward. The advantage of this approach is that the count data are converted into a continuous variable that, provided it is reasonably normally distributed, can be used in standard linear regression. The disadvantage is that the differing population sizes (and hence the accuracy of the rate estimates) are ignored by such a process. Figure 4.2 in appendix one shows the histograms of the CHD mortality and hospitalisation rates after direct age-standardisation to the European Standard Population to account for differing age structures. The male and female mortality rate variables follow close to a normal distribution, albeit with some positive skew. The hospitalisation variables show much more positive skew and deviate further from the normal distribution (for example, the

kurtosis of a perfect normal distribution is three, whereas the kurtosis of the hospitalisation rate distributions is over six).

The possibility of transforming these variables to improve the normality of the distributions was explored. Log transformation of the CHD rates improved the normality of the hospitalisation variables but had little impact on the mortality variables. In order for the hospitalisation and mortality models to be directly comparable, it was decided to use the non-transformed outcome variables in the regression analyses described in this thesis. This should allow for easier interpretation of the results and is unlikely to result in poorly fitted models since regression analysis is fairly robust to deviations from normality of the outcome variable provided the dataset is large enough, as is the case here (Vittinghoff et al., 2005).

Multi-level regression analysis

Multi-level regression analysis deviates from standard linear regression by allowing the random error element of the regression model to vary at more than one analytical level. The advantage of using a multi-level model for this thesis is that it allows for the spatial structure of the ecological dataset to be appropriately modelled and it allows for detection of whether the fitted models are successful at explaining 'large scale' geographic variation in CHD (modelled by residual variance between all local authorities in England) or 'small scale' geographic variation in CHD (modelled by residual variance between wards in a local authority) or both.

Extending the standard linear regression model to a two-level structure (wards nested in local authorities) involves introducing a second error term. A similar process of parameter estimation is then performed using equation 4.2 below, where j is used to denote wards, k is used to denote local authorities, u_{jk} is an error term that varies at ward level, and v_k is an error term that can only vary at local authority level.

$$(4.2)\quad y = \alpha + \beta_1 x_{1jk} + \ldots + \beta_n x_{njk} + u_{jk} + v_k$$

Such a model is described as a 'random intercept' model, because the introduction of the local authority level error term has the effect of changing the α parameter for each local authority, but the slope parameters (β_1, …, β_n) remain constant. As is the case with standard linear regression, if the model is appropriately fitted then both the ward and local authority error terms will be normally distributed with mean of zero. The distributions of the error terms produced by the final multi-level model reported in table 8.3 in chapter eight are shown in figures 4.3 and 4.4 in appendix one. They are reasonably normal and have a mean of zero, suggesting that the multi-level models built for this thesis were well-fitted and did not violate crucial assumptions underpinning linear regression modelling.

Multi-level regression analyses using models derived from equation 4.2 with age-standardised CHD mortality and hospitalisation rates as the outcome variables are reported in chapters five, seven and eight. The tables reporting the results provide estimates of the α parameter and each of the β parameters (plus associated standard errors and p values), and a calculation of the amount of ward-level and local authority-level

variance that has been explained by the model. The α parameter is the intercept – that is, it is the CHD rate that would be predicted if the values for all of the explanatory variables were zero. Since many of the explanatory variables could not practically take a value of zero (it would be unpleasant to live in a ward that had zero annual hours of sunshine), the value of α is mostly meaningless but is provided for completeness. The values of the β parameters are the estimate of the change in CHD rate that would be expected for a unit change in the explanatory variable after adjustment for all other explanatory variables. Therefore a positive β value indicates a positive association between the variable and CHD rates and vice versa. Because of the different units used to measure the explanatory variables, comparison of the size of the β parameters is not advised – for example, the β parameter for rainfall estimates the change in CHD rates that would be expected for a 1m change in annual rainfall, whereas the β parameter for smoking estimates the change in CHD rates associated with a 1% increase in the prevalence of smoking. There is no reason to suggest that these two changes in the explanatory variables are in any sense equivalent.

The amount of ward-level and local authority-level variance that is explained by the model is calculated as ((baseline variance – model variance) / baseline variance) x 100% where the baseline variance refers to the amount of ward-level or local authority-level variance in the error terms in the model with no explanatory variables and the model variance refers to the amount of ward-level or local authority-level variance in the error terms in the investigated model. For example, table 8.3 (chapter eight) reports that the investigated female mortality model explains 17% of the ward-level variance and 76% of

the local authority-level variance. This implies that the variance of the ward-level error term in the investigated model was 17% smaller than in the baseline model, and the variance of the local authority-level error term was 76% smaller than in the baseline model. Another way of interpreting this is that the explanatory variables in the investigated model successfully explained 17% of the small scale geographic variance and 76% of the large scale geographic variance that was unexplained by the baseline model.

Spatial error regression analysis

As discussed above, standard linear regression modelling techniques are inappropriate for the ecological data used in this thesis as they do not take account of the spatial structure of the dataset. Multi-level modelling accounts for the different levels of geographic variation that exist in CHD rates, but it does not account for spatial autocorrelation and the bias that it can cause. This bias is appropriately accounted for by fitting a spatial error regression model.

Spatial autocorrelation is the phenomenon that areas in close proximity to each other are likely to have similar measures of both outcome and explanatory variables, and hence their residuals from regression models are likely to be similar. This violates the assumption that residuals are randomly distributed and independent of each other, and the result is a bias of the analysis away from the null hypothesis and to underestimates of standard errors (Loftin and Ward, 1983). Spatial error regression models account for this

potential bias by introducing a further parameter into the regression model that adjusts the modelled estimate of the outcome variable for an area to take account of the average value of the outcome variable in all neighbouring areas. The regression equation is described in equation 4.3, where λ is the additional 'spatial error' parameter.

$$(4.3)\quad y = \alpha + \beta_1 x_1 + ... + \beta_n x_n + \lambda + \varepsilon$$

The spatial error parameter is included in the systematic part of the model, and should not be confused with the random error term, ε.

Spatial error regression modelling depends upon a definition of 'neighbouring' areas and a software package that accounts for the spatial positioning of each of the areas in the analysis. For the analyses conducted in this thesis, the software package GeoDa (Anselin, 2003) and a shape file describing the boundaries of all wards in England downloaded from UKBORDERS (Edina, 2008) were used. There are a number of ways to define 'neighbouring' wards: for the purpose of this thesis a "queen's 1st order" matrix was used, as is reasonably standard in spatial error analyses (Cliff and Ord, 1973). Such a matrix defines neighbouring wards as all wards that have either common border lines or a common border point (i.e. that are adjacent to each other, or that meet at a single point).

Using this definition it is possible to calculate the degree of spatial autocorrelation in the dataset by calculating Moran's Global I statistic. This is calculated by plotting a variable against the average value of the same variable for all neighbouring wards, and then calculating Pearson's correlation coefficient for the plot – therefore, a Global I statistic of 1 indicates that ward values are perfectly correlated with neighbouring values and hence

there is perfect spatial autocorrelation, and a value of 0 shows no correlation and hence no spatial autocorrelation. Table 4.1 in appendix one shows the Global Moran's I statistic for each of the outcome variables and all explanatory variables used in the thesis. Significant spatial autocorrelation was shown, indicating a large potential for bias if this autocorrelation is not properly addressed.

Ideally, multi-level spatial error regression models would have been used for this thesis, thereby accounting for both spatial autocorrelation and the spatial structure of the dataset simultaneously. However, software packages are not currently available that can build such models. Therefore, the results of the spatial error regression models in this thesis are provided for comparison with the multi-level models - substantial deviation between the two sets of models was interpreted as evidence of spatial autocorrelation bias in the multi-level models. The tables reporting the spatial error regression models are similar to those reporting the multi-level models, but contain an extra column which indicates whether the parameter estimates are similar to those found in the equivalent multi-level models. In addition, the tables also report the r^2 value for the spatial error regression models – a measure of the amount of variance in the outcome variables that has been explained by variance in the explanatory variables included in the model. However, this measure includes variance that has been explained by the spatial error term (λ), therefore not all of this explained variance is due to the environmental variables and behavioural risk factor profiles of populations. The r^2 value also does not separate the variance into small scale and large scale geographic variation as is the case for the multi-level models.

Histograms showing the distribution of the residuals from the spatial error regression models reported in table 8.4 (chapter eight) are shown in figure 4.5 in appendix one. They show that the variance of the error terms follow a reasonably normal distribution with mean zero, suggesting that the models were well-fitted.

Synthetic estimation

As described in the previous chapter, the technique of synthetic estimation of the prevalence of a behavioural risk factor for a ward involves building a logistic regression model that estimates the probability of an individual having the risk factor in question given both individual-level variables (e.g. age, sex) and area-level variables (e.g. ward-level measures of deprivation or unemployment rate). This logistic regression model is then applied to all individuals within a ward to produce an estimate of the number of people in the ward who have the risk factor.

For synthetic estimation, it is assumed that the prevalence of a risk factor in any population is driven by underlying tendencies (e.g. the socio-demographic mix of the population) but also has a random element (Heady et al., 2003). The synthetic estimation process results in estimates of only the systematic (non-random) part of the prevalence, as opposed to direct survey-based measures which estimate both the fixed and random element for each population. Synthetic estimates are not (necessarily) biased even though they only estimate the systematic element of the prevalence, since it is assumed that the set of all random elements for a large geographic area are normally distributed with mean

of zero. Therefore the random element has expectation of zero, and the set of all synthetic estimates for a large area are unbiased.

The multi-level logistic regression model supporting synthetic estimates is shown below, where p is the probability that an individual has a risk factor based on individual-level variables $x_1, \ldots, x_m$ and area-level variables $z_1, \ldots, z_n$. The parameters to be estimated are $\alpha, \beta_1, \ldots, \beta_m, \psi_1, \ldots, \psi_n$, u is the random area-level term and e is the random individual-level term,

$$(4.4)\ \log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \ldots + \beta_m x_m + \varphi_1 z_1 + \ldots + \varphi_n z_n + u + e$$

The parameters from equation 4.4 are calculated using data from a national survey supplemented with area-level variables from the census. Rather than use equation 4.4 to compare the odds of different sets of individuals having the risk factor in question (which is the usual output from a logistic regression analysis), the equation is solved for p, giving equation 4.5 below.

$$(4.5)\ p = \frac{e^{\alpha+\beta_1 x_1+\ldots+\beta_m x_m+\varphi_1 z_1+\ldots+\varphi_n z_n}}{1+e^{\alpha+\beta_1 x_1+\ldots+\beta_m x_m+\varphi_1 z_1+\ldots+\varphi_n z_n}}$$

Equation 4.5 is then applied to all individuals living within a ward, which provides an estimate of the number of people in the ward who have the risk factor (and hence the prevalence rate can be calculated). Note that equation 4.5 does not include the two

random terms u and e. This implies that synthetic estimates are actually estimates of the *expected prevalence* of a risk factor given the socio-demographic profile of the population of a ward, rather than a measure of the *actual prevalence*.

Trend surface modelling

The technique of trend surface modelling assumes that the variable that is being modelled varies smoothly over the entire geographic area for which estimates are being modelled. The technique is therefore appropriate for environmental variables such as climate or air pollution, but inappropriate for aggregate measures such as mortality rates. The technique involves taking measurements of the variable from a set of defined spatial points within the geographic area and extrapolating the measurements to all other points within the geographic area by building a regression model with spatial co-ordinates as the explanatory variables, and hence depends upon the assumption that the set of spatial points from which measurements are available are equivalent to a random sample of points drawn from the larger population of all spatial points within the geographic area.

Trend surface modelling can be conducted using any order of explanatory variables - a first order model uses just the two-dimensional spatial co-ordinates as explanatory variables and assumes that the variable changes linearly over the geographic area; a second order model includes squared co-ordinate terms and assumes that the variable changes quadratically over the geographic area, and so on. The choice of what order model to apply depends upon the nature of the environmental variable and the number of

measurement points available. For the case of the climate variables used in this thesis, second order trend surface modelling was chosen for two reasons. Firstly, climate variables are large scale environmental variables that do not vary radically in small regions and are unlikely to produce complex surfaces such as saddle points, sharp peaks or steps that would require higher order modelling. Secondly, the climate data were taken from only 37 weather stations, which would not allow for higher order modelling since higher order models have more explanatory variables and require greater statistical power.

The equation used to model the climate variables is shown below, where z is the climate variable, x is the Eastings UK grid reference, y is the Northings UK grid reference, α, $\beta_1,\ldots\beta_5$ are the parameters to be estimated and ε is a normally distributed random error term with mean zero.

$$(4.6)\quad z = \alpha + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 xy + \varepsilon$$

Such a model was constructed for each of the climate variables, and for every month between January 2000 to December 2002. The models were then used to construct estimates of the climate variables for each ward in England, using the grid reference of the weighted population centres of the wards as the inputs for the models. The grid references of the weighted population centres were calculated by averaging the centroid grid references of all post codes within each ward (using data provided from the All Fields Postcode Directory (Office for National Statistics, 2002)). The modelled estimates

for each month were then aggregated to create a mean modelled estimate of each of the climate variables for the period 2000-2002. The assumption that the weather stations are equivalent to a randomly selected sample of data points seems reasonable since the weather stations provide good coverage of England (97% of the wards in England are within 50km of at least one weather station), and this distribution was selected by the Meteorological Office so that they could estimate trends in climate variables using similar extrapolation techniques.

Principal components analysis

One of the uses of regression analysis is to derive an estimate of the magnitude and direction of an association between an explanatory variable and an outcome variable after adjustment for a set of other variables that may be explanatory or confounding. This is achieved by generating a transformation of the set of explanatory and confounding variables (what is usually referred to as the 'model') that most closely resembles the outcome variable, and hence reduces the sum of the squared difference between the outcome variable and this transformation to a minimum. The magnitude and direction of the association between the explanatory variable and the outcome variable is derived from the coefficient of the explanatory variable in this transformation.

In general this process is reasonably robust, in that small changes to the system (e.g. removal of some data points or addition of extra non-confounding explanatory variables) have little impact on the generated model. However, this is not the case when a number

of the explanatory variables are highly correlated. Since highly correlated explanatory variables essentially have the same role in regression models, it is possible that substantially different models can produce similarly small sums of squared residuals – for example, by simply exchanging the coefficients on two highly correlated variables. This leaves the model vulnerable to small changes in the dataset, and hence reduces confidence in the estimates of association produced by the model. This is the situation when the sets of synthetic estimates used for this thesis are used as explanatory variables, as they are mostly highly correlated with each other.

One method of addressing this problem is to transform the problem variables to create a new set of orthogonal variables (i.e. variables that are not at all correlated with each other). There are a number of algorithms that could be used to generate such transformations – the process used in this thesis is known as principal components analysis. Here, the covariance matrix of the problem variables *(C)* is calculated and a transformation matrix *(T)* is constructed such that *C*T* is a diagonal matrix where the magnitude of the values decreases along the diagonal. The transformations to be performed on the problem variables are described by the row vectors of *T*, and the resultant diagonal matrix *(C*T)* is the covariance matrix of the transformed variables. Since *C*T* is diagonal the covariance of any two of the transformed variables is zero, and hence the transformed variables are all uncorrelated. The values along the leading diagonal are the amount of variance in the transformed variables and can be used to calculate how much of the variance in the original set of variables is included in the transformed variables.

One use of principal components analysis is to decrease the number of explanatory variables that are under investigation, in order to aid interpretation of the regression model. Suppose there are 5 highly correlated variables that are included in a principal components analysis. Then *C*, *T* and *C*T* are all 5x5 matrices and hence the result is five transformed variables. Suppose the first transformed variable explains 80% of the variance in the five original variables, the second 12%, the third 5%, the fourth 2% and the fifth 1%. In this situation, the third, fourth and fifth transformed variables do not provide much information about the variance of the explanatory variables and would usually be dropped from further analysis. Therefore the five explanatory variables have now been transformed into two new explanatory variables. The transformation matrices and amount of variance described by the transformed variables for the five sets of synthetic estimates used for this thesis are shown in table 4.2 in appendix one.

Chapter 5: The association between environmental variables and coronary heart disease mortality and hospitalisation rates in England

INTRODUCTION

This chapter covers the first of the analyses conducted for this thesis – an exploration of the association between environmental variables and coronary heart disease (CHD) mortality and hospitalisation rates in England. The analyses were designed to answer the following questions:

RQ1: Which environmental variables are significantly associated with CHD mortality and hospitalisation rates in England?

RQ2: How much of the large scale and small scale geographic variation in CHD mortality and hospitalisation rates can be explained by environmental variables alone?

The environmental variables that are included in this chapter are measures of the climate (average maximum daily temperature, average minimum daily temperature, total annual hours of sunshine, and total annual rainfall), air pollution and urbanicity (a categorical variable splitting wards into 'coastal and countryside', 'urban' and 'metropolitan'). Chapter two describes the literature supporting the hypothesised relationship between these variables and CHD rates and the conceptual framework that underpins these

analyses is shown in figure 2.1 (chapter two). A detailed description of the data sources and modelling applied to the environmental variables is described in chapters three and four.

METHODS

For each of the analyses reported here four sets of models were built, which considered different outcome variables: male CHD mortality rates, female CHD mortality rates, male CHD hospitalisation rates and female CHD hospitalisation rates – all directly age-standardised to the European Standard Population (West Midlands Public Health Observatory, 2009). Both multi-level regression modelling and spatial error regression modelling were conducted, using all 7,929 English wards (nested in the 354 English local authorities for the multi-level modelling) as the units of analysis. The results of the multi-level modelling were used to estimate the amount of small scale and large scale geographic variation in CHD rates that was explained by the environmental variables, modelled by the ward-level variance (the mean variance between wards within a single local authority) and local authority-level variance (the variance between all local authorities in England) respectively.

The results of the spatial error modelling were assessed for supporting evidence of the relationships displayed in the multi-level models after accounting for spatial autocorrelation in both the explanatory and outcome variables which has the potential to bias the results of the multi-level models. For each of the environmental variables, the

spatial error and multi-level models were said to 'agree' if the estimated association with CHD predicted by the models (i.e. the beta coefficients) were both of the same sign and the statistical accuracy of the estimates (i.e. the p value) were either both higher or both lower than 0.01.

Initially exploratory data analysis techniques were used to investigate correlations between the explanatory variables and check the distribution of the outcome variables (see appendix one). Three stages of modelling were then conducted. Baseline models were built that did not include any explanatory variables to determine the amount of small scale and large scale geographic variation in CHD rates after adjustment for sex and age only. Secondly, each of the environmental variables were added to the models individually to assess the crude association with CHD rates. Finally, models were constructed that included all environmental variables that displayed a statistically significant ($p < 0.05$) association in the univariate analyses. Additionally, the residuals of the spatial error regression models were assessed for spatial autocorrelation to determine whether the models had appropriately accounted for potential spatial autocorrelation bias. The spatial error regression modelling was conducted using the GeoDa software package (Anselin, 2003), and the multi-level modelling was conducted using MLwiN v2.02 (Rasbash et al., 2003). Throughout this thesis, the estimation technique used for the multi-level modelling was iterative generalised least squares (IGLS), and the spatial error modelling used maximum likelihood techniques. This should ensure that the results of the multi-level models and the spatial error models are comparable, since IGLS is an extension of the maximum likelihood process to incorporate the structure of the dataset.

In order to ensure this comparability was maintained, Monte Carlo Marcov Chain estimation techniques were not used for the multi-level modelling.

RESULTS

Exploratory data analysis

Table 5.1 shows descriptive statistics for the outcome and explanatory variables included in the analyses for this chapter, and provides a correlation matrix for the continuous explanatory variables. Both the female mortality and hospitalisation rate variables include wards that had zero events, but only eight wards had zero mortality events and one ward had zero hospitalisations. The climate variables showed little variance, as would be expected in a temperate country. The maximum temperature, minimum temperature and sunshine variables were all reasonably collinear. The rainfall variable was less correlated with the other three climate variables – this is because the rainfall variable shows a predominately West-East gradient with greater rainfall in the West of the country, whereas the two temperature variables demonstrate a North-South gradient, and the sunshine variable demonstrates a North East – South West gradient.

Table 5.1 Summary statistics and correlation matrix for outcome variables (CHD mortality and hospitalisation rates) and explanatory variables (environmental risk factors) (wards; n = 7,929)

Outcome variables

Variable	*Range*	*Interquartile range*	*Standard deviation*	*Mean*	*Median*
CHD mortality rate, men†	24.4 – 525.3	142.5 – 212.1	53.6	179.9	174.9
CHD mortality rate, women†	0.0 – 336.2	63.0 – 100.6	29.7	83.6	80.5
CHD hospitalisation rate, men†	151.0 – 3,486.9	688.8 – 1,048.9	291.4	892.1	854.7
CHD hospitalisation rate, women†	0.0 – 1,743.2	253.4 – 449.1	164.8	368.4	339.7

Continuous explanatory variables

Variable	*Range*	*Interquartile range*	*Standard deviation*	*Mean*	*Median*
Mean max. temp (°C)	11.2 – 14.4	13.5 – 14.4	0.6	13.9	14.1
Mean min. temp (°C)	5.3 – 9.8	5.8 – 6.9	0.7	6.4	6.5
Sunshine (000s hrs / yr)	1.3 – 1.7	1.4 – 1.6	0.1	1.5	1.5
Rainfall (m / yr)	0.8 – 1.3	0.9 – 1.0	0.1	0.9	0.9
Air quality index (SDs)	0.4 – 2.2	0.9 – 1.3	0.3	1.1	1.1

Categorical explanatory variable

Category	*Number of wards*	*%*
Coastal and countryside	2,737	35
Urban	4,708	59
Metropolitan	484	6
Total	**7,929**	**100**

Correlation matrix (Pearson's r) of continuous explanatory variables

	Mean max. temp	Mean min. temp	Sunshine	Rainfall	Air quality index
Mean max. temp	1.00				
Mean min. temp	0.72	1.00			
Sunshine	0.63	0.88	1.00		
Rainfall	-0.42	0.18	0.26	1.00	
Air quality index	0.21	-0.14	-0.07	-0.31	1.00

† Age-standardised rate per 100,000

Multi-level regression modelling

The baseline models displayed around 75% of the total geographic variation in CHD mortality rates at ward-level, and around 55% of the total geographic variation in CHD

hospitalisation rates at ward-level, with the remainder at local authority-level (see table 5.2). This suggests that the average variance in CHD rates for wards within a local authority was higher than the variance between local authorities within England, and hence that small scale geographic variations in CHD rates are larger than large scale geographic variations.

Table 5.2 Residual variance at ward-level (n = 7,929) and local authority-level (n = 354) for baseline and final multi-level models

		BASELINE		*FINAL*	
		Variance	*Standard Error*	*Variance*	*Standard Error*
Mortality models					
MEN	Ward-level	2,096.4	34.1	2,001.6	32.5
	LA-level	779.7	66.3	340.6	32.8
WOMEN	Ward-level	660.8	10.7	641.1	10.4
	LA-level	226.8	19.5	92.6	9.3
Hospitalisation models					
MEN	Ward-level	48,594.9	789.6	44,696.1	726.2
	LA-level	37,958.9	3,034.8	24,983.0	2,042.5
WOMEN	Ward-level	14,884.6	241.8	13,867.8	225.4
	LA-level	12,618.2	1,004.3	7,573.5	617.1

All of the environmental variables showed a statistically significant association with both mortality and hospitalisation rates in univariate analyses, with the exception of total annual rainfall, which was not associated with hospitalisations for either men or women. As would be expected, the climate variables were successful at explaining large scale geographic variation in the univariate analyses, and the air quality index and urbanicity variables were successful at explaining small scale geographic variation.

Table 5.3 shows the models that included all environmental variables that were significantly associated with CHD rates in the univariate analyses. The mortality models

shown in table 5.3 explained a considerable amount of large scale geographic variation (local authority-level variance: nearly 60%) but little of the small scale geographic variation. The hospitalisation models were less successful at explaining large scale geographic variation (around 40%) and again explained little of the small scale geographic variation.

Table 5.3 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against environmental variables (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	663.4			318.7		
Mean max. temp (°C)	-34.5	4.4	<0.001	-16.3	2.3	<0.001
Mean min. temp (°C)	11.2	4.0	0.005	2.7	2.1	0.215
Sunshine (000s hrs / yr)	-93.4	24.8	<0.001	-41.8	13.2	0.001
Rainfall (m / yr)	5.0	26.5	0.849	10.8	14.1	0.441
Air quality index (SDs)	43.8	3.9	<0.001	19.9	2.1	<0.001
Urban†	12.6	1.3	<0.001	5.4	0.7	<0.001
Metropolitan†	31.0	3.4	<0.001	15.0	1.9	<0.001
Ward-level variance explained:		5%			3%	
LA-level variance explained:		56%			59%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	3,784.1			2,010.6		
Mean max. temp (°C)	-227.6	19.3	<0.001	-126.0	10.6	<0.001
Mean min. temp (°C)	220.6	25.6	<0.001	79.4	14.1	<0.001
Sunshine (000s hrs / yr)	-1,077.7	173.6	<0.001	-441.8	95.9	<0.001
Rainfall (m / yr)						
Air quality index (SDs)	368.5	21.9	<0.001	209.5	12.2	<0.001
Urban†	74.6	6.1	<0.001	34.2	3.4	<0.001
Metropolitan†	232.4	17.4	<0.001	105.6	9.7	<0.001
Ward-level variance explained:		8%			7%	
LA-level variance explained:		34%			40%	

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations

Total annual rainfall (which was not included in the hospitalisation models as it did not display a significant association in univariate analyses) was not significantly associated with CHD mortality rates after adjustment for other environmental variables. Mean daily minimum temperature was positively associated with both CHD mortality and hospitalisation rates after adjustment for other environmental variables (non-significantly in the female mortality model). This suggests that CHD rates are higher in areas of higher minimum temperatures, after adjustment for maximum temperatures, and is a reversal of the results found in the univariate analyses. Both total annual sunshine and mean daily maximum temperatures remained negatively associated with CHD rates in the adjusted models, and air quality and urbanicity remained positively associated.

Spatial error regression modelling

As was the case for the univariate multi-level models, all of the environmental variables were significantly associated with CHD rates in univariate spatial error models, with the exception of rainfall in the hospitalisation rates models. The results of the fully adjusted models are shown in table 5.4, and suggest that the multi-level models were not particularly affected by spatial autocorrelation bias. Only the parameter estimates for the average daily minimum temperature variable in the male mortality models showed any substantial difference – the parameter estimate was significantly different from zero in the multi-level model ($p = 0.005$) but not so in the spatial error model ($p = 0.050$).

Residuals from the spatial error models had very low spatial autocorrelation (results not shown).

Table 5.4 Spatial error regression models of (1) CHD mortality rates and (2) CHD hospitalisation rates against environmental variables (wards, n = 7,929)

(1) Mortality rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi-level model**	*Beta*	*SE*	*p*	*Agree with multi-level model**
Constant	577.8				283.6			
Mean max. temp (°C)	-27.0	3.4	<0.001	✓	-13.6	1.9	<0.001	✓
Mean min. temp (°C)	5.8	3.0	0.050	x	0.2	1.6	0.922	✓
Sunshine (000s hrs / yr)	-88.9	18.6	<0.001	✓	-34.0	10.1	0.001	✓
Rainfall (m / yr)	26.2	20.2	0.195	✓	16.8	11.0	0.126	✓
Air quality index (SDs)	36.8	3.6	<0.001	✓	17.9	2.0	<0.001	✓
Urban†	11.5	1.3	<0.001	✓	4.9	0.7	<0.001	✓
Metropolitan†	25.1	3.4	<0.001	✓	12.0	1.9	<0.001	✓
Spatial error	0.4	0.0	<0.001		0.4	0.0	<0.001	
Model r^2:	0.27				0.25			

(2) Hospitalisation rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi-level model**	*Beta*	*SE*	*p*	*Agree with multi-level model**
Constant	3,601.2				1,864.4			
Mean max. temp (°C)	-198.2	14.4	<0.001	✓	-108.2	7.9	<0.001	✓
Mean min. temp (°C)	212.5	17.2	<0.001	✓	72.9	9.5	<0.001	✓
Sunshine (000s hrs / yr)	-1,157.7	126.7	<0.001	✓	-470.0	69.8	<0.001	✓
Rainfall (m / yr)								
Air quality index (SDs)	336.7	24.6	<0.001	✓	204.4	13.6	<0.001	✓
Urban†	64.3	6.0	<0.001	✓	27.9	3.4	<0.001	✓
Metropolitan†	203.5	18.6	<0.001	✓	80.5	10.4	<0.001	✓
Spatial error	0.6	0.0	<0.001		0.6	0.0	<0.001	
Model r^2:	0.47				0.48			

* ✓ = Direction of association and whether association is significant (p<0.01) agrees with equivalent multi-level regression models; x = otherwise.

† Compared to baseline of 'Coastal and Countryside'

SDs: Standard Deviations

In contrast to the multi-level models, the spatial error models explained more of the variance in hospitalisation rates (nearly 50%) than in mortality (around 25%). This is likely to be due to the greater amount of spatial autocorrelation that is apparent in the hospitalisation rates than in the mortality rates (see table 4.1, appendix one). Since the hospitalisation rates are more autocorrelated than the mortality rates, more of the variance in the hospitalisation rates can be explained by rates in neighbouring wards – this is precisely the association that is modelled by inclusion of the spatial error term.

The parameter estimates in the spatial error models tended to be closer to zero than in the multi-level models, demonstrating that spatial autocorrelation (when unaccounted for) tends to result in a bias away from the null hypothesis. The difference in the parameter estimates was generally in the region of around 10% to 20%.

DISCUSSION

In the crude models displayed in this chapter (unadjusted for deprivation and the behavioural risk factor profile of populations) around 5% of the small scale geographic variation in CHD mortality and hospitalisation rates was explained by environmental variables, specifically urbanicity and air quality. Around 60% of the large scale geographic variation in mortality rates and 40% of the large scale geographic variation in hospitalisation rates was explained by climatic differences in temperature and exposure to sunlight. These findings are unlikely to be a result of spatial autocorrelation bias.

However, the models that are presented in this chapter are crude – that is, variables that may confound the association between the environmental variables and CHD rates have not been included in the modelling process, and the results must be interpreted in this context. Deprivation, in particular, may be an important confounding factor. It is well-established that deprivation is associated with CHD outcomes (Law and Morris, 1998; Romeri, 2006), that deprivation is higher in more urban areas (Department of the Environment, Transport and the Regions, 2000), and that deprivation follows a broad North-South gradient similar to some of the climate variables studied in this chapter. As deprivation is also associated with prevalence rates of behavioural risk factors for CHD further analyses that include both deprivation and behavioural risk factor profiles of populations are required to confirm the relationship between environmental variables and CHD. Such analyses are reported in chapter eight.

The results displayed in table 5.3 suggest that a number of climatic variables are associated with CHD rates independently of each other, including both average maximum and minimum daily temperature, the regression coefficients of which have opposite signs. One possible explanation of such a finding is that the collinearity between the maximum temperature and minimum temperature variables has resulted in spurious statistical associations, and the positive association found for minimum temperature is actually an adjustment of the negative association found for maximum temperature. Another explanation is that the results represent a genuine relationship between temperature and CHD: that high CHD rates are found in areas with a *low* maximum temperature, but a relatively *high* minimum temperature. It has previously been observed that the impact of

cold air temperature on cardiovascular mortality is affected by social and cultural norms in housing standards and thermoregulatory behaviour (e.g. wearing appropriate outdoor clothing); so mortality rates in Norway increase sharply once the air temperature drops below 0°C, whereas a similar sharp increase occurs in Ireland once the air temperature drops below 10°C (Mercer, 2003), and countries in the South of Europe with relatively mild winters tend to suffer higher excess winter mortality rates than Northern European countries (Keatinge et al., 1997). Therefore people living in areas with high minimum temperatures (i.e. that never get very cold) may be less prepared for cold weather than others, and are hence more susceptible to it. Given that climatic conditions do not vary radically throughout England, where the entire range of maximum temperature is only around 4°C, it seems more plausible to accept that the findings are due to spurious statistical findings, particularly since the results of the spatial error modelling suggest that the minimum temperature variable is not significantly associated with male CHD mortality rates after adjustment for maximum temperature.

The independent associations of maximum temperature and sunshine may also be due to spurious statistical findings, since these two variables are also reasonably collinear. However, in the final models displayed in tables 5.3 and 5.4 both variables retain the significant negative associations with CHD rates that they showed in univariate analyses. These independent associations are supported by the literature which suggests different physiological mechanisms for the relationship between temperature and sunshine and CHD (Toledano et al., 2005; Grimes et al., 1996).

The models displayed in tables 5.3 and 5.4 are broadly similar in their results. The main differences are that the regression parameters in the spatial error models are slightly reduced from those reported in the multi-level models. This suggests that spatial autocorrelation bias has not had a large effect on the multi-level models, so we can have some faith in the findings of these models. The results of the multi-level models are particularly important as they assess the amount of ward-level geographic variation that remains after accounting for LA-level geographic variation, i.e. to assess whether there was an additional small scale area effect of climate, say, after accounting for the large scale area effects of climate.

In general, the results of this chapter concur with other studies of the effect of environmental variables on the geographic variation in CHD rates. Higher CHD rates have previously been found in places that are colder (West and Lowe, 1976), less sunny (Morris et al., 2001), urban (Law and Morris, 1998) or have high air pollution (Maheswaran et al., 2005). In general, the authors of these studies adjusted their analyses for confounding variables such as deprivation; a comparison of the results of this thesis with the literature is therefore better placed in chapter eight, where the results of further modelling of the impact of environmental variables on CHD that include adjustments for both deprivation and the behavioural risk factor profile of populations are included.

CONCLUSIONS

The analyses reported in this chapter have shown that temperature, sunshine, air pollution and urbanicity are significantly associated with both CHD mortality and hospitalisation

rates, and that they should therefore be considered in analyses aimed at explaining geographic variations in CHD rates. The results suggest that the relationship between outdoor air temperature and CHD rates can be adequately modelled with only one temperature variable, and therefore inclusion of both mean daily minimum and maximum temperature variables in further analyses is unjustified. Since the maximum temperature variable showed a stronger relationship with CHD in univariate analyses than the minimum temperature variable, only maximum temperature was included in the analyses reported in chapter eight. Similarly, since total annual rainfall did not show an independent association with either mortality or hospitalisation rates in this chapter, the rainfall variable was not included in the analyses reported in chapter eight.

Chapter 6: Validation of synthetic estimates of the prevalence of behavioural risk factors for coronary heart disease at the ward-level in England

INTRODUCTION

This is the second analytical chapter of the thesis. The following two chapters report analyses of geographic variation of coronary heart disease (CHD) rates in England that include model-based estimates of the prevalence of behavioural risk factors for CHD at ward level as explanatory variables. This chapter is concerned with the validity of these model-based estimates. The model-based estimates considered in this chapter were drawn from different sources and utilise different modelling techniques, but the general term for the techniques is 'synthetic estimation' (Heady et al., 2003) and for simplicity they are henceforth known as 'synthetic estimates'. An outline of the theoretical concepts underlining synthetic estimation is provided in chapter four. This chapter answers the following research questions:

RQ1: Are the sets of synthetic estimates identified in chapter three valid and accurate estimates of the prevalence of behavioural risk factors for CHD at ward-level in England?

RQ2: Which of the sets of synthetic estimates identified in chapter three are the most appropriate for use as explanatory variables in analyses of the geographic variation of CHD rates in England?

Synthetic estimation techniques were developed to allow for small area estimation of variables studied by national surveys that have been designed to only produce estimates for large areas or populations. Such small area estimates are increasingly required by government departments, principally to inform more directed resource allocation for problems such as poor health, poor housing conditions, unemployment and low pay (Heady et al., 2003). Different techniques for small area estimation were developed in the early 1990s (Ghosh and Rao, 1994), and by the late 1990s synthetic estimates of the prevalence of behavioural risk factors for CHD for electoral wards were being developed by health geographers (Twigg et al., 2000). The Office for National Statistics (ONS) developed a Small Area Estimation Project in April 1998, the aim of which was to develop the methods behind synthetic estimation, and also to produce methods for testing the accuracy and validity of developed estimates. The ONS published a report of the work conducted by the Small Area Estimation Project in 2003, which included the description of diagnostic tests of bias and 'goodness of fit' for the models that support synthetic estimates (Heady et al., 2003).

In addition to the Small Area Estimation Project, the differing synthetic estimation techniques were also assessed in a European project known as EURAREA (EURAREA Consortium, 2004). The aim of this project was to evaluate whether synthetic estimates outperform survey-based estimates for small areas by comparing their performances at different estimation levels (e.g. local authority (LA)-level and ward-level). Census data from six European countries were used to create an artificial population, in which the true values for area means for each variable were known. A number of different samples were

then drawn from this population, and the different techniques were used to estimate small area means for three target variables, which were then compared against the known true values. The report concluded that:

> *"At NUTS5 [e.g. electoral ward] level model-based estimators [i.e. synthetic estimates] substantially outperform design-based methods, even when the models have fairly limited predictive power, but of course the gain from using model-based methods increases substantially with the availability of explanatory covariates that are highly correlated with the target variable… Theoretical expectations regarding the performance of these model-based estimators under sampling from fixed populations seem to be broadly supported"*
>
> (Page E-2, EURAREA Consortium, 2004).

Hence, the methodology supporting synthetic estimation has been validated and is well established, and it is not the purpose of this chapter to re-examine it. Rather, the aim is to validate individual sets of ward-level synthetic estimates of the prevalence of behavioural risk factors for CHD in England. According to the Small Area Estimation Project and EURAREA reports, these estimates should be valid and accurate provided that a) the risk factor in question is strongly associated with individual-level and area-level covariates, and b) the developed model is well-fitted. If the first criterion is satisfied then it is possible to create a model that explains a large proportion of the variance in the prevalence of behavioural risk factors. If the second criterion is satisfied then the developed model accurately describes the relationship between uptake of the risk factor and the individual-level and area-level covariates. Due to constraints in the modelling process it is not straightforward that the second criterion should be met. For example, the individual-level variables used in the models are constrained by the provision of UK census data, which are generally displayed in tables stratified by no more than three

individual-level variables (e.g. population of each ward by age-sex-ethnicity groups, although further tables stratified by more individual-level variables may be available through commission), therefore the models are restricted to a maximum of three individual-level covariates. Further, the models depend on the quality of data collected by the national survey, which is variable. Data for models of raised cholesterol, for example, are dependent upon the data collected in the blood sample section of the Health Survey for England (HSfE), which suffers from a response rate slightly lower than 50% (Department of Health, 2000b).

The sets of synthetic estimates that are considered in this chapter are described in table 3.4 in chapter three. For ease of reference, this table is repeated below, with abbreviated names for each set of synthetic estimates for use throughout this chapter.

Table 6.1 Synthetic estimation models to be assessed in this chapter

Short ID	Synthetic estimate of percentage of population…	Reference
FV01	eating less than 5 portions of fruit and vegetables per day, 2001 model	(Dibben et al., 2004)
FV03	eating less than 5 portions of fruit and vegetables per day, 2003 model	(Dibben et al., 2004)
PA01	doing under five hours of physical activity in a week, 2001 model	(Dibben et al., 2004)
PA03	doing under five hours of physical activity in a week, 2003 model	(Dibben et al., 2004)
ALC	consuming greater than weekly recommended average intake of alcohol	(Twigg et al., 2000)
SMOK00	who are current smokers, 2000 model	(Twigg et al., 2000)
SMOK04	who are current smokers, 2004 model	(Twigg et al., 2004)
SMOK01	who are current smokers, 2001 model	(Dibben et al., 2004)
SMOK03	who are current smokers, 2003 model	(Dibben et al., 2004)
OBES01	with BMI>=30kg/m^2, 2001 model	(Dibben et al., 2004)
OBES03	with BMI>=30kg/m^2, 2003 model	(Dibben et al., 2004)
BP01	with SBP>=160mmHg or DBP>=95mmHg, 2001 model	(Dibben et al., 2004)
BP03	with SBP>=160mmHg or DBP>=95mmHg, 2003 model	(Dibben et al., 2004)
CHOL01	with total cholesterol >=6.5mmol/l, 2001 model	(Dibben et al., 2004)
CHOL03	with total cholesterol >=6.5mmol/l, 2003 model	(Dibben et al., 2004)
DIAB	with diagnosed and undiagnosed type 1 and type 2 diabetes	(YHPHO, 2005)

BMI = Body Mass Index, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure. The Health Poverty Index synthetic estimates (Dibben et al., 2004) were originally developed in 2001 and then updated in 2003 (hence the two models for each risk factor).

During their development, the sets of synthetic estimates were submitted to various testing procedures, similar to those described by the Small Area Estimation Project report. Table 6.2 provides details about this work.

To date, there has been a limited amount of published validation work on the synthetic estimates. No evidence was found of any validation assessment for the synthetic estimates developed by Dibben et al. for the Health Poverty Index website, and very little evidence was found for the Yorkshire & Humberside Public Health Observatory diabetes synthetic estimates. The smoking and problem drinking synthetic estimates developed by Twigg and colleagues at University of Portsmouth (models SMOK00, SMOK04 and ALC) have been assessed for validation in a number of ways. This work suggests that the sets of synthetic estimates of the prevalence of smoking are accurate – the results are less promising for the synthetic estimates of the prevalence of problem drinking. As yet, no work has been done to assess the predictive validity of any of the sets of synthetic estimates.

If there is to be any confidence in the results of the analyses reported in chapters seven and eight, the synthetic estimates that are used must be accurate and validated. The different synthetic estimates are currently used by health professionals and policy makers to identify areas where resources should be allocated to tackle high levels of unhealthy behaviour, and the validity assessments used here can also be used to evaluate whether these decisions on resource allocation are being made appropriately.

Table 6.2 Previous attempts at validation of the sets of synthetic estimates assessed in this chapter

Model	Forms of validity tested	Details of testing	Results of testing	Reference
FV01 FV03 PA01 PA03 SMOK01 SMOK03 OBES01 OBES03 BP01 BP03 CHOL01 CHOL03	None	Not applicable	Not applicable	Not applicable
ALC SMOK00	Face	Model variables were assessed for expected associations with the outcome variables.	*"The resulting models uphold conventional wisdoms concerning both smoking and problem drinking. Male gender and being single are particularly important factors in both cases, the former notably so in the case of problem drinking"*	(Twigg et al., 2000)
ALC SMOK00	Construct	Geographic variation of synthetic estimates compared against expected geographic variation.	*"It is possible to generate small-area predictions of health-related behaviours, which conform to expected patterns and vary in an expected way around established national and regional means"*	(Twigg et al., 2000)

Table 6.2 (cont.)

Model	Forms of validity tested	Details of testing	Results of testing	Reference
ALC SMOK00	Convergent	Synthetic estimates were compared against local survey-based estimates. Three local surveys were used: Health of the Welsh (1996); Newcastle and South Tyne Health Survey (1994); Health Quest Portsmouth and South East Hampshire (1993)	Correlation between synthetic and survey estimates: SMOK00: Wales r=0.49 Portsmouth r=0.55 Newcastle r=0.75. All are statistically significant. ALC: Wales r=-0.13 Portsmouth r=0.08 Both are non-significant. The range of estimates was much smaller for the synthetic estimates in comparison to the survey estimates.	(Twigg and Moon, 2002)
SMOK04	Face	The models were examined for 'goodness of fit' (by comparing the deviance statistic of the null and full models), and for how accurately they predict individual survey responses.	The SMOK04 model achieves a 14% reduction in deviance, in comparison to the null model. The null smoking model correctly predicts 50% of survey responses. The SMOK04 model increases this to 60%.	(Twigg et al., 2004)
SMOK04	Convergent / construct	The synthetic estimates were combined to produce estimates for strategic health authorities, and compared against survey estimates using similar HSfE data (each drawn from 1998 to 2001)	Correlation between synthetic and survey estimates, r=0.49 (p=0.009). *"The difference between the synthetic and direct [i.e. survey] estimates is above 5% in only three of the SHAs; in half, the difference is less than 1%."*	(Twigg et al., 2004)
DIAB	Construct	The synthetic estimates were compared to gender specific national estimates.	*"For total diabetes, the model estimates higher prevalence for women than for men. This is consistent with some major studies from the literature."*	(YHPHO, 2005)

METHODS

The validity assessments reported here cover four areas of validity: face, construct, convergent and predictive validity. The face validity assessments considered whether the models supporting the synthetic estimates are well-fitted. The remaining assessments consider whether the synthetic estimates accurately describe patterns in the prevalence of behavioural risk factors for CHD. The validity terms have been generally defined elsewhere (Last, 2001). For the purpose of this chapter, the terms refer to the following:

- Face validity: The extent to which the models supporting the synthetic estimates are well-fitted - that is, they adequately describe the relationship between the risk factor and individual-level and area-level covariates.
- Construct validity: The extent to which the synthetic estimates satisfy a theoretical construct (that synthetic estimates aggregated to the level of Government Office Region (GOR) concur with those produced by a survey designed to produce accurate estimates at this level).
- Convergent validity: The extent to which the synthetic estimates correlate with external measures of the same attribute, in this case with small-area behavioural risk factor prevalence estimates produced by local health surveys.
- Predictive validity: The extent to which the synthetic estimates predict an external criterion of the behavioural risk factors, in this case ward-level CHD mortality rates.

Face validity

The face validity of the synthetic estimates was assessed by investigating the models that generate the estimates. The following areas were investigated, and are described below:

- Model variables
- Bias
- Heteroskedasticity
- Spatial grouping of residuals

Model variables

The individual-level variables that are included in each of the synthetic estimation models were examined to assess whether the relative size and sign (positive / negative) of the associated coefficients is in accord with national surveys that consider the determinants of behavioural risk factors for CHD. For example, it is well established that average blood pressure levels increase with age (Department of Health, 2004) – the coefficients for the categorical age variables in the BP01 and BP03 models should therefore increase as the age categories increase. Deviations from well established relationships between the individual-level variables and the modelled behaviour were considered a sign of invalidity of the estimates.

Bias

Presence of bias was assessed by comparing the synthetic estimates with survey-based estimates of the prevalence of behavioural risk factors drawn from the HSfE 2000 to

2002. The survey estimates are (largely) unbiased[1], but underpowered for small areas. Therefore, if the synthetic estimates are unbiased then the regression line of the survey-based estimates on the synthetic estimates will be linear and equivalent to the y = x line (albeit with large variance around the regression line). Significant deviations of the regression line from y = x were considered indications of invalidity.

Data on the LA of residence for each respondent of the 2000, 2001 and 2002 HSfEs (supplied by the National Centre for Social Research specifically for this thesis) were added to a combination of the 2000 to 2002 datasets. The combined years dataset was used to derive survey estimates of the prevalence of behavioural risk factors for LAs in England. The ward level synthetic estimates were aggregated to LA level for the analysis. No data were collected on blood cholesterol levels in the 2000, 2001 and 2002 Health Surveys for England; consequently it was not possible to include the CHOL01 and CHOL03 models in the bias assessment (and also assessments for heteroskedasticity and spatial grouping of residuals, which also use the survey estimates generated from the combined years dataset).

[1] The clustered random sample of the HSfE series uses postcode sectors as the primary sampling unit. The survey estimates are therefore not representative of the LAs, as they are based on a sample drawn only from the intersection of the sampled postcode sectors and the LA of interest. Therefore individual survey estimates are likely to be biased. However, since the postcode sectors are sampled at random any positive or negative bias introduced for individual LAs is likely to be evened out when the entire set of LAs are considered. The whole set of survey estimates is therefore likely to be largely unbiased (Heady et al., 2003).

Because of the sampling frame of the HSfE, respondents were only drawn from those LAs that contain postcode sectors that were randomly sampled. The assessment of bias was therefore necessarily conducted on a subset of LAs in England. Since the different risk factors were not all included in each of the 2000, 2001 and 2002 surveys (for example, respondents were asked whether or not they smoked in all three surveys, whereas the in-depth questions used to assess physical activity levels were only included in the 2002 survey) the assessment of bias (and also heteroskedasticity and spatial clustering) were conducted on a different number of LAs for different sets of synthetic estimates.

Heteroskedasticity

If the synthetic estimation model is properly fitted then there should be no relationship between the size of the synthetic estimates and the size of the variance around the estimates: in other words, a plot of residuals against synthetic estimates should have a best fit line of $y = 0$ and have equal variance along this line. If the size of the variance increases or decreases as the size of the synthetic estimates increases then the model is said to be heteroskedastic, which is a sign of model mis-specification, and usually occurs when linear models have been chosen to describe a non-linear relationship, or important explanatory variables have not been included in the model.

The 'residuals' that were used to test the heteroskedasticity of the models were defined as the synthetic estimate minus the survey estimate (as described above). Because of this, the plot of residuals against synthetic estimates is a translation of the plot of synthetic

estimates against survey estimates around the y = x line. Therefore, the fit line of the residuals against synthetic estimates plot will be y = 0 if and only if the fit line of the synthetic estimates against survey estimates plot is y = x (which is assessed in the test for bias). Therefore the heteroskedasticity test only assessed whether the residuals have equal variance for all values of the synthetic estimates (and not for deviations of the best fit line from y = 0). As with the test for bias, the heteroskedasticity test was conducted at the LA level.

Spatial grouping of residuals

If the synthetic estimation models were correctly fitted, then the residuals should be randomly distributed geographically. Deviation from this random distribution was assessed by calculating the global Moran's I statistic for each set of residuals. A global Moran's I statistic that was significantly greater than zero was considered a sign of invalidity. The threshold for statistical significance was set at p = 0.01, since the large number of local authorities included in the assessments means that a small degree of spatial autocorrelation would achieve statistical significance at the p = 0.05 threshold. In addition, maps of the residuals were produced and examined for evidence of systematic clustering. As with the tests for bias and heteroskedasticity, the spatial grouping of residuals test was conducted at the LA level.

Construct validity

The following theoretical construct was used to assess the construct validity of the synthetic estimates:

Construct: The synthetic estimates of the prevalence of behavioural risk factors aggregated to the GOR-level (n = 9) should be ranked in approximately the same order as survey-based estimates drawn from the HSfE series.

The construct was set at the GOR-level since the HSfE sampling frame is designed to produce representative and unbiased estimates of behavioural risk factor prevalence rates at this level. In order to test the construct, the synthetic estimates were aggregated to GOR-level. GOR-level estimates of the prevalence of risk factors were derived using a combination of the HSfE general population datasets from 1999 to 2003 inclusive (details of the variables used are given in appendix two). The derived GOR-level estimates were neither gender-stratified nor age-standardised, so that they were comparable with the synthetic estimates.

The synthetic estimates were considered to have achieved construct validity if the ranking of the GORs by the synthetic estimates matched a ranking of the survey estimates that was achievable if the survey estimates were allowed to vary across their 95% confidence intervals. The 95% confidence intervals for the survey estimates were calculated in the standard way for a proportion (Altman, 1991). In addition, the rank correlation between the synthetic estimates and the survey estimates was calculated.

Convergent validity

Convergent validity was assessed by comparing the synthetic estimates with external measurements of the same target variable. In this case, the external measurements were made by a local health survey that was designed to estimate the prevalence of behavioural risk factors for CHD at LA-level (WMPHO, 2006), and a local health survey which has provided (under-powered) prevalence estimates at ward-level (NEPHO, 2002).

The external measurement is not required to be a gold standard, and it is therefore sufficient to show merely an association between the measurement of interest and the external measurement of the target variable. The outcome used here was the Pearson correlation coefficient (Altman, 1991): if a set of synthetic estimates showed a positive correlation with the survey-based estimates that was significantly different to zero then it achieved convergent validity.

The West Midlands Regional Lifestyle Survey 2005 (WMPHO, 2006) – which provided data for the LA-level comparison - estimated the prevalence of binge drinking, obesity, low fruit and vegetable consumption, moderate physical activity and smoking for all LAs in the West Midlands government office region (n = 34). All data were collected by self-completed questionnaire. Individuals were selected using a stratified sampling frame; the strata were LAs and deprivation quintiles (defined using the Index of Multiple Deprivation 2004 (ODPM, 2004)). The survey received a total of 54,773 responses and

an overall response rate of 32%. Response rates for the LAs ranged from 27% (Sandwell) to 39% (Wychavon). Data were weighted for non-response.

The Durham and Darlington Health and Lifestyle Survey (NEPHO, 2002) – which provided data for the ward-level comparison – estimated the prevalence of low fruit and vegetable consumption, low physical activity, drinking above weekly recommendations, smoking, obesity, raised blood pressure and diabetes. All data were collected by self-completed questionnaire. The randomised sample for the survey was drawn from the GP registration system and was stratified by primary care trust. The survey received a total of 8,630 responses and an overall response rate of 29%. The average number of respondents per ward was 49. Prevalence estimates have been provided for all wards within the Durham and Darlington primary care trusts (n = 116).

Predictive validity

For the synthetic estimates to achieve predictive validity they must be able to predict an external criterion of the phenomenon under study – in this case, mortality rates for CHD. The correlations for the predictive validity analysis used ward-level age-standardised gender specific synthetic estimates and the CHD mortality rates used in the analyses described in chapter five.

Assessments of predictive validity are conducted on longitudinal datasets, where the measure under assessment is tested for association with an outcome after a certain

amount of follow-up. If the measure under assessment is an aggregate measure (as is the case here) then the longitudinal dataset should follow the progress through time of the cohorts that the measure is aggregated to. The ideal assessment of predictive validity for the synthetic estimates would therefore be to test for association between the synthetic estimates for wards and the CHD mortality rate since 2001 of all people who were aged 16 or over in 2001, by residence of ward in 2001. Unfortunately, these data are not available. A weaker alternative approach (which is used here) is to test for association between the synthetic estimates and the CHD mortality rate for wards using data on deaths collected between 1999 and 2004 (i.e. centred on 2001). This is described here as an assessment of predictive validity under the assumptions that a) the synthetic estimates of risk factor prevalence rates are reasonable proxies of past prevalence rates, and b) that the influence of migration between wards is negligible.

The cross-sectional nature of the assessment and the problems involved with the assumptions imply that this validity assessment has more in common with convergent validity testing than predictive validity testing. The synthetic estimates were therefore assessed for validity in the same way: synthetic estimates were assumed to have achieved validity if they were positively associated with CHD mortality rates (for both men and women separately), and the correlation coefficient was significantly different to zero. In contrast to the convergent validity assessment, the synthetic estimates were all assessed against the same data. Since the synthetic estimates are all measured on the same scale (namely a prevalence rate of between 0 and 100%) it was possible to directly compare the coefficients of correlation between the different estimates and CHD mortality, allowing

for a further test of validity: synthetic estimates with a correlation coefficient that is unrealistic in terms of the medical literature were considered invalid (e.g. if the correlation coefficient for any of the set of synthetic estimates is, say, twenty times higher than the other coefficients).

RESULTS

Face validity

Model variables

Table 6.3 provides details of the variables included in the synthetic estimation models, the sign of the coefficient involved for each, and whether or not interaction terms were included in the modelling. With the exception of the DIAB model (which does not have a baseline category), the models all used three individual-level categorical variables and the coefficients reflect the increase (or decrease) in odds of uptake of the behaviour associated with change in characteristics from the baseline category. All of the models used sex and age as individual-level variables. The third individual-level variables that were used were ethnicity, social class or marital status. The baseline categories were always young and female, and depending on the third individual-level variable selected were white, low social class or not single.

Table 6.3 Variables included in the synthetic estimation models, signs of the coefficients and use of interaction terms

Individual-level variables	***FV 01***	***FV 03***	***PA 01***	***PA 03***	***ALC***	***SMOK 00***	***SMOK 04***	***SMOK 01***	***SMOK 03***	***OBES 01***	***OBES 03***	***BP 01***	***BP 03***	***CHOL 01***	***CHOL 03***	***DIAB***
Male	+	+	–	–	+	+	+	+	+	–	–	+	+	–	–	–
Increasing age	–	–	+	+	–	–	–	–	–	+	+	+	+	+	+	+
Single	.	.	.	.	+	+	+	.	.	.	.	.	.	.	.	.
Black	.	+	.	+	.	.	.	.	.	.	.	.	.	.	.	+
South Asian	.	+	.	+	.	.	.	.	.	.	.	.	.	.	.	+
Chinese	.	–	.	+	.	.	.	.	.	.	.	.	.	.	.	Equal
High social class	–	.	+	.	.	.	.	–	–	–	–	–	–	–	–	.
Area-level variables																
% high social class	–	.	.	+	+	.	–	–	–	–	–	–	–	.	.	.
% households with 2+ cars	.	.	.	.	+	–	.	.	.	.	.	.	.	.	.	.
% households with 6+ rooms	.	.	.	.	.	.	–	.	.	.	.	.	.	.	.	.
% households with no car	.	.	.	.	.	+	.	.	.	.	.	.	.	.	.	.
% income support recipient	.	.	–	.	.	.	.	+	+	.	.	.	.	.	.	.
% privately renting	.	.	.	.	+	+	+	.	.	.	.	.	.	.	.	.
% males economically inactive	.	.	.	.	.	.	–	.	.	.	.	.	.	.	.	.
Increasing deprivation	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	+
% Asian population	–	.	.	.	.	.	.	–	–	–	–	.	.	.	.	.
% Black population	.	.	.	.	.	.	.	.	.	+	.	.	.	–	.	.
% non-white population	.	.	.	.	.	.	–	.	.	.	.	.	.	.	.	.
% households with dependent children	.	.	.	.	.	.	+	.	.	.	.	.	.	.	.	.
% living alone	–	.	.	.	.	.	.	.	.	–	–	.	.	.	.	.
Includes interaction terms	NO	NO	NO	NO	YES	YES	YES	NO	NO	NO	NO	NO	NO	NO	NO	YES

+ = positive coefficient; - = negative coefficient; . = variable was not included in the logistic regression model.

The area-level variables that are included in the various models could be split into measures of deprivation (e.g. % high social class, % households with 6+ rooms) and cultural measures (e.g. % Asian population, % living alone). None of the models that included area-level deprivation measures did so in such a way that would appear contradictory (e.g. by having positive coefficients for measures of low and high deprivation in the same model) apart from the ALC model ('% high social class', '% households with 2+ cars' and '% privately renting' all positive).

In general, the relationships between the individual-level variables and the different risk factor prevalence models are in line with data from national surveys, but this is not the case for all of the models. The Expenditure and Food Survey of 2004/05, for example, suggests that fruit and vegetable intake is higher amongst those of Asian ethnicity and lower amongst those of Chinese ethnicity than the White population, contradicting the FV03 model (Office for National Statistics, 2006b). The DIAB model suggests that the prevalence of diabetes is lower for men than women for all ethnicities except the Asian population – this contradicts the results from the HSfE for both diagnosed and undiagnosed diabetes (Department of Health, 2004).

The HSfE 2003 suggests that social trends in physical inactivity are different for men and women; men in higher social classes are more likely to be physically inactive than men in lower social classes, whilst this social pattern is reversed in women (Department of Health, 2004). These differing patterns should be modelled by introducing an interaction term between gender and social class; such a term is absent from the PA01 model. The

social trend in cholesterol levels for both men and women is not linear – the prevalence of raised cholesterol in higher and intermediate social classes is similar, but the prevalence in lower social classes is higher (Department of Health, 2004). This non-linear trend is reasonably modelled by both CHOL01 and CHOL03, where the individual social class variable included in the models is 'income support recipient'.

Bias

Table 6.4 describes how the regression lines of survey-based estimates on synthetic estimates deviate from the y = x line (i.e. intercept of zero, and coefficient of one). Here, the 'Intercept' and 'Coefficient' columns describe the regression line through the survey-based estimates and the synthetic estimates. If, for a set of synthetic estimates, the intercept is not statistically different from zero and the coefficient is not statistically different from one (indicated in both instances by the 95% confidence intervals), then the regression line is not statistically different from the y = x line, and hence there is no strong evidence of bias.

In some cases (FV03, PA01, PA03, ALC, DIAB) the synthetic estimates were uncorrelated with the survey-based estimates. The regression lines for these synthetic estimates were not found to be statistically different to the y = x line, but this was only a result of large confidence intervals around the regression parameters. Of the remaining synthetic estimates, only two of the smoking models (SMOK00, SMOK03), both the obesity models, and one blood pressure model (BP01) were found to be unbiased. The variance of each of the sets of synthetic estimates was far smaller than the variance of the

survey estimates. The scatter plots of the synthetic estimates versus LA-level HSfE estimates are shown in figure 6.1a-n, appendix two.

Table 6.4 Intercept, coefficient and 95% confidence intervals for the regression lines of survey-based estimates versus synthetic estimates (local authorities)

Synthetic estimates	n	Intercept*	95% confidence interval - intercept	Coefficient*	95% confidence interval - coefficient	r^2
FV01	345	-0.59	(-1.02, -0.15)	1.88	(1.26, 2.49)	0.09
FV03	345	0.46	(0.08, 0.84)	0.54	(-0.22, 1.29)	0.00
PA01	249	0.24	(-0.91, 1.39)	-0.02	(-2.42, 2.36)	0.00
PA03	249	-0.38	(-1.43, 0.68)	1.40	(-1.03, 3.83)	0.00
ALC	346	0.13	(-0.05, 0.31)	0.46	(-0.38, 1.30	0.00
SMOK00	346	-0.02	(-0.08, 0.04)	0.95	(0.73, 1.17)	0.17
SMOK04	346	-0.10	(-0.19, -0.01)	1.62	(1.22, 2.02)	0.15
SMOK01	346	-0.25	(-0.38, -0.11)	1.56	(1.12, 2.00)	0.12
SMOK03	346	-0.09	(-0.19, 0.01)	1.30	(0.91, 1.69)	0.11
OBES01	345	0.05	(-0.01, 0.11)	1.07	(0.66, 1.48)	0.07
OBES03	345	-0.05	(-0.15, 0.05)	1.14	(0.69, 1.59)	0.06
BP01	345	-0.02	(-0.06, 0.02)	1.39	(0.93, 1.85)	0.09
BP03	345	-0.04	(-0.09, 0.01)	1.64	(1.11, 2.17)	0.10
DIAB	345	0.03	(0.00, 0.06)	0.44	(-0.30, 1.17)	0.00

* The y = x line has intercept 0 and coefficient 1.
Assessments conducted at local authority level.

Heteroskedasticity

The scatter plots of residuals against synthetic estimates are provided in appendix two (figures 6.2a-n). An assessment of the plots suggests that the synthetic estimates do not display heteroskedasticity, with the possible exception of FV03 and SMOK00 where the variance of the residuals slightly increases for larger estimates.

Spatial grouping of residuals

Table 6.5 shows the degree of spatial autocorrelation of the residuals for each set of synthetic estimates.

Table 6.5 Spatial autocorrelation (measured by global Moran's I statistic) of residuals for each set of synthetic estimates (local authorities)

Synthetic estimates	n	Global Moran's I statistic	p
FV01	345	0.08	0.014
FV03	345	0.11	0.001
PA01	249	-0.06	0.139
PA03	249	-0.06	0.104
ALC	346	0.09	0.006
SMOK00	346	-0.01	0.480
SMOK04	346	0.02	0.225
SMOK01	346	-0.00	0.143
SMOK03	346	0.01	0.376
OBES01	345	0.00	0.076
OBES03	345	-0.00	0.436
BP01	345	0.15	0.001
BP03	345	0.15	0.001
DIAB	345	0.05	0.080

Assessments conducted at local authority level.

Only the FV03, BP01, BP03 and ALC synthetic estimates showed a significantly positive spatial autocorrelation of residuals ($p < 0.01$). Maps of these residuals are shown in figure 6.3a-n in appendix two, and these largely confirm the absence of spatial clustering of the residuals with the exception of the residuals for prevalence of blood pressure which show some clustering of high residuals in the North of England, suggesting that the synthetic estimates of blood pressure may systematically over-estimate prevalence rates in the North.

Construct validity

Table 6.6 shows the prevalence rates of behavioural risk factors for CHD by government office region (GOR) that have been generated by combining data from the 1999 to 2003 HSfEs. 'Allowable ranking' refers to all of the ranks that a GOR could occupy if the survey estimates were allowed to vary over the 95% confidence intervals of the prevalence estimates. For example, the prevalence of problem drinking in East Midlands is significantly lower than the prevalence in the North East, North West and Yorkshire and the Humber (and hence the East Midlands could not be ranked any higher than 4 of the 9 GORs), but significantly higher than the prevalence in London (and hence could not be ranked any lower than 8 of the 9 GORs). There is no significant difference in the prevalence rate between the East Midlands, West Midlands, East of England, South East and South West. Therefore, the allowable ranking for the East Midlands is 4-8.

Table 6.6 Prevalence of behavioural risk factors for CHD by Government Office Region, HSfE data, 1999-2003 (years of data collection in brackets)

Government Office Region	Smoking (99-03) Prev %	Smoking (99-03) Allowable ranking	Fruit and veg consumption (01-03) Prev %	Fruit and veg consumption (01-03) Allowable ranking	Physical inactivity (99; 02-03) Prev %	Physical inactivity (99; 02-03) Allowable ranking
North East	29.0	1-3	81.2	1-3	42.0	1-9
North West	27.8	1-4	78.3	2-6	40.1	2-9
Yorks & Humber	28.0	1-4	78.8	1-5	40.4	1-9
East Midlands	25.5	4-9	76.5	3-7	42.1	1-8
West Midlands	24.8	4-9	76.9	2-7	43.8	1-5
East of England	25.2	4-9	75.6	4-7	39.8	2-9
London	26.1	2-9	72.2	8-9	40.8	1-9
South East	24.4	4-9	72.5	8-9	38.4	3-9
South West	25.0	4-9	77.4	2-7	39.5	2-9

Government Office Region	Blood pressure (99-03) Prev %	Blood pressure (99-03) Allowable ranking	Cholesterol (99; 03) Prev %	Cholesterol (99; 03) Allowable ranking	Obesity (99-03) Prev %	Obesity (99-03) Allowable ranking
North East	10.4	1-4	30.8	1-4	17.9	1-6
North West	7.8	2-8	21.8	2-9	16.7	1-7
Yorks & Humber	8.9	1-5	23.2	2-8	17.3	1-7
East Midlands	7.2	4-8	23.6	2-8	18.1	1-6
West Midlands	8.5	1-7	26.8	1-8	18.6	1-4
East of England	6.9	5-9	23.3	2-8	16.5	2-7
London	5.6	8-9	17.2	8-9	14.7	7-9
South East	7.4	4-8	25.0	1-8	14.9	7-9
South West	9.3	1-4	26.6	1-8	15.6	4-9

Government Office Region	Problem drinking (99-02) Prev %	Problem drinking (99-02) Allowable ranking	Diabetes (99-03) Prev %	Diabetes (99-03) Allowable ranking
North East	26.8	1-3	2.6	1-9
North West	27.3	1-3	2.0	3-9
Yorks & Humber	28.1	1-3	2.3	1-9
East Midlands	23.0	4-8	2.4	1-9
West Midlands	21.6	4-9	2.6	1-7
East of England	21.6	4-9	2.7	1-7
London	20.3	6-9	2.5	1-9
South East	23.5	4-8	2.0	3-9
South West	21.9	4-9	2.3	1-9

Definitions of variables provided in appendix two.

Table 6.7 displays the rank correlation between the synthetic estimates at GOR-level and the associated HSfE GOR estimates, and displays whether the ranking of the synthetic estimates matches an allowable ranking of the survey estimates.

Table 6.7 Rank correlation between synthetic estimates and associated HSfE based prevalence estimates, and achievement of ranking allowed by HSfE data (Government Office Regions, n = 9)

Synthetic estimates	Spearman's rho	p	Achieve the allowable ranking?
FV01	0.93	<0.001	YES
FV03	-0.33	0.381	NO
PA01	-0.72	0.030	NO
PA03	-0.52	0.154	NO
ALC	-0.67	0.050	NO
SMOK00	0.79	0.013	NO
SMOK04	0.88	0.001	YES
SMOK01	0.82	0.007	YES
SMOK03	0.73	0.025	YES
OBES01	0.85	0.004	YES
OBES03	0.83	0.005	YES
BP01	0.92	<0.001	YES
BP03	0.90	<0.001	YES
CHOL01	0.55	0.125	YES
CHOL03	0.65	0.058	YES
DIAB	0.27	0.488	NO

The rank correlation between the synthetic estimates generated by the SMOK00 model and the survey estimates for smoking was high (rho = 0.79), but the ranking of the synthetic estimates was incompatible with the survey estimates, as the synthetic estimates predicted that the prevalence of smoking is highest in London (which has a significantly lower prevalence rate than the North East according to the HSfE data).

Convergent validity

The scatter plots of the synthetic estimates against the local health survey estimates are shown in figures 6.4a-k and figures 6.5a-n in appendix two. Each set of synthetic

estimates was assessed against LA-level estimates from the West Midlands Regional Lifestyle Survey and ward-level estimates from the Durham and Darlington Health and Lifestyle Survey, and the resultant correlation coefficients are shown in table 6.8. If the correlation coefficient is negative (indicating a negative association between the survey-based estimates and the synthetic estimates) or not statistically different from zero ($p < 0.05$) then the set of synthetic estimates displayed invalidity.

Table 6.8 Correlation coefficients (Pearson's r) for correlation between synthetic estimates and estimates derived from the West Midlands Regional Lifestyle Survey 2005 (local authorities, n = 34), and the Durham and Darlington Health and Lifestyle Survey 2002 (wards, n= 116)

	West Midlands Regional Lifestyle Survey *		Durham and Darlington Health and Lifestyle Survey	
Synthetic estimates	**Pearson's r**	**p**	**Pearson's r**	**p**
FV01	0.52	0.002	-0.46	<0.001
FV03	0.60	<0.001	-0.02	0.861
PA01	-0.84	<0.001	-0.09	0.356
PA03	-0.32	0.061	-0.04	0.689
ALC	-0.06	0.752	-0.14	0.143
SMOK00	0.57	<0.001	0.60	<0.001
SMOK04	0.72	<0.001	0.59	<0.001
SMOK01	0.66	<0.001	0.62	<0.001
SMOK03	0.66	<0.001	0.63	<0.001
OBES01	0.75	<0.001	0.26	0.005
OBES03	0.66	<0.001	0.36	<0.001
BP01	-	-	0.29	0.002
BP03	-	-	0.27	0.004
DIAB	-	-	0.20	0.037

* Assessments conducted at local authority level.

Only the PA01, PA03 and ALC synthetic estimates failed to achieve a significantly positive association with the LA-level local survey estimates. For ward-level estimates, the FV01, FV03, PA01, PA03 and ALC synthetic estimates were not positively correlated

with the survey estimates. In every case the range of the synthetic estimates was smaller than the range of the local survey estimates. In some cases the difference was small (the SMOK00 synthetic estimates ranged from 24% to 34%, compared to the local authority survey smoking estimates which ranged from 11% to 27%), whereas in some cases the difference was substantial (the ALC synthetic estimates ranged from 20% to 23%, compared to the local authority binge drinking estimates that ranged from 18% to 33%).

Predictive validity

Table 6.9 provides details of the strength of association between the age-standardised synthetic estimates and the age-standardised CHD mortality rate at ward-level, measured separately by gender (n = 7,929). The 'beta' column is the regression coefficient, indicating the strength of the association between the sets of synthetic estimates and CHD mortality rates. The 'r' column displays the correlation coefficient between the synthetic estimates and the CHD mortality rates, and the p values display the degree of statistical significance of the correlation. Scatter graphs of the relationship are shown in figure 6.6a-q in appendix two.

Table 6.9 Regression coefficient (beta) and correlation coefficient (r) between age-standardised synthetic estimates and age-standardised CHD mortality rate per 100,000 (wards, n = 7,929)

	MEN			WOMEN		
Synthetic estimates	**Beta**	**r**	**p**	**Beta**	**r**	**p**
FV01	8.7	0.38	<0.001	4.4	0.38	<0.001
FV03	19.3	0.16	<0.001	8.5	0.15	<0.001
PA01	-17.2	-0.51	<0.001	-8.3	-0.46	<0.001
PA03	1.2	0.03	0.011	-0.0	0.00	0.966
ALC	-18.7	-0.36	<0.001	-8.7	-0.42	<0.001
SMOK00	5.5	0.49	<0.001	2.3	0.38	<0.001
SMOK04	6.9	0.53	<0.001	2.8	0.38	<0.001
SMOK01	10.0	0.50	<0.001	5.1	0.46	<0.001
SMOK03	8.4	0.52	<0.001	4.2	0.47	<0.001
OBES01	14.7	0.45	<0.001	7.1	0.45	<0.001
OBES03	16.2	0.52	<0.001	8.1	0.48	<0.001
BP01	44.2	0.47	<0.001	24.6	0.44	<0.001
BP03	75.5	0.50	<0.001	41.5	0.47	<0.001
CHOL01	5.3	0.03	0.002	3.7	0.05	<0.001
CHOL03	497.5	0.49	<0.001	228.0	0.45	<0.001
DIAB	20.4	0.44	<0.001	9.8	0.40	<0.001

With the exception of the PA01, PA03 and ALC synthetic estimates, significant positive associations with the CHD mortality rates were shown. The positive beta coefficients were generally in the range of 1 to 80 for men and 2 to 40 for women, with the exception of the CHOL03 synthetic estimates where the beta coefficient was nearly 500 for men and nearly 230 for women. The beta coefficients represent the increase in mortality rate per 100,000 that would be expected for a 1% increase in prevalence of risk factor: a beta coefficient of 500 would therefore imply that an increased prevalence rate of 2% should result in an increase of 1,000 deaths per 100,000 for the ward. To place this in context, the highest male mortality rate for a ward in the dataset was 525 deaths per 100,000. These high correlation coefficients were considered a sign of invalidity.

Summary of results

The results described in this chapter cover different methods for assessing different aspects of validity of the synthetic estimates. The face validity assessments consider whether the logistic models used to describe the relationship between the risk factor and individual-level and area-level covariates are well-fitted. The construct, convergent and predictive validity assessments consider whether the resulting estimates are accurate, in so far as the existing evidence will allow. Although poorly fitted models are unlikely to result in accurate estimates, well-fitted models are not guaranteed to result in accurate estimates (since uptake of the modelled behaviour may have little association with the covariates included in the modelling process). Table 6.10 summarises the results of the various validity assessments described in this chapter.

Table 6.10 Achievement of different validity criteria

Synthetic estimates	Face validity – model variables	Face validity - bias	Face validity – heterosked-asticity	Face validity – spatial clustering of residuals	Construct validity	Convergent validity	Predictive validity
FV01	✓	x	✓	✓	✓	x	✓
FV03	x	x	x	x	x	x	✓
PA01	x	x	✓	✓	x	x	x
PA03	✓	x	✓	✓	x	x	x
ALC	✓	x	✓	x	x	x	x
SMOK00	✓	✓	x	✓	x	✓	✓
SMOK04	✓	x	✓	✓	✓	✓	✓
SMOK01	✓	x	✓	✓	✓	✓	✓
SMOK03	✓	✓	✓	✓	✓	✓	✓
OBES01	✓	✓	✓	✓	✓	✓	✓
OBES03	✓	✓	✓	✓	✓	✓	✓
BP01	✓	✓	✓	x	✓	✓	✓
BP03	✓	x	✓	x	✓	✓	✓
CHOL01	✓	N/A	N/A	N/A	✓	N/A	✓
CHOL03	✓	N/A	N/A	N/A	✓	N/A	x
DIAB	x	x	✓	✓	x	✓	✓

x: Synthetic estimates showed sign of invalidity
✓: Synthetic estimates did not show sign of invalidity
N/A: Synthetic estimates not included in the assessment.

DISCUSSION

The synthetic estimates have shown mixed evidence of face, construct, convergent and predictive validity (albeit with a weak test for predictive validity). Only the Health Poverty Index synthetic estimates for obesity (both the 2001 and 2003 versions) and the 2003 Health Poverty Index synthetic estimates for smoking showed evidence of validity in all of the assessments reported here. The remaining sets of synthetic estimates showed evidence of invalidity in at least one of the assessments.

The FV03, PA01 and DIAB synthetic estimates are based on logistic regression models that do not adequately describe the relationship between the individual-level variables and the risk factors. As a result, the models are mis-specified and generate inaccurate estimates. For the FV03 and PA01 synthetic estimates, this is due to non-inclusion of important interaction terms. For the DIAB synthetic estimates, this is because the model is based on two surveys conducted in the early 1990s from Coventry and Brent (Simmons et al., 1991; Chaturverdi et al., 1993), the results of which do not appear to be representative of the situation in England around 2001.

The ALC and PA03 synthetic estimates did not show any evidence of model mis-specification, but showed little agreement with HSfE or local survey estimates. This may be due to the non-standard definitions used to define ‘physical inactivity’ and ‘problem drinking’ (see table 6.2). As a result, the comparisons with HSfE and local survey

estimates were based on similar but different definitions, and should be viewed with caution.

The two sets of cholesterol synthetic estimates developed for the Health Poverty Index (CHOL01 and CHOL03) showed evidence of invalidity in the predictive validity assessment. This is because the age and sex variables dominate the related logistic models, so there is very little variance in the gender stratified age-standardised synthetic estimates, which leads to the over-inflated estimates of the impact of raised cholesterol rates on CHD mortality rates. The finding that raised cholesterol levels do not vary much in England after adjustment for sex and age may well be genuine: the prevalence of raised cholesterol does not vary much by either social class or GOR in England (Department of Health, 2004).

There was only weak evidence of invalidity for the two sets of smoking synthetic estimates developed by Twigg and colleagues (SMOK00 and SMOK04), and the 2001 Health Poverty Index smoking synthetic estimates. The SMOK00 synthetic estimates appeared to over-estimate the prevalence of smoking in London, which could be a problem of mis-specification of the related logistic model (for example, a dummy variable indicating residence in London could be included), or it could be due to random variation in the HSfE data which were used for comparison. The other two sets of smoking synthetic estimates showed evidence of a slight bias, the nature of which was to over-estimate low prevalence rates and under-estimate high prevalence rates. As a result, the synthetic estimates for individual areas are likely to be inaccurate but the sets of

estimates for all areas accurately describe the geographic trend – albeit with a reduced variance. This bias is evident to some extent in all of the sets of synthetic estimates that showed some correlation with survey-based estimates (indicated by a slope coefficient greater than 1 in table 6.3), except the SMOK00 synthetic estimates. A previous assessment of the bias of synthetic estimates at ward-level concurred with the findings reported here: that synthetic estimates tend to under estimate for areas with high prevalence rates, and over estimate for areas with low prevalence rates (Heady et al., 2003). Conversely, the synthetic estimates developed by the National Centre for Social Research showed very little signs of bias when compared against ward-level HSfE estimates (Pickering et al., 2005). The development of these estimates was aided by access to data on the ward of residence for HSfE respondents, and also population-level data from non-census datasets (e.g. the proportion of residents in a ward who claim benefits). The additional data available to the researchers is likely to have resulted in better-fitted logistic models, which would reduce the degree of bias displayed in these analyses. Future attempts at synthetic estimation would be aided by access to such data.

The various analyses that are reported in this chapter have utilised statistically powerful datasets from combined years of the HSfE series. They allow for comparable validity assessments for sets of synthetic estimates designed by different researchers for different purposes. Many of the sets of synthetic estimates that are included in this chapter have not previously undergone any validation assessments of this kind. However, it has not been possible to conduct certain analyses on model mis-specification that are described elsewhere in the synthetic estimation literature; for example, testing for the amount of

between-area variance that is explained by the models, and testing for the goodness-of-fit of the models (Twigg et al., 2004). Such tests would require data on the ward of residence of HSfE respondents and it has not been possible to obtain such data.

The convergent validity assessments relied upon data gathered from local health surveys, which for pragmatic reasons could not replicate the data collection methods of the HSfE. Both local surveys relied on a mail-out of questionnaires, as compared to the HSfE method of interviewers visiting sampled households. As a result, the response rates for the local surveys were far lower than that of the HSfE (around 30% as compared to 60-70%), and data collection was not standardised. Differential response by age, sex and social class is likely to be higher for the local surveys, which would affect the aggregated responses for local areas. Further, the definitions of the variables in the local surveys were not always the same as for the HSfE, particularly for physical inactivity (local survey definition: prevalence of individuals failing to achieve 30 minutes physical activity on at least 5 days a week; PA01, PA03 definition: prevalence of individuals failing to achieve five hours of physical activity per week) and problem drinking (West Midlands Health Survey definition: prevalence of individuals exceeding daily benchmark for heavy drinking; ALC definition: prevalence of individuals exceeding weekly benchmark for safe drinking). Similarly, data for obesity and hypertension were generated by direct measurement in the HSfE, but by self-completed questionnaire in the local surveys. Despite this, the degree of correlation between the synthetic estimates and the local survey estimates was generally high, and results using the two different local surveys were generally comparable. This is reassuring as it suggests that the synthetic

estimates are consistent in two different locations in England. The fruit and vegetable synthetic estimates (FV01 and FV03) are the exceptions to this finding - reasonable correlations were found with the West Midlands local survey, but not with the Durham and Darlington local survey. This may be due to the low number of respondents within each ward in the Durham and Darlington local survey – the scatter graphs (figures 6.4a and b, appendix two) reveal that the majority of wards had a survey estimate of around 100% of people not eating five portions of fruit and vegetables today. Because of this, the correlations are susceptible to large deviation by the presence of outliers.

The assumptions required for the predictive validity assessment (regarding current prevalence levels being good proxies of past prevalence rates, and levels of migration between wards) are unlikely to be sound. Regarding the first of the assumptions, large scale geographic trends in smoking, alcohol consumption and poor diet have been fairly consistent in England over the last ten years (Allender et al., 2008), but it is unclear whether these geographic trends are stable at smaller geographic levels. Trend data for the prevalence of risk factors for CHD at ward level is not readily available. One source which allows the examination of trends at ward level is the South Tyneside Health and Lifestyle Survey (Snowdon et al., 2004; Tyler et al., 1995). Here a standard self-reported health questionnaire was used to collect data on behavioural risk factors for CHD, first in 1992 and then again in 2003. The data were examined at ward-level, for all wards in the South Tyneside LA (n = 20), and prevalence rates for various health-related behaviour were reported in quartiles. Comparable results for the prevalence of people who have never smoked, and for the prevalence of self-reported overweight (BMI>=25kg/m^2) are

reported for both the 1992 and 2003 surveys. When the two sets of results were compared, they achieved rank correlation scores of 0.22 and 0.40 for the two health behaviours respectively, suggesting that agreement between the prevalence rates in 1992 and 2003 was not high. But these results are based on a small number of comparisons and there was a substantial difference in the response rates for the two survey years (69% in 1992, 27% in 2003).

Routinely collected data on internal migration within the UK suggest that the second assumption is also unlikely to be sound. Four per cent of people within the UK moved residence to a new location over 10km away in 2000, suggesting that the rate of migration between wards is at least one in 25 per year (Office for National Statistics, 2007a). This estimated rate excludes migration to and from England, which would increase the degree of flux for wards.

The different validation techniques reported here all rely on the assumed accuracy of survey-based estimates of the prevalence of behavioural risk factors for CHD, which varies from risk factor to risk factor. For example, the HSfE collects data on fruit and vegetable consumption, alcohol consumption, levels of physical activity, smoking and diabetes from a questionnaire completed by a trained interviewer, who also collects height and weight measurements (for the obesity data). A nurse visit is then conducted to obtain data on blood pressure, and a blood sample is obtained for measurements of cholesterol levels. In 2003, the response rate for the questionnaire was 66%, for the height/weight measurements was 60%, for blood pressure measurements was 50%,

whereas the response rate for the blood sample was only 40% (Department of Health, 2004). The low response rates for the blood pressure and cholesterol measurements could have introduced a bias if the non-response was differential (by ethnicity, for example). Therefore, although the synthetic estimates of the prevalence of raised blood pressure and raised cholesterol showed signs of validity in the comparisons with survey-based data reported in this chapter, they may still not reflect the actual geographic variation in prevalence rates if survey-based estimates are not accurate reflections of the true geographic pattern.

The results reported here concur with a previous assessment of the convergent validity of the SMOK00 and ALC synthetic estimates (Twigg et al., 2002). In the previous assessment the synthetic estimates were compared with ward-level estimates of similar variables generated by local health surveys in North London, Portsmouth and Wales. As reported here, the SMOK00 synthetic estimates showed a high degree of association with the local survey estimates, whereas the ALC synthetic estimates did not show a significant positive association with the local survey estimates.

The synthetic estimates of the prevalence of smoking and obesity that have been examined here (Twigg et al., 2000; Twigg et al., 2004; Dibben et al., 2004) have been shown to describe geographic variations at small area levels accurately, albeit with reduced variance. The reduced variance implies that care must be taken in interpreting the results of ecological analyses where the synthetic estimates are used as explanatory variables, as is the case in the following chapter. As shown by the results of the predictive

validity assessment, using the synthetic estimates in such a way could result in an over-estimation of the impact of the risk factor on the ecological phenomenon under investigation.

CONCLUSIONS

The sets of synthetic estimates that were deemed valid and accurate enough to be used as explanatory variables in chapter seven were the smoking estimates developed for *The Smoking Epidemic in England* (Twigg et al., 2004), and the 2001 estimates of fruit and vegetable consumption, obesity, hypertension and raised cholesterol developed for the Health Poverty Index website (Dibben et al., 2004). The smoking and obesity synthetic estimates only showed small signs of invalidity in the assessment of bias, which must be addressed in the interpretation of the results of the analyses in the following chapter. The fruit and vegetable consumption and hypertension synthetic estimates showed small signs of invalidity in the convergent validity and spatial clustering of residuals assessments respectively. The fruit and vegetable synthetic estimates produced a negative correlation with ward-level results from the Durham and Darlington Health and Lifestyle Survey, but the results at local authority level (where the number of survey respondents for each area was far higher) were more positive. The hypertension synthetic estimates achieved a very high rank correlation with the ordering of high blood pressure levels for Government Office Regions (Spearman's rho = 0.92), so it is unlikely that the clustering of the residuals has resulted in systematic over or under estimation in any region of the country.

The FV03, PA01, PA03, ALC, SMOK00, CHOL03 and DIAB synthetic estimates were not used in the analyses reported in chapter seven and eight as they displayed signs of invalidity in one or more of the analyses described here. The selected sets of synthetic estimates allow for reasonable coverage of individual-level risk factors for CHD, as shown in the conceptual framework in figure 2.1 (chapter two). Elements of diet that have not been included (saturated fat intake, energy intake, salt intake) are covered by the medical conditions that they cause (namely raised cholesterol, obesity and raised blood pressure): to a lesser extent this is also true for physical inactivity and problem drinking. The unavailability of valid synthetic estimates for diabetes prevalence and high stress levels is problematic but unavoidable.

Chapter 7: The association between behavioural risk factor profiles of populations and coronary heart disease mortality and hospitalisation rates in England

INTRODUCTION

This chapter reports on analyses exploring the association between ward-level prevalence estimates of behavioural risk factors for coronary heart disease (CHD) and CHD mortality and hospitalisation rates. The analyses used similar techniques to those reported in chapter five, but were concerned with the geographic variation in CHD rates that is *compositional* rather than *contextual.* The sets of synthetic estimates that were explored in the previous chapter and found to be reasonably valid and accurate were used as explanatory variables in the analyses reported here. The analyses were designed to answer the following questions:

RQ1: Are ward-level synthetic estimates of the prevalence of behavioural risk factors for CHD appropriate for use as explanatory variables in regression analyses with CHD mortality and hospitalisation rates as outcome variables?

RQ2: How much of the large scale and small scale geographic variation in CHD mortality and hospitalisation rates can be explained by differences in the behavioural risk factor profiles of populations?

The relationships between behavioural risk factors for CHD and both CHD mortality and hospitalisation have been well-established at the individual-level (Stamler, 2005).

However, just because the established risk factors are powerful predictors of CHD at individual-level does not necessarily mean that they will be powerful predictors of the geographic variation of CHD: for example, if the prevalence of smoking was uniform for all wards in England then smoking would have no influence on the geographic variation of CHD in England despite the fact that the INTERHEART study estimates that smoking alone causes nearly 30% of myocardial infarctions in Western Europe (Yusuf et al., 2004). Indeed, the area-level relationship between the prevalence of risk factors and CHD rates can be surprising. The British Regional Heart Study found that geographic variation in CHD incidence rates in men were *negatively* associated with the prevalence of raised cholesterol, after other individual-level risk factors for CHD had been taken into account (Morris et al., 2001). The fact that the association between cholesterol levels and CHD has been shown to be different at the individual-level and the area-level illustrates the danger of interpreting results regarding this association using data collected at only one level – there is the potential for either 'ecological fallacy' or 'individualistic fallacy' (Subramanian et al., 2009). These dangers are discussed further in chapter nine.

The associations between prevalence rates of behavioural risk factors for CHD and CHD mortality and hospitalisation rates at the area-level are worthy of investigation. Since the models included in this chapter use ecological data (specifically aggregate data for both outcome and explanatory data) the interpretations of the model must also be restricted to the area-level. This means that the results can only provide information about the geographic variation in CHD rates in England, and not about the relationship between behavioural risk factors and CHD in individuals. The models built for this

chapter included only the behavioural risk factor profiles of populations as explanatory variables – a further exploration of geographic variation including behavioural risk factor profile of populations, environmental variables and deprivation is reported in chapter eight.

METHODS

The analyses that are reported here followed the same structure as those reported in chapter five. Initially, exploratory data analysis searched for collinearities amongst the explanatory variables. Then both multi-level and spatial error regression models of the four outcome variables (male and female CHD mortality rates, male and female CHD hospitalisation rates) were built to explore how much of the geographic variation in CHD rates can be explained by each of the individual explanatory variables (in univariate analyses) and the set of all explanatory variables (in combined multivariate analyses). In addition, principal components analysis (PCA) was conducted on the sets of synthetic estimates of male and female prevalence of risk factors for CHD in order to produce a set of orthogonal explanatory variables that are more suited to regression analyses. The technique of PCA is described in chapter four. The PCA was conducted using Stata v10 (StataCorp, 2007), the spatial error regression modelling was conducted using the GeoDa software package (Anselin, 2003), and the multi-level modelling was conducted using MLwiN v2.02 (Rasbash et al., 2003).

The five sets of synthetic estimates used as explanatory variables in this chapter are as follows:

- prevalence of consuming less than five portions of fruit and vegetables per day
- prevalence of obesity (body mass index greater than or equal to 30kg/m^2)
- prevalence of raised blood pressure (systolic blood pressure greater than or equal to 160mmHg, or diastolic blood pressure greater than or equal to 95mmHg)
- prevalence of raised cholesterol (total blood cholesterol greater than or equal to 6.5mmol/l)
- prevalence of current smoking

The first four sets of synthetic estimates were developed for the Health Poverty Index website (Dibben et al., 2004). This website was developed by researchers from the universities of St Andrews and Oxford, and is funded by the Department of Health. The aim of the website is to provide public health practitioners with a graphical tool that can compare the degree of 'health poverty' between different areas in Britain. Some of the domains of 'health poverty' are provided by the synthetic estimates that are used in this chapter. All of the Health Poverty Index synthetic estimates are derived from the Health Survey for England (HSfE) series, specifically the surveys conducted between 1998 and 2001.

The smoking synthetic estimates were developed for the Health Development Agency as part of an investigation of the impact of smoking on mortality in England (Twigg et al., 2004). The synthetic estimates are based on the 'current cigarette smoking' response to the HSfE, and are derived from the 1998 to 2001 surveys. During the development of the synthetic estimates, the researchers were given access to protected

HSfE data on the ward of residence of each respondent, which was used to generate a three-level model (individuals nested in wards nested in Government Office Regions) that generates the ward-level estimates. This allowed for the development of a more sophisticated model than those developed for the Health Poverty Index (where access to the 'ward of residence' HSfE variable was restricted).

All of the sets of synthetic estimates were age-standardised to the European Standard Population (West Midlands Public Health Observatory, 2009) using ten-year age bands. Prevalence rates for men and women were generated separately.

RESULTS

Exploratory data analysis

Table 7.1 displays summary statistics for the explanatory variables used in these analyses (summary statistics of the outcome variables are shown in table 5.1, chapter five).

Table 7.1 Summary statistics and correlation matrices of synthetic estimates (wards, n = 7,929)

Explanatory variables (sets of synthetic estimates)

Variable	*Range*	*Interquartile range*	*Standard deviation*	*Mean*	*Median*
Fruit and vegetable consumption, men (%)	61.8 – 78.6	71.0 – 74.0	2.4	72.4	72.5
Fruit and vegetable consumption, women (%)	58.3 – 75.8	67.2 – 70.5	2.6	68.7	68.8
Obesity, men (%)	7.6 – 18.3	12.1 – 14.4	1.7	13.2	13.2
Obesity, women (%)	8.9 – 21.1	14.0 – 16.5	1.9	15.3	15.2
Raised blood pressure, men (%)	6.0 – 9.7	7.8 – 8.6	0.6	8.2	8.1
Raised blood pressure, women (%)	6.0 – 9.0	7.1 – 7.9	0.5	7.5	7.5
Raised cholesterol, men (%)	11.7 – 16.5	16.1 – 16.2	0.4	16.1	16.1
Raised cholesterol, women (%)	14.7 – 20.3	19.8 – 19.9	0.4	19.8	19.9
Smoking, men (%)	13.3 – 45.3	25.5 – 32.5	4.6	28.9	29.0
Smoking, women (%)	9.7 – 45.0	24.2 – 31.1	4.7	27.7	27.5

Correlation matrix (Pearson's r) of explanatory variables (male prevalence estimates)

	Fruit and vegetable consumption	Obesity	Raised blood pressure	Raised cholesterol	Smoking
Fruit and vegetable consumption	1.00				
Obesity	0.82	1.00			
Raised blood pressure	0.66	0.93	1.00		
Raised cholesterol	0.35	0.15	0.20	1.00	
Smoking	0.59	0.55	0.54	-0.02	1.00

Correlation matrix (Pearson's r) of explanatory variables (female prevalence estimates)

	Fruit and vegetable consumption	Obesity	Raised blood pressure	Raised cholesterol	Smoking
Fruit and vegetable consumption	1.00				
Obesity	0.83	1.00			
Raised blood pressure	0.66	0.93	1.00		
Raised cholesterol	0.35	0.16	0.21	1.00	
Smoking	0.77	0.57	0.53	0.32	1.00

The distributions of the synthetic estimates for the prevalence of cholesterol were highly right-skewed and 95% of the male estimates were between 15.5% and 16.5%. The correlation matrices for the sets of synthetic estimates are also shown in table 7.1 and show that in most cases the sets of synthetic estimates were highly correlated: the Pearson correlation coefficient for the raised blood pressure and obesity estimates was 0.93 for both men and women. The correlation coefficients between the sets of synthetic estimates were all highly significant ($p < 0.001$), with the exception of the correlation between the prevalence of raised cholesterol and smoking levels in men ($p = 0.087$). This suggests that, in general, high prevalence rates for risk factors tend to cluster in certain areas – so areas with a high smoking rate are also likely to have a high obesity rate, and so on.

The synthetic estimates were standardised, and PCA was performed on their z scores to produce orthogonal (uncorrelated) independent variables. Full details of this PCA are shown in appendix one. Two of the transformed PCA variables were included in the analyses reported in this chapter – these two variables explained around 85% of the total variance of the sets of synthetic estimates for both male and female prevalence estimates, and their transformation factors allowed for simple interpretation of results (see table 4.2, appendix one). For simplicity, the two PCA variables were named *unhealthy lifestyle 1* and *unhealthy lifestyle 2*. These two variables broadly measure the following features: *unhealthy lifestyle 1* – increased prevalence rate of low fruit and vegetable consumption, obesity, raised blood pressure and smoking; *unhealthy lifestyle 2* – increased prevalence of raised cholesterol and reduced prevalence rate of smoking (in men) or obesity and raised blood pressure (in women). The transformation factors

show the weighting of each of the sets of synthetic estimates that make up the PCA variables. The interpretation of the two variables provided above is subjective. Here, a set of synthetic estimates was deemed to have a strong influence on the derived variable if the transformation factor for that set of synthetic estimates was greater than 0.2. Additionally, the greater the transformation factor, the stronger the influence of the set of synthetic estimates. So the male unhealthy lifestyle 1 variable is roughly equally influenced by low fruit and vegetable consumption, obesity, raised blood pressure and smoking (with transformation factors between 0.41 and 0.54), and is not particularly influenced by raised cholesterol (transformation factor of only 0.17).

In order to use the results of the PCA in the multilevel and spatial error models, it is necessary to generate ward-level estimates of the *unhealthy lifestyle 1* and *unhealthy lifestyle 2* variables. This is achieved by applying the appropriate transformation factors to each ward-level standardised synthetic estimate and then summing the results. For example, the standardised synthetic estimates for the prevalence of low fruit and vegetable consumption, obesity, raised blood pressure, raised cholesterol and smoking for men in the North Sunderland ward were 0.75, 1.08, 1.46, 0.41 and 0.39 respectively. This indicates that the North Sunderland ward has a higher than average prevalence rate of all these risk factors, with raised blood pressure deviating furthest from the national average. The *unhealthy lifestyle 1* variable for North Sunderland was calculated as (0.51*0.75)+(0.54*1.08)+(0.51*1.46)+(0.17*0.41)+(0.41*0.39) = 1.94.

A summary of these two PCA variables is shown in table 7.2. Since the PCA variables were derived from standardised sets of synthetic estimates they have a mean of zero.

Hence, a value greater than zero for the unhealthy lifestyle variable implies that the general lifestyle of the population in the ward is more unhealthy than the England average, and vice versa.

Table 7.2 Summary statistics for the two principal components analysis variables (wards, n = 7,929)

PCA variables

Variable	*Range*	*Interquartile range*	*Standard deviation*	*Mean*	*Median*
Unhealthy lifestyle 1, men (SDs)	-6.7 – 5.3	-1.2 – 1.2	1.8	0.0	-0.1
Unhealthy lifestyle 1, women (SDs)	-6.2 – 5.6	-1.3 – 1.3	1.8	0.0	-0.1
Unhealthy lifestyle 2, men (SDs)	-11.9 – 1.1	0.0 – 0.4	1.0	0.0	0.2
Unhealthy lifestyle 2, women (SDs)	-12.5 – 1.5	-0.2 – 0.4	1.0	0.0	0.2

Transformation factors (see table 4.2, appendix one)

	Fruit and veg consumption	Obesity	Raised blood pressure	Raised cholesterol	Smoking
Unhealthy lifestyle 1, men	0.51	0.54	0.51	0.17	0.41
Unhealthy lifestyle 1, women	0.51	0.51	0.48	0.22	0.45
Unhealthy lifestyle 2, men	0.14	-0.08	-0.06	0.92	-0.36
Unhealthy lifestyle 2, women	0.05	-0.31	-0.28	0.89	0.17

SDs = Standard Deviations

Multi-level regression modelling

The residual variance at ward-level and local authority-level in the baseline model, final model with synthetic estimates as explanatory variables and final model with the PCA variables as explanatory variables is shown in table 7.3. This information is given for completeness – the percentage of both ward-level and local authority-level variance that

is explained by the two sets of final models are shown in table 7.4, which describes the final models.

Table 7.3 Residual variance at ward-level (n = 7,929) and local authority-level (n = 354) for baseline and final multi-level models

		BASELINE		*FINAL - SYNTH ESTIMATES*		*FINAL - PCA VARIABLES*	
		Variance	*Standard Error*	*Variance*	*Standard Error*	*Variance*	*Standard Error*
Mortality models							
MEN	Ward-level	2,096.4	34.1	1,618.3	26.3	1,726.2	28.1
	LA-level	779.7	66.3	328.5	30.3	333.6	31.2
WOMEN	Ward-level	660.8	10.7	576.9	9.4	585.0	9.5
	LA-level	226.8	19.5	110.8	10.4	110.4	10.4
Hospitalisation models							
MEN	Ward-level	48,594.9	789.6	34,309.7	557.5	37,034.9	602.1
	LA-level	37,958.9	3,034.8	21,847.6	1,767.9	25,186.3	2,014.8
WOMEN	Ward-level	14,884.6	241.8	11,139.5	181.0	11,672.4	189.7
	LA-level	12,618.2	1,004.3	9,866.8	783.1	7,819.1	628.2

All of the sets of synthetic estimates and the two PCA variables showed a significant ($p < 0.05$) association with both CHD mortality and hospitalisation rates in univariate analyses and hence were included in the multivariate models (shown in table 7.4). The four models that included the sets of synthetic estimates as explanatory variables produced some erratic results: prevalence of obesity, low fruit and vegetable consumption and raised cholesterol were not significantly associated with male mortality rates but were with female mortality rates; prevalence of raised blood pressure was negatively associated with male hospitalisation rates but positively associated with female hospitalisation and both male and female mortality rates; prevalence of raised cholesterol was positively associated with male hospitalisation rates but negatively associated with female hospitalisation rates; and far less large scale geographic variation in female hospitalisation rates (22%) was explained by the sets of synthetic estimates than for male hospitalisation rates (42%).

Table 7.4 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against behavioural risk factor profiles of populations (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	-139.8			83.3		
Fruit and veg (%)	-0.5	0.8	0.484	-1.3	0.5	0.017
Obesity (%)	-1.9	1.9	0.313	4.7	1.1	<0.001
Blood pressure (%)	33.2	4.3	<0.001	9.7	2.9	<0.001
Cholesterol (%)	-2.1	2.7	0.430	-5.4	1.3	<0.001
Smoking (%)	5.0	0.2	<0.001	1.8	0.1	<0.001
Ward-level variance explained:		23%			13%	
LA-level variance explained:		58%			51%	
Constant	180.1			83.7		
PCA: Unhealthy lifestyle 1 (SDs)	17.0	0.4	<0.001	8.4	0.2	<0.001
PCA: Unhealthy lifestyle 2 (SDs)	-10.2	0.7	<0.001	-3.8	0.5	<0.001
Ward-level variance explained:		18%			11%	
LA-level variance explained:		57%			51%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	482.0			-163.2		
Fruit and veg (%)	-53.1	3.7	<0.001	-13.9	2.5	<0.001
Obesity (%)	139.9	8.8	<0.001	40.2	5.1	<0.001
Blood pressure (%)	-59.2	23.1	0.011	123.3	15.5	<0.001
Cholesterol (%)	123.3	13.6	<0.001	-18.9	6.6	0.004
Smoking (%)	31.3	0.9	<0.001	11.7	0.7	<0.001
Ward-level variance explained:		29%			25%	
LA-level variance explained:		42%			22%	
Constant	898.0			371.9		
PCA: Unhealthy lifestyle 1 (SDs)	102.8	2.2	<0.001	57.7	1.2	<0.001
PCA: Unhealthy lifestyle 2 (SDs)	-44.4	4.1	<0.001	-21.4	2.4	<0.001
Ward-level variance explained:		24%			22%	
LA-level variance explained:		34%			38%	

SDs = Standard Deviations

The models with the two PCA variables as explanatory variables were far more stable: each of the four models showed a strongly significant positive association between the unhealthy lifestyle 1 variable and the outcomes and a strongly significant negative association between the unhealthy lifestyle 2 variable and the outcomes. Around 55% of the large scale geographic variation in mortality rates was explained by the PCA variables, and around 35% of large scale geographic variation of hospitalisation rates. Between 10% and 25% of small scale geographic variation in mortality and hospitalisation rates was explained by the PCA variables. The amount of geographic variation explained by the multivariate PCA models was only slightly higher than that of the univariate models including only the unhealthy lifestyle 1 PCA variable, suggesting that the unhealthy lifestyle 2 PCA variable added little to the explanation of geographic variance in CHD rates.

This very small increase in explanatory power of the models when the unhealthy lifestyle 2 variable is added suggests that the second PCA variable does little to explain geographic variation in CHD rates once the first variable has been accounted for. This seems paradoxical, since the two PCA variables have been designed to be uncorrelated with each other, but there is a logical explanation for these findings. The PCA variables were built from five input variables (the five sets of synthetic estimates) and can be considered to be a column vector in five dimensional space (with the weightings of each of the sets of synthetic estimates displayed in table 7.2 as the values of the column vectors). By definition, these two column vectors are orthogonal. However, the five sets of synthetic estimates are not all associated with CHD rates with the same strength. In

fact, the univariate analyses showed that the association between raised cholesterol prevalence rates and CHD mortality and hospitalisation rates was not very strong. This is supported by similar findings reported from the British Regional Heart Study (Morris et al., 2001) and it is because the prevalence of raised cholesterol does not vary much around England (see table 7.1) once the age and sex structure of populations has been accounted for. Because of this lack of association, when the two PCA variables are included together in the models the cholesterol elements of the variables are almost irrelevant. Therefore, in this instance the variables are virtually equivalent to four dimensional column vectors. But these four dimensional vectors are not orthogonal, they are strongly negatively correlated, with the unhealthy lifestyle 1 variable describing increased prevalence rates of smoking, raised blood pressure, obesity and low fruit and vegetable consumption and the unhealthy lifestyle 2 variable describing decreased prevalence rates of the same behavioural risk factors. Therefore, in this instance the unhealthy lifestyle 2 variable does not add any explanatory power to the model after the unhealthy lifestyle 1 variable has been accounted for.

Spatial error regression modelling

As was the case for the multi-level models, all of the synthetic estimates and PCA variables were significantly associated with the outcomes in univariate analyses, and hence all were included in the multivariate models, which are displayed in table 7.5.

Table 7.5 Spatial error regression models of (1) CHD mortality rates and (2) hospitalisation rates against behavioural risk factor profiles of populations (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN				*WOMEN*			
Variable	*Beta*	*SE*	*p*	*Agree with multi-level model**	*Beta*	*SE*	*p*	*Agree with multi-level model**
Constant	66.3				54.8			
Fruit and veg (%)	-2.6	0.6	<0.001	x	-0.4	0.4	0.300	✓
Obesity (%)	4.3	1.6	0.006	x	3.0	0.9	0.001	✓
Blood pressure (%)	15.0	3.5	<0.001	✓	8.9	2.3	<0.001	✓
Cholesterol (%)	-1.6	2.5	0.517	✓	-5.0	1.2	<0.001	✓
Smoking (%)	5.0	0.2	<0.001	✓	1.5	0.1	<0.001	✓
Spatial error	0.3	0.0	<0.001		0.4	0.0	<0.001	
Model r^2		0.38				0.29		
Constant	179.5				83.4			
PCA: Unhealthy lifestyle 1 (SDs)	15.5	0.4	<0.001	✓	7.4	0.2	<0.001	✓
PCA: Unhealthy lifestyle 2 (SDs)	-8.8	0.7	<0.001	✓	-3.5	0.4	<0.001	✓
Spatial error	0.4	0.0	<0.001		0.4	0.0	<0.001	
Model r^2		0.35				0.29		

(2) Hospitalisation rates models

	MEN				*WOMEN*			
Variable	*Beta*	*SE*	*p*	*Agree with multi-level model**	*Beta*	*SE*	*p*	*Agree with multi-level model**
Constant	354.5				43.4			
Fruit and veg (%)	-39.0	3.3	<0.001	✓	-2.6	2.0	0.192	x
Obesity (%)	107.2	8.0	<0.001	✓	19.9	4.5	<0.001	✓
Blood pressure (%)	-94.9	18.2	<0.001	x	56.6	11.6	<0.001	✓
Cholesterol (%)	115.1	14.3	<0.001	✓	-25.6	6.6	<0.001	✓
Smoking (%)	29.9	0.9	<0.001	✓	10.3	0.5	<0.001	✓
Spatial error	0.6	0.0	<0.001		0.6	0.0	<0.001	
Model r^2		0.57				0.56		
Constant	896.7				369.1			
PCA: Unhealthy lifestyle 1 (SDs)	91.4	2.2	<0.001	✓	51.4	1.2	<0.001	✓
PCA: Unhealthy lifestyle 2 (SDs)	-37.6	4.4	<0.001	✓	-13.4	2.5	<0.001	✓
Spatial error	0.7	0.0	<0.001		0.7	0.0	<0.001	
Model r^2		0.55				0.55		

* ✓ = Direction of association and whether association is significant (p<0.01) agrees with equivalent multi-level regression models; x = otherwise.
SDs = Standard Deviations

There were a number of disagreements between the spatial error models and multi-level models that included the sets of synthetic estimates as explanatory variables. This is likely to be due to the erratic nature of the models caused by the collinearities of the sets of synthetic estimates rather than by spatial autocorrelation bias. The multi-level models and spatial error models that included the PCA variables as explanatory variables showed very good agreement. In all cases the parameter estimates in the two different models were very similar, suggesting that the multi-level models have not been substantially affected by spatial autocorrelation bias.

DISCUSSION

In the crude models reported in this chapter (unadjusted for environmental variables and deprivation) around 55% of the large scale geographic variation in CHD mortality rates and 35% of large scale geographic variation in hospitalisation rates in England can be explained by differences in the behavioural risk factor profiles of populations: i.e. the prevalence of unhealthy lifestyle behaviours (such as smoking and poor diet) and related medical conditions (hypertension and obesity). Differences in behavioural risk factor profiles of populations also explain between 10% and 25% of the small scale geographic variation. The prevalence of raised cholesterol does not vary substantially around England, and hence does not appear to have an impact on variation in CHD mortality and hospitalisation rates. These results are not substantially affected by spatial autocorrelation bias.

High prevalence rates of behavioural risk factors tend to cluster in certain areas and populations, and it is therefore not possible using this study design to attribute the geographic variation in CHD rates to individual risk factors. This is especially the case here, since the synthetic estimates of the prevalence of behavioural risk factors are estimated with error, therefore even when all of the risk factors are adjusted against each other in the multivariate final models the estimated strength of association for each risk factor may still be biased by residual confounding (Davey Smith and Phillips, 1992; Phillips and Davey Smith, 1991). However, an 'unhealthy lifestyle' variable, developed for this analysis, shows a strong positive association with CHD mortality and hospitalisation rates. The models reported here suggest that an increase in one standard deviation of this index is associated with an increase in age-standardised CHD mortality of 17.0 male deaths per 100,000, and 8.4 deaths per 100,000 in women. The difference between the best and worst ward in England is 12 standard deviations, representing a difference of 204 male deaths per 100,000 and 101 female deaths per 100,000 – about the range from the 5^{th} to 95^{th} percentile of age-standardised CHD mortality rate.

Viewing the results of this chapter alongside those reported in chapter five (which showed that environmental variables explained approximately the same amount of large scale geographic variance in CHD rates as behavioural risk factor profiles of populations, but only around 5% to 10% of the small scale geographic variance) could lead to the following conclusions. Firstly, the North-South CHD gradient in England is due in equal parts to differences in behavioural risk factor profiles of populations and differences in the climate, and interventions to reduce geographical inequalities should

therefore focus on reducing the impact of the colder Northern climate (e.g. improving household insulation) as much as improving the lifestyle of the Northern population. Secondly, small scale differences in CHD mortality and hospitalisation rates are only partially explained by urbanicity, air pollution and behavioural risk factor profiles of populations, and more work is needed to identify the cause of local differences. These conclusions are premature, however, as it is has yet to be determined whether the associations shown in this chapter and in chapter five confound each other, or whether they are a result of other potentially confounding variables such as socio-cultural factors which can be proxied with some success by deprivation indices. The issue of confounding is explored in the following chapter which reports on models of CHD rates that include environmental variables, behavioural risk factor profiles of populations and deprivation as explanatory variables.

In this chapter the 'unhealthy lifestyle' variable – constructed from a number of different estimates of the prevalence of behavioural risk factors – has been interpreted as an index of unhealthy lifestyle, in a similar way as deprivation indices that are constructed on the basis of a number of individual measures of specific social or economic deprivation. Such an index could be useful if it can provide additional explanatory power to models of health outcomes (not solely restricted to CHD, since the behavioural risk factors included in the index are risk factors for many chronic diseases) beyond that provided by an indirect measure of unhealthy behaviour such as a deprivation index. The potential usefulness of this variable is also explored in the following chapter.

The sets of synthetic estimates selected for these analyses have been shown in chapter six to be valid and reasonably accurate. One of the problems with synthetic estimates identified in the previous chapter is that they tend to reduce the true variance in the variable being estimated, by over-estimating in areas of low prevalence and under-estimating in areas of high prevalence. The introduction of the standardised measure of unhealthy behaviour removes the risk of inflated parameter estimates that would be a result of this reduced variance, but it also removes the possibility of direct interpretation of the results of the models. The sets of synthetic estimates used in these analyses allow for a reasonable coverage of the behavioural and medical risk factors shown in the conceptual framework for this thesis (figure 2.1, chapter two). However, there are some direct routes to CHD displayed in the framework which it has not been possible to test due to lack of valid synthetic estimates, such as the impact of stress and diabetes. It is unclear whether the unhealthy lifestyle variable would also be correlated with the prevalence of either diabetes or high stress levels, therefore it is possible that some of the geographic variation remaining in CHD mortality and hospitalisation rates could be due to differences in these conditions.

Ecological designs of the kind used in this chapter are ideal for examining area-level variables, but they are not ideal for the study of individual-level variables. This is because the collapsing of the lifestyle experience of a population of individuals to an aggregated measure results in a loss of information. This can be for a number of reasons: firstly, the single measure can disguise a variety of different age distributions, so that a ward with a high percentage of elderly smokers, but few young smokers would be indistinguishable from a ward with a high percentage of young smokers but few

elderly smokers. It has been suggested that increased cardiovascular risk is related to risk exposure over the life course (Davey Smith and Hart, 2002), therefore these two distributions could produce very different CHD outcomes. Secondly, the prevalence rate method inevitably involves collapsing continuous individual-level variables to binary measures (such as the obesity measure of BMI>=30kg/m^2), which again could hide a variety of underlying distributions. These problems, and the lack of inclusion of estimates of the prevalence of diabetes and raised stress, are limitations which suggest that the true amount of geographic variation in CHD rates that is explained by behavioural risk factor profiles of populations may be higher than the amount predicted by the models from this chapter.

Although the sets of synthetic estimates used in this chapter have been validated, their use as explanatory variables in regression analyses is problematic, and there is an important limitation that must be acknowledged. Synthetic estimates of the prevalence of behavioural risk factors are associated with confidence intervals related to the accuracy of the estimates. However, an assumption of regression analysis is that the explanatory variables are not associated with any error. If this assumption is severely violated, then the results of a regression analysis can be biased. If the error associated with the explanatory variables is measurement error, and is randomly distributed with a mean of zero, then the effect is to *increase* the total variance of the explanatory variables, and hence to bias the results of the regression analysis towards the null hypothesis (Cheng and Van Ness, 1999). Such measurement error can be accounted for by using measurement error regression modelling techniques, but these techniques are not appropriate here because the sets of synthetic estimates used in these analyses are

biased towards the mean, and hence have a *reduced* total variance. This implies that the error terms associated with the estimates are not randomly distributed, but rather are more likely to be positive if the estimate is below the mean, and more likely to be negative if the estimate is above the mean. As mentioned above, the importance of this limitation is reduced when the standardised 'unhealthy lifestyle' variable is used as the explanatory variable.

The analyses reported in this chapter found that raised cholesterol rates had little association with CHD mortality or hospitalisation rates, because the variance in the prevalence of raised cholesterol was small. This is a finding that is replicated in both the British Regional Heart Study (BRHS) and the British Women's Heart and Health Study (BWHHS). The BRHS found that including cholesterol levels in the final model *reduced* the amount of variance in CHD incidence rates that can be explained by differences in risk factors (Morris et al., 2001), and the BWHHS showed that the geographic pattern in the prevalence of raised cholesterol was opposite to the pattern displayed in CHD prevalence, with the highest raised cholesterol rates in the South of England, and the lowest rates in Scotland (Lawlor et al., 2003). These consistent findings that raised cholesterol levels have little association with CHD rates within Britain validate the decision not to include the standardised 'unhealthy lifestyle 2' PCA variable in further analyses reported in the following chapter.

CONCLUSION

A large proportion of the large scale geographic variation in mortality and hospitalisation rates for CHD in England (and a smaller part of the small scale

variation) can be explained by differences in behavioural risk factor profiles of populations around the country. It is unclear, however, if this relationship is confounded by area-level deprivation and environmental variables. The following chapter will report on the modelling of CHD mortality and hospitalisation rates that include both environmental variables and behavioural risk factor profiles of populations, and also deprivation, which should provide evidence of whether the association between the index of unhealthy lifestyle and CHD rates is independent of area-level deprivation, and provide an estimate of the amount of variation in CHD rates that is due to both environmental and behavioural determinants.

Chapter 8: The impact of confounding on the relationship between environmental variables, behavioural risk factor profiles of populations and coronary heart disease mortality and hospitalisation rates in England

INTRODUCTION

This is the final analytical chapter of the thesis. The previous chapters have explored the amount of geographic variation in coronary heart disease (CHD) that can be explained by environmental variables, and the amount that can be explained by behavioural risk factor profiles of populations, and introduced an 'unhealthy lifestyle' index, derived from synthetic estimates of the prevalence of behavioural risk factors for CHD, which was shown to be positively associated with CHD mortality and hospitalisation rates for both men and women. Table 8.1 summarises the results of the multi-level analyses conducted in chapters five and seven, exploring the amount of geographic variation in CHD mortality and hospitalisation rates that is explained by environmental variables and behavioural risk factor profiles of populations. Both the environmental variables and the behavioural risk factor profiles of populations were successful at explaining a substantial amount of the large scale geographic variance in both CHD mortality and hospitalisation rates, whereas they were less successful at explaining small scale geographic variance, particularly for mortality rates.

Table 8.1 Large scale and small scale geographic variation in CHD mortality and hospitalisation rates explained by environmental variables and behavioural risk factor profiles of populations, drawn from previous analyses (see chapters 5 and 7)

		Small scale variance in CHD rates explained		Large scale variance in CHD rates explained	
		Men	Women	Men	Women
MORTAL-ITY	Environmental variables	5%	3%	56%	59%
	Behavioural risk factor profile	18%	11%	57%	51%
HOSPITAL-ISATIONS	Environmental variables	8%	7%	34%	40%
	Behavioural risk factor profile	24%	22%	34%	38%

The models that were developed in the previous chapters and that generated the results displayed in table 8.1 did not fully allow for potential confounding. The two sets of variables (environmental and behavioural risk factor profiles of populations) are likely to confound each other to some degree, and also be confounded with deprivation, since climate, unhealthy behaviour and deprivation all have similar North-South gradients (Joint Health Surveys Unit, 2008; Department of the Environment, Transport and the Regions, 2000). It is essential to determine the extent of this confounding so that accurate interpretations can be made from the results of the analyses presented in this thesis. The impact of this confounding is addressed in this chapter, which answers the following research question:

RQ1: How much of the large scale and small scale geographic variation in CHD mortality and hospitalisation rates can be explained by environmental variables, behavioural risk factor profiles of populations and deprivation, after adjustment for each other?

METHODS

The general structure for the analyses reported here is similar to those reported in chapters five and seven: multi-level and spatial error regression models of male and female age-standardised CHD mortality and hospitalisation rates. The explanatory variables included in the models were those that were considered to be worthy of further investigation in the previous chapters. These include a subset of the environmental variables that were used in chapter five (mean daily maximum temperature, total annual hours of sunlight, air quality index, and a categorical urbanicity variable), and the standardised 'unhealthy lifestyle' variable generated from the sets of synthetic estimates of the prevalence of behavioural risk factors for CHD in chapter seven.

In addition, the models include a measure of area-level deprivation - the Carstairs index (Carstairs and Morris, 1990), generated using data from the 2001 census at ward level (Morgan and Baker, 2006). The variables included in the index are: % economically active men who are unemployed; % individuals who live in overcrowded accommodation; % people in households where the head of the household is in the semi-skilled or unskilled social class; % individuals with no access to a car. This particular

measure of deprivation has previously been shown to be strongly positively associated with cardiovascular disease rates in England (Romeri et al., 2006).

In order to disentangle the effects of confounding between the three sets of explanatory variables (environmental, behavioural risk factor profiles of populations and deprivation), models were constructed that contained any possible combination of the explanatory variables (i.e. three models that contained only one set of explanatory variables; three models that contained two sets of explanatory variables; one model that contained all three sets of explanatory variables). The comprehensive results of all multi-level and spatial error models are shown in appendix three (tables 8.5 to 8.16). Only the final models that include all three sets of explanatory variables are shown in this chapter, which provide most of the evidence for the interpretations regarding the amount of geographic variation in CHD rates that can be attributed to environmental variables, behavioural risk factor profiles of populations and deprivation. The intermediary models shown in the appendix were necessary in order to draw conclusions about the nature of the confounding between the three sets of explanatory variables. These conclusions were drawn in a systematic way by reference to each developed model as follows:

- Initially, the models with two sets of explanatory variables *('B models')* were compared with the complementary models containing only one set of explanatory variables *('A models')*. Beta coefficients (measuring the strength of the association between the individual variables and the CHD outcome, adjusted for all other variables in the model) for each of the variables were compared between the B models and the A models. Substantial reduction of the beta coefficient (or

reversal of the sign) in the B model was interpreted as evidence of confounding of the relationship between the variable and CHD by the other set of explanatory variables present in the B model.

- The amount of variance in CHD rates explained by the B models was compared to the amount explained by the two complementary A models. The amount of variance explained in the B model could not be lower than for either of the A models, but if the value was not substantially higher than for either of the A models than the other set of explanatory variables was considered to have contributed little to the explanation of variance in the B model.
- The final models that contained all three sets of explanatory variables *('C models')* were then compared with each of the B models. Again, the size and sign of the beta coefficients in the C models were compared with results from each of the B models, and the amount of variance explained by the C models was compared with the amount explained by each of the B models.

RESULTS

Exploratory data analysis

Summary statistics for the deprivation variable are shown in table 8.2, along with a correlation matrix of the continuous variables used as explanatory variables in the analyses reported here, and a summary of the mean of the continuous variables for each category of the urbanicity variable.

Table 8.2 Summary statistics, correlation co-efficient matrix of the continuous explanatory variables, and mean of continuous variables by urbanicity category (wards, n = 7,929)

Summary of deprivation variable

Variable	*Range*	*Interquartile range*	*Standard deviation*	*Mean*	*Median*
Deprivation (SDs)	-5.7 – 16.5	-5.4 – 15.1	3.5	-0.1	-1.0

Correlation co-efficient matrix

	Mean max. temp	Sunshine	Air quality index	Unhealthy lifestyle, men	Unhealthy lifestyle, women	Deprivation
Mean max. temp	1.00					
Sunshine	0.63	1.00				
Air quality index	0.21	-0.07	1.00			
Unhealthy lifestyle, men	-0.43	-0.40	-0.07	1.00		
Unhealthy lifestyle, women	-0.44	-0.39	-0.12	0.99	1.00	
Deprivation	-0.19	-0.17	0.42	0.57	0.51	1.00

Mean of continuous variables by urbanicity category

Variable	*Coastal & countryside*	*Urban*	*Metropolitan*	*p for trend*
Mean max. temp (°C)	13.8	14.0	14.7	<0.001
Sunshine (000s hrs / yr)	1.5	1.5	1.6	<0.001
Air quality index (SDs)	0.9	1.2	1.6	<0.001
Unhealthy lifestyle, men (SDs)	-0.1	0.2	-1.4	0.001
Unhealthy lifestyle, women (SDs)	-0.1	0.3	-2.1	<0.001
Deprivation (SDs)	-1.8	0.3	6.0	<0.001

SDs = Standard Deviations

The climate variables showed a strong negative association with the unhealthy lifestyle variables, indicating that both these sets of variables share a North-South gradient. Deprivation was also associated with the unhealthy lifestyle variable, although the

strength of the association (r = 0.57 for men and r = 0.51 for women) suggests that the variables are reasonably independent of each other. Surprisingly, the unhealthy lifestyle variable showed *lower* levels of unhealthy lifestyle in metropolitan wards than both urban and coastal and countryside wards, which is an opposite finding to the relationship between deprivation and urbanicity. As expected, the air quality was worse in more urban wards, which explains why the air quality index showed a reasonably strong association with deprivation.

Multi-level regression modelling

Univariate analyses showed that deprivation was strongly positively associated with both CHD mortality and hospitalisation rates. Deprivation alone explained around 20% of the small scale geographic variation in mortality and 30% of the small scale variation in hospitalisations, and around 45% of the large scale geographic variation in both mortality and hospitalisation rates (table 8.10, appendix three and tables 8.5 and 8.6 below).

The residual variance at ward-level and local authority-level in the baseline model and final model is shown in table 8.3. This information is given for completeness – the percentage of both ward-level and local authority-level variance that is explained by the final model and all the sets of intermediary models are shown in tables 8.5 and 8.6.

Table 8.3 Residual variance at ward-level (n = 7,929) and local authority-level (n = 354) for baseline and final multi-level models

		BASELINE		*FINAL*	
		Variance	*Standard Error*	*Variance*	*Standard Error*
Mortality models					
MEN	Ward-level	2,096.4	34.1	1,580.2	25.7
	LA-level	779.7	66.3	166.1	18.1
WOMEN	Ward-level	660.8	10.7	547.6	8.9
	LA-level	226.8	19.5	53.5	6.0
Hospitalisation models					
MEN	Ward-level	48,594.9	789.6	32,561.3	529.2
	LA-level	37,958.9	3,034.8	17,443.4	1,428.2
WOMEN	Ward-level	14,884.6	241.8	9,928.1	161.3
	LA-level	12,618.2	1,004.3	4,352.7	362.8

Table 8.4 shows the multi-level models that include all of the explanatory variables. The models were very successful at explaining large scale geographic variation in CHD rates – around 60% of the large scale variation in hospitalisation rates and nearly 80% of the large scale variation in mortality rates was explained by the environmental variables, behavioural risk factor profiles of populations and deprivation. The models were less successful at explaining small scale geographic variation – only 33% for hospitalisation rates and around 20% for mortality rates. Both the deprivation and the behavioural risk factor profile of populations variables showed consistent associations with CHD rates in each of the models. Both of the variables displayed highly significant positive associations, and the deprivation variable was in all cases more strongly associated with CHD rates than the behavioural risk factor profile of populations (the parameter estimates and the range of the two variables suggest that the deprivation variable could explain more of the variation in CHD rates).

Table 8.4 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against environmental variables, behavioural risk factor profiles of populations and deprivation (wards, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	387.4	19.2		207.9	11.0	
Mean max. temp (°C)	-12.5	1.8	<0.001	-7.9	1.1	<0.001
Sunshine (000s hrs / yr)	-27.3	11.6	0.019	-14.3	6.7	0.032
Air quality index (SDs)	5.8	3.2	0.067	5.7	1.9	0.002
Urban†	1.9	1.2	0.105	0.4	0.7	0.549
Metropolitan†	-8.0	3.5	0.023	1.2	2.1	0.576
Unhealthy lifestyle (SDs)	5.0	0.6	<0.001	3.3	0.3	<0.001
Deprivation (SDs)	7.2	0.3	<0.001	3.0	0.2	<0.001
Ward-level variance explained:		25%			17%	
LA-level variance explained:		79%			76%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	1,126.9	160.4		812.8	81.3	
Mean max. temp (°C)	-34.7	14.7	0.018	-33.3	7.5	<0.001
Sunshine (000s hrs / yr)	82.5	96.1	0.390	-32.9	48.9	0.503
Air quality index (SDs)	100.2	19.1	<0.001	61.3	10.3	<0.001
Urban†	17.5	5.3	<0.001	1.2	3.0	0.697
Metropolitan†	74.8	17.2	<0.001	12.9	9.7	0.180
Unhealthy lifestyle (SDs)	45.9	3.3	<0.001	21.3	1.7	<0.001
Deprivation (SDs)	34.1	1.4	<0.001	21.2	0.7	<0.001
Ward-level variance explained:		33%			33%	
LA-level variance explained:		54%			66%	

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations

The results for the environmental variables were less clear. The maximum temperature variable showed a significant negative association with CHD rates in all four of the models, whereas hours of sunshine was only weakly significantly negatively associated with mortality rates, and even showed a positive (non-significant) association with male

hospitalisation rates. The air quality index variable showed only a small association with CHD mortality rates after adjustment for deprivation and behavioural risk factor profiles of populations (this association was non-significant for men) whereas there was a strong positive association between air pollution and hospitalisation rates for both men and women. The effect of urbanicity on CHD rates showed a clear gender divide, with no significant association between urbanicity and female mortality or hospitalisation rates. In contrast, living in a metropolitan ward was associated with lower male mortality rates, but higher male hospitalisation rates.

Apportioning the amount of small scale and large scale geographic variation that is explained by the environmental variables, the behavioural risk factor profiles of populations, and deprivation is not straightforward. The intermediary models (A models with one set of explanatory variables and B models with two sets of exposure variables) are displayed in tables 8.8 to 8.13 in appendix three. A summary of these tables is provided in tables 8.5 and 8.6 below. These tables display the beta parameters that were estimated for each of the explanatory variables in each of the intermediary models, and also the amount of ward-level and local authority-level variance explained by the intermediary models, and allows for an assessment of confounding between the explanatory variables.

Table 8.5 Summary of beta parameters, explanation of ward-level variance and explanation of local authority-level variance for all intermediary models of mortality rates (wards nested in local authorities, n = 7,929)

	MEN							*WOMEN*						
	A			*B*			*C*	*A*			*B*			*C*
Mean max temp (°C)	-31			-16	-13		-12	-16			-9	-9		-8
Sunshine (000s / yr)	-33			18	-47		-27	-21			3	-26		-14
Air quality index (SDs)	40			33	0		6	19			17	1		6
Urban†	13			6	2		2	5			2	0		0
Metropolitan†	31			41	-21		-8	15			24	-9		1
Unhealthy lifestyle (SDs)		18		16		7	5		8		8		4	3
Deprivation (SDs)			9		9	7	7			4		4	3	3
Ward-level variance explained	4%	16%	24%	19%	24%	25%	25%	3%	11%	17%	14%	16%	17%	17%
LA-level variance explained	55%	49%	46%	77%	76%	67%	79%	59%	45%	40%	77%	73%	62%	76%

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations
A: Models with only one set of exposure variables; *B:* Models with two sets of exposure variables; *C:* Models with all three sets of exposure variables

Table 8.6 Summary of beta parameters, explanation of ward-level variance and explanation of local authority-level variance for all intermediary models of hospitalisation rates (wards nested in local authorities, n = 7,929)

	MEN							*WOMEN*						
	A			*B*			*C*	*A*			*B*			*C*
Mean max temp (°C)	-156			-47	-48		-35	-100			-43	-39		-33
Sunshine (000s / yr)	60			324	-78		82	-32			103	-108		-33
Air quality index (SDs)	336			243	52		100	196			153	37		61
Urban†	74			34	17		18	34			12	1		1
Metropolitan†	237			285	-27		75	107			161	-45		13
Unhealthy lifestyle (SDs)		108		106		38	46		55		56		20	21
Deprivation (SDs)			50		49	40	34			28		28	23	21
Ward-level variance explained	8%	23%	31%	28%	31%	33%	33%	7%	21%	32%	26%	32%	33%	33%
LA-level variance explained	20%	24%	45%	48%	50%	53%	54%	35%	33%	48%	62%	61%	60%	66%

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations
A: Models with only one set of exposure variables; *B:* Models with two sets of exposure variables; *C:* Models with all three sets of exposure variables

For the mortality models, it would seem that the environmental variables contribute little to the explanation of small scale variation: the male mortality model containing only the environmental variables explained 4% of the small scale variation, via the air quality index and urbanicity variables (see table 8.8, appendix three), but the regression parameters for these variables were much reduced or even reversed after the introduction of deprivation (see table 8.12, appendix three). However, the climate variables clearly contribute to the explanation of large scale variance in mortality, even after adjustment for behavioural risk factor profiles of populations and deprivation: the models containing only behavioural risk factor profiles of populations and deprivation explained around 65% of the large scale variation in male mortality rates (see table 8.13, appendix three), whereas this increased to nearly 80% in the final fully-adjusted model (table 8.4). The regression parameters for the climate variables in the fully-adjusted models were substantially smaller than in the models containing only environmental variables, suggesting that climate does not make the major contribution to explanation of large scale variation in CHD mortality rates.

Similarly, the environmental variables make a small but independent contribution to the explanation of large scale geographic variation to CHD hospitalisation rates. In contrast to the mortality models, air quality and urbanicity also make a small but independent contribution to explanation of the small scale variation in hospitalisation rates, indicated by the significant regression parameters displayed in table 8.4, despite the fact that the fully adjusted models explained the same amount of small scale geographic variation as

the hospitalisation models containing only deprivation and behavioural risk factor profiles of populations.

Deprivation and behavioural risk factor profiles of populations together are responsible for most of the explanation of the small scale variation, and much of the large scale variation in both CHD mortality and hospitalisation rates. When separately included in models with the environmental variables (tables 8.11 and 8.12, appendix three), the regression parameters of the unhealthy lifestyle variables and deprivation remained very similar to those found in the univariate models containing only behavioural risk factor profiles of populations or deprivation (tables 8.9 and 8.10, appendix three). Due to their mutual confounding, it is impossible to say exactly how much of the small scale and large scale variation in CHD rates is due to behavioural risk factor profiles of populations and deprivation individually. It is likely that deprivation is responsible for more of the variation in both mortality and hospitalisation models since the deprivation variables were more strongly associated with CHD rates in the fully-adjusted models.

Spatial error regression modelling

The spatial error models that contained either one or two sets of explanatory variables (tables 8.14 – 8.19, appendix three) showed good agreement with the multi-level models, suggesting that spatial autocorrelation bias in the multi-level models was small. The spatial error regression models that included all of the explanatory variables are shown in table 8.7, which seems to indicate that there is a good deal of disagreement between the

spatial error and multi-level mortality models. However, much of this ‘disagreement’ is due to the arbitrary choice of threshold of significance that was used for demonstration in the tables. If the significance threshold was set at $p = 0.05$ rather than $p = 0.01$ then there would be no disagreement for the sunshine and metropolitan variables, for example, and the difference between regression parameter estimates for the air quality index in the male mortality rate model was not substantial (multi-level model estimate 5.8, $p = 0.07$; spatial error model estimate 8.3, $p < 0.01$). The parameter estimates in the spatial error and multi-level models were generally within 10% of each other, but this was not always the case – for example, the parameter estimates for maximum temperature and the air quality index in the male hospitalisation rate spatial error model were around half of those predicted in the equivalent multi-level model.

Table 8.7 Spatial error regression models of (1) CHD mortality rates and (2) CHD hospitalisation rates against environmental variables, behavioural risk factor profiles of populations and deprivation (wards, n = 7,929)

(1) Mortality rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	384.1				206.8			
Mean max. temp (°C)	-12.1	1.3	<0.001	✓	-8.1	0.8	<0.001	✓
Sunshine (000s hrs / yr)	-30.6	8.5	<0.001	x	-13.4	5.0	0.007	x
Air quality index (SDs)	8.3	2.7	0.002	x	8.0	1.6	<0.001	✓
Urban†	2.2	1.2	0.056	✓	0.7	0.7	0.277	✓
Metropolitan†	-12.6	3.4	<0.001	x	-1.4	2.1	0.512	✓
Unhealthy lifestyle (SDs)	4.7	0.5	<0.001	✓	3.1	0.3	<0.001	✓
Deprivation (SDs)	6.9	0.3	<0.001	✓	2.8	0.2	<0.001	✓
Spatial error	0.2	0.0	<0.001		0.2	0.0	<0.001	
Model r^2:	0.41				0.35			

(2) Hospitalisation rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	1,109.3				759.2			
Mean max. temp (°C)	-18.8	10.8	0.082	✓	-25.7	5.4	<0.001	✓
Sunshine (000s hrs / yr)	-16.4	71.4	0.819	✓	-72.1	35.6	0.043	✓
Air quality index (SDs)	53.8	20.0	0.007	✓	67.8	10.3	<0.001	✓
Urban†	17.1	5.4	0.001	✓	1.1	3.0	0.703	✓
Metropolitan†	52.4	18.1	0.004	✓	-6.1	10.2	0.547	✓
Unhealthy lifestyle (SDs)	37.3	3.2	<0.001	✓	19.2	1.6	<0.001	✓
Deprivation (SDs)	36.1	1.5	<0.001	✓	21.3	0.8	<0.001	✓
Spatial error	0.6	0.0	<0.001		0.6	0.0	<0.001	
Model r^2:	0.59				0.60			

* ✓ = Direction of association and whether association is significant (p<0.01) agrees with equivalent multi-level regression models; x = otherwise.

† Compared to baseline of 'Coastal and Countryside'

SDs: Standard Deviations

DISCUSSION

In previous chapters the relationships between CHD and environmental variables and behavioural risk factor profiles of populations have been assessed separately, and the tentative conclusions were that environmental variables have a big impact on large scale geographic variation in CHD rates and a modest impact on small scale variation, whereas the behavioural risk factor profile of populations of areas explained a large proportion of both the small scale and large scale variation in CHD mortality and hospitalisation rates. The analyses reported in this chapter suggest that the relationship between the environmental variables and CHD is largely confounded by both deprivation and behavioural risk factor profiles of populations, whilst the association between behavioural risk factor profiles of populations and CHD is unaffected by environmental variables, but confounded by deprivation. However, the climate has a small but independent impact on both CHD mortality and hospitalisation rates – the models estimate that male mortality rates in the coldest areas of England are 40 deaths per 100,000 higher than in the warmest areas – and air quality has a small but independent impact on both hospitalisation and mortality rates. In addition, urbanicity is independently associated with increased levels of hospitalisation in men, but not women.

The conceptual framework for the analyses reported in this chapter is shown in figure 2.1 (chapter two), although it was not possible to include all of the variables in the framework in the final models (specifically water hardness, socio-cultural factors and the behavioural and medical risk factors for CHD for which there were no valid sets of synthetic estimates). In most of the cases, the direct relationships with CHD that are

suggested in figure 2.1 were confirmed by the analyses. However, only a weak relationship was found between air pollution and CHD mortality rates, and no direct association between urbanicity and mortality rates or female hospitalisation rates was found. The relationship between air pollution and hospitalisation rates was strongly significant, even after adjustment for both urbanicity and deprivation which confound the relationship, suggesting that air pollution may increase hospitalisation rates for CHD, but not significantly increase mortality rates. This may suggest that air pollution is a risk factor for milder forms of CHD that result in hospitalisation but are usually non-fatal - the pathophysiological mechanisms between air pollution and CHD are unclear but it has been suggested that pollution may increase heart rate variability which would require medical attention but is usually non-fatal (Pope, 2005).

The direct association between urbanicity and both CHD mortality and hospitalisation rates is via greater access to health services in more urban areas which in turn generates higher demand (Black et al., 1995). The conceptual framework suggests that, after deprivation and the behavioural risk factor profile of populations has been taken into account, individuals living in more urban areas are more likely to be hospitalised for CHD, and are hence less likely to die from CHD. Increased hospitalisation rates and lower mortality rates in metropolitan areas have only been demonstrated for men, but not for women, suggesting a gender inequality in access to health services that increases in more urban areas.

The models reported here suggest that colder, less sunny areas are likely to have higher CHD mortality and hospitalisation rates than warm, sunny areas, independent of the effect of both deprivation and the behavioural risk factor profile of the individuals living in the area. This observation may be due to residual confounding by untested variables: specifically, cultural differences that exist between the North and South of England. Because of the collinearity between these cultural differences and climate variables, an ecological analysis of this kind cannot effectively control for the cultural differences, however the (weak) independent association of hours of sunshine and CHD rates shown here provides some evidence that the results are not entirely due to residual confounding since the sunshine gradient in England is North West-South East, with lower levels of sunshine in the West of the country, which is often covered with cloud coming in from the Atlantic Ocean. Further analyses of CHD mortality rates in different countries where the cultural and temperature variables are not collinear will increase the understanding of the relationship between climate and CHD.

Behavioural risk factor profiles of populations and deprivation were shown to have independent associations with CHD mortality and hospitalisation rates in the analyses reported in this chapter. The conceptual framework suggests that deprivation does not have a direct association with CHD rates, therefore the independent association between deprivation and CHD rates has been interpreted as residual association due to a) non-inclusion of potentially explanatory variables that are associated with deprivation and b) inaccuracies in the measurement of the explanatory variables included in the model. This is likely to be the case since the behavioural risk factor profiles of populations used in

these analyses did not include a number of behavioural risk factors for CHD known to be associated with deprivation (e.g. diabetes, physical inactivity, high stress levels (Allender et al., 2008)), the standardised 'unhealthy lifestyle' index used to measure the behavioural risk factor profiles of populations is derived from synthetic estimates of risk factor prevalence rates that are known to contain measurement error (see chapter 6) and is based on aggregated individual-level estimates that do not consider the underlying distribution of individual-level risk factors (e.g. distribution of BMI in the population, distribution of smoking levels by age group etc.). However, the index has proved to be a useful indicator of the general health of a population that explains a large amount of geographic variation in CHD rates independently of deprivation.

The systematic approach to building models that was utilised here (developing models that incorporate environmental variables, behavioural risk factor profiles of populations and deprivation separately, and then all possible combinations) allowed for a comprehensive assessment of the impact of confounding, and for some disentanglement of the amount of geographic variation that is explained by the different sets of explanatory variables. Due to the multiple collinearities and associations that are apparent in the dataset, it was not possible to entirely disentangle how much of the variation in CHD mortality and hospitalisation rates is due to individual factors.

The results of the spatial error modelling suggest that the multi-level models were not subject to large biases despite the fact that they did not account for the spatial auto-correlation that was present in the dataset, since the two sets of results were in broad

agreement. This was not the case for the regression parameter estimates for the air quality index variable, which were substantially biased away from the null hypothesis in the results from the multi-level models. This is due to the fact that the air quality index variable was highly clustered (table 4.1, appendix one), following a broad rural / urban trend, and CHD rates in inner city areas are also likely to be clustered due to high inner city deprivation. Therefore, the ward level measure for the relationship between air pollution and CHD may have been inappropriately small in this case, resulting in the spatial autocorrelation bias. Despite the climate variables being highly clustered (table 4.1, appendix one) there was little evidence of spatial autocorrelation bias in the relationship between maximum temperature and CHD rates because the outcome variables were not clustered at the same geographical scale as the climate variables.

The results presented here are in general agreement with the UK literature on geographic variation in CHD rates, in that not all of the variation in CHD rates can be explained by compositional factors. Results from the British Regional Heart Study (BRHS) suggest that accounting for individual-level risk factors and social class explains between 50% and 80% of the geographic variance in CHD incidence rates in men (Morris et al., 2001) – similar to the 67% of large scale geographic variation in male CHD mortality rates explained by the behavioural risk factor profiles of populations and deprivation reported here (table 8.11, appendix three). The British Women's Heart and Health Study (BWHHS) estimated that accounting for individual-level risk factors and social class accounted for a reduction in the increased risk of CHD prevalence in the North of England by around 50% in women (Lawlor et al., 2003). An ecological study using all

local authorities in England and Wales found that the increased risk of CHD mortality from living in the North of England was reduced by 45% once differences in smoking, alcohol consumption and cold climate were accounted for (Law and Morris, 1998), which is less than the estimates found here (table 8.11, appendix three). The difference may be due to the geographic level of risk factor estimates used – Law and Morris used survey-based estimates at the regional-level rather than small area estimates such as those used here.

The BRHS provides the most comparable results for the impact of climate on geographic variation in heart disease in England, despite the widely differing methodology employed in the study compared with the analyses reported here. Analysis of phase one of the BRHS suggested that climate variables have a modest effect on variation in local CHD mortality rates after adjustment for individual-level variables (Pocock et al., 1980), which is a similar result to those reported here. Phase two of the study showed that after 17 years of follow-up maximum temperature explains around 30% of the between-towns variance in CHD incidence rates that remained after adjustment for social class and individual-level risk factors (Morris et al., 2001). Again, this is in broad agreement with the results reported here – that the climate has a modest effect on CHD rates after adjustment for differences in the behavioural risk factor profile of populations and socio-economic status.

A recent analysis of mortality rates for circularity disease between 2002 and 2004 for all super output areas in England found that the influence of urbanicity was removed after

adjustment for deprivation (Gartner et al., 2008), which supports the results presented here. Recent analyses of hospitalisation rates for CHD by urbanicity and deprivation are not available, but a study that investigated revascularisation rates in a number of health districts in the UK in 1992-93 found that rates were significantly higher in areas close to specialist cardiology centres (which were predominantly urban areas) independent of the deprivation of the area (Black et al., 1995). A similar finding was reported from a cross-sectional study of angiography and revascularisation rates from 180 Nottinghamshire general practices in 1993-97: those practices that were nearer to revascularisation centres had significantly higher angiography and revascularisation rates (Hippisley-Cox and Pringle, 2000). These two studies did not present gender-stratified results, but an earlier study of ward-level coronary artery bypass graft rates showed that rates in areas near a cardiothoracic unit were about 50% higher than rates in wards that are not near a cardiothoracic unit, after adjustment for deprivation, with no differences for male and female rates (Ben-Shlomo and Chaturvedi, 1995), which does not support the findings reported here. Gender inequalities in access to CHD health services in the United Kingdom have previously been reported (MacLeod et al., 1999; Gatrell et al., 2002; Petticrew et al., 1993; Dong et al., 1998), but such inequalities would not explain the findings in this chapter unless the inequalities were differential by urbanicity. The results reported here regarding the different association between hospitalisations and urbanicity between men and women are not supported by the literature, and may be due to the conflicting associations between deprivation and the unhealthy lifestyle variable with urbanicity (see table 8.2). This gender differential warrants further investigation.

The models reported in this chapter suggest that air pollution is strongly associated with hospitalisations for CHD, but only weakly associated with CHD mortality. A recent UK study using a similar ecological design found contradictory results: namely that air pollution (specifically high nitrous oxide levels) is associated with CHD mortality but not hospital admissions, and the authors speculated that this may be because air pollution-related CHD is more likely to cause sudden death (Maheswaran et al., 2005). The study used a finer geographical scale than the analyses reported here (census enumeration districts, areas with a mean population of around 150). In addition, the study was restricted to an industrial city (Sheffield); the results are therefore not entirely comparable to those reported here (which incorporated rural areas in the analysis). The contradictory results suggest, though, that the air pollution results reported here should be treated with caution, as they may be a result of confounding with urbanicity and hence proximity to hospitals.

CONCLUSIONS

Large scale geographic variation in CHD mortality and hospitalisation rates is mostly due to large scale differences in both socio-economic deprivation and prevalence of unhealthy lifestyles, but some of the remaining variation is explained by differences in the climate, primarily outdoor air temperature. The environmental variables included here have little impact on small scale variation in CHD mortality rates but both the degree of urbanicity and air pollution levels explain some of the small scale variation in hospitalisation rates, independently of behavioural risk factor profiles of populations and deprivation. The

index of unhealthy lifestyle of populations developed for these analyses has been shown to be strongly associated with CHD rates independently of deprivation and is therefore a potentially useful measure for future studies of geographic variation in chronic disease in England.

INTRODUCTION

The analyses for this thesis were designed to explore the impact of environmental variables on geographic variation in coronary heart disease (CHD) rates after accounting for the behavioural risk factor profile of populations and other potential confounding factors (proxied by area-level deprivation). The research questions stated in chapter three require a quantification of the amount of geographic variation in CHD mortality and hospitalisation rates in England that is explained by differences in environmental variables. This chapter explores the extent to which these questions have been answered, and describes the important contributions to scientific knowledge that the various analyses have provided. The general strengths and limitations of the study, and recommendations for future work are then discussed.

Results of the thesis

Two local climate measures (mean daily maximum temperature and total hours of sunshine), a measure of air pollution, and a rural / urban / metropolitan categorisation were found to explain - without accounting for other factors - nearly 60% of large scale geographic variation in CHD mortality rates, and around 30% of CHD hospitalisation rates on their own, but did little to explain small scale geographic variations in CHD rates. Their association with CHD rates was substantially reduced when both deprivation

and behavioural risk factor profiles of populations were added as explanatory variables, particularly for those that affected small scale geographic variation (air pollution and urbanicity). A substantial amount of large scale geographic variation in CHD rates is explained by environmental variables even after adjustment for deprivation and behavioural risk factor profiles of populations – at least 15% of large scale variation in mortality rates (see tables 8.3, chapter eight and 8.10, appendix three). Environmental variables are less effective at explaining small scale variations in CHD rates: air pollution was not significantly associated with male mortality rates, urbanicity was not significantly associated with female mortality or hospitalisation rates after adjustment for deprivation and behavioural risk factor profiles of populations.

The impact of environmental variables on large scale geographic variations in CHD predicted by the analyses in this thesis is likely to be an over-estimate because the climate variables used as explanatory variables follow a similar North-South gradient as the outcome variables. Therefore, the climate variables are likely to be correlated with any remaining variance in the modelled estimates that has not been explained by either deprivation or behavioural risk factor profiles of populations because of either missing variables that could have an impact on CHD (e.g. prevalence of diabetes) or poor measurement of the variables that were included in the models. Poor measurement of the behavioural risk factor profiles of populations is likely to have produced a lot of residual 'noise' – as discussed in previous chapters, aggregate measures for populations do not account for the underlying distribution of individual-level risk factors (e.g. the distribution of body mass index (BMI) for an area is lost in a simple measure of the

prevalence of BMI ≥ 30kg/m^2 as was used in these analyses), and synthetic estimates of prevalence rates at ward level are likely to result in a substantial amount of measurement error, even though only synthetic estimates that were found to be valid and accurate were included in the analyses in chapters seven and eight. A measure of deprivation was included in the models in order to explain the residual variance that was left due to the problems with the behavioural risk factor profile of populations measures, but the deprivation variable itself can only be a proxy of the missing risk factor variables, or of the risk factor elements lost in measurement error, and it is therefore not likely to be able to fully account for all such residual variance – some is likely to have been accounted for by the environmental variables.

This over-estimation is demonstrated by the small yet significant associations found between sunlight and CHD mortality rates in the final models. One proposed mechanism for the influence of sunlight on CHD is that sunlight acts as a catalyst to convert a precursor of cholesterol (squalene) into vitamin D, and hence reduces cholesterol levels in the blood (Grimes et al., 1996). If this mechanism is correct, then cholesterol levels should fall on the causal chain between sunlight exposure and CHD mortality rates. Therefore an adequate accounting for the role of cholesterol levels in geographic variations in CHD mortality rates should remove the correlation between sunlight and CHD mortality rates.

It was the aim of the analyses of this thesis to estimate only the *direct* impact of environmental variables on CHD rates. Therefore, associations between environmental

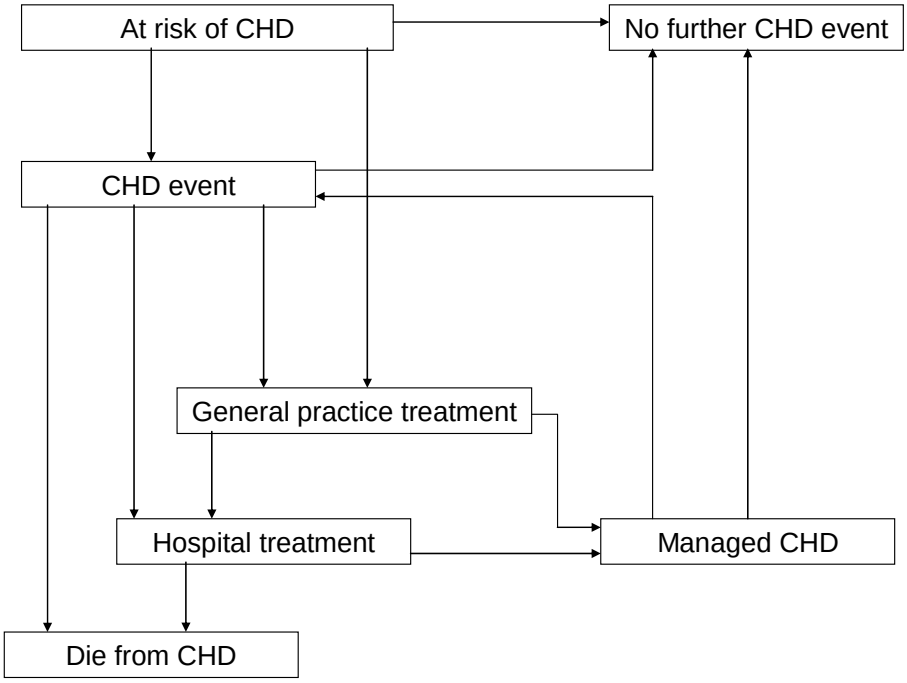
variables and CHD that are mediated by risk factors for CHD should not be estimated by the regression models described in chapter eight. Macintyre et al. argue that this method of including individual-level factors that are on the causal pathways between environmental variables and health outcomes as 'confounding' factors inevitably leads to an under-estimation of the total (direct and indirect) impact of environmental variables (Macintyre et al., 2002). The analyses in chapter five do not include behavioural risk factor profiles of populations or deprivation, so could be interpreted as an assessment of the total impact of environmental variables, yet the amount of genuine confounding with differences in behavioural risk factor profiles of populations that are not on the causal pathway between environmental variables and CHD is likely to be large, hence the results presented in those models are likely to be substantially over-estimated.

Alternatively, since the synthetic estimates of the prevalence of behavioural risk factors for CHD should actually be interpreted as *the prevalence rate that would be expected given the social, economic and demographic structure of the population* it could be argued that the estimates of the impact of environmental variables on CHD derived from this thesis actually include indirect effects of environmental variables. This is because any increase in cholesterol levels that are a result of low exposure to sunlight, say, will not be included in the synthetic estimate of prevalence of cholesterol which is only based on social, economic and demographic variables for the population. But this interpretation is problematic, since the synthetic estimates are based on logistic models that were built using national survey data and have been validated against national and local survey data, and survey-based data does not distinguish between increased prevalence of risk factors

that are a result of environmental or social, economic and demographic factors. Indeed, some of the models built to generate the synthetic estimates used in this thesis include variables that modify the prevalence estimates on the basis of government office region of residence (Twigg et al., 2004), which explicitly accounts for locality as well as social, economic and demographic data.

The models presented in this thesis had two different outcomes: CHD mortality rates and hospitalisation rates. These different outcomes warrant separate investigation since the geographic patterns in CHD mortality and hospitalisation rates in England are different (Scarborough et al., 2008). Also, studying both of these outcomes allows for some interpretation of the form and severity of CHD that is associated with the different environmental variables. Figure 9.1 shows a framework for the development and treatment of CHD for individuals who are at risk. Each of the arrows in the framework represents a potential path from one state to another and the probability of following these paths can be influenced by a number of variables. For example, it is known that prescribing rates for CHD drugs in England show significant geographic variation (Ward et al., 2004), which could affect the possibility of moving from 'Managed CHD' to 'CHD event' or 'No further CHD event' in figure 9.1.

Figure 9.1 Framework for development and treatment of CHD



By comparing the models for hospitalisation rates and mortality rates it is possible to estimate whether the environmental variables influence the potential pathways from one state to another. This allows for inferences to be made regarding how the environmental variables influence the course of CHD. This is most apparent for the urbanicity variable, where metropolitan wards had significantly lower male mortality rates but significantly higher male hospitalisation rates when compared with coastal and countryside wards. This is likely to be due to greater access to health services in more urbanised wards. This increases hospitalisation rates by reducing out-of-hospital CHD deaths (since transport to hospital is quicker) and by lowering GP referral rates to hospital because of increased supply which in turn influences demand (this has been shown to be a factor for referral

rates for CHD investigations and revascularisations (Hippisley-Cox and Pringle, 2000; Black et al., 1995; Ben-Shlomo and Chaturvedi, 1995)). The reduced mortality rates are likely to be a result of lower case fatality rates due to the increased level of treatment. The regression parameters for air pollution were much stronger (explaining more of the variance) for hospitalisation rates than for mortality rates, suggesting that air pollution causes less severe forms of CHD that are less likely to result in out-of-hospital sudden death, whereas the regression parameters for temperature were stronger for mortality rates, suggesting that a cold climate increases the chances of sudden out-of-hospital mortality.

Contribution to knowledge

This thesis contributes to the scientific understanding of geographic variations in CHD rates in England in three important ways. Firstly, it extends the results of the British Regional Heart Study (BRHS) to include men and women of all ages and all areas within England. Secondly, it provides an assessment of the validity of a series of synthetic estimates of the prevalence of behavioural risk factors for CHD that are currently used to guide healthcare resources in England. Thirdly, it introduces an important standardised index of unhealthy lifestyle for wards in England that could be used for future ecological studies of chronic disease.

Confirmation and extension of results of the British Regional Heart Study

The BRHS is the most in-depth investigation of geographic variation in coronary heart disease in Britain that has ever been conducted. Phase one of the study gathered aggregated data for 253 towns in Britain identified to have a population of between 50,000 and 100,000 at the 1971 census (Pocock et al., 1980). Standardised mortality ratios for cardiovascular disease for men and women aged 35-74 were constructed and compared with data on water hardness taken from local samples; rainfall, mean temperature and total sunshine gathered from 164 nearby weather stations; air pollution taken from 134 smoke and sulphur dioxide reporting sites; and fifteen socioeconomic measures taken from the 1971 census. Five of these measures were found to have statistically significant associations with cardiovascular disease rates in multivariate analyses - water hardness, rainfall, mean temperature, percentage of manual workers and car ownership – and these five variables explained 78% of the variance in the cardiovascular disease rates. Phase two of the BRHS recruited a cohort of 7,735 men aged 40-59 between 1978-80 from 24 small towns in Britain (Morris et al., 2001). Measurements of blood pressure, height, weight and total cholesterol, and questionnaire responses regarding smoking, alcohol intake, physical activity and social class were collected at baseline and at a number of subsequent screenings. Measures of water hardness were taken from samples, and weather stations located close to the 24 towns provided data on maximum daily temperature, minimum daily temperature, total rainfall and total sunshine. After fifteen years of follow-up, 79% of the between-towns variance of CHD incidence was explained by smoking, social class, blood pressure, physical

activity, height and body mass index (although further inclusion of alcohol intake and cholesterol levels reduced the explanatory power of the model). Between 9% and 30% of the remaining between-towns variance was explained by the different environmental variables (each considered separately).

The analyses described in this thesis broadly concur with the results from both phase one and phase two of the BRHS, in that climate variables were found to have a substantial significant association with CHD rates in analyses unadjusted for behavioural risk factor profiles of populations (phase one) and environmental variables were shown to explain a small but significant amount of geographic variation in CHD rates after adjustment for behavioural risk factor profiles of populations (phase two). The major difference in the models explored in the BRHS and those reported here is the presence of water hardness as an explanatory variable. However, a recent analysis of the BRHS dataset suggests that water hardness does not protect against cardiovascular disease (Morris et al., 2008), so the absence of water hardness from the models reported here is unlikely to be a major drawback.

The results of this thesis extend the results of phase one of the BRHS in the following ways: all wards in England were included in the analysis; a measure of the behavioural risk factor profile of populations of areas was included; an exploration of both small scale and large scale geographic variation in CHD rates was conducted; including wards from rural areas allowed for urbanicity to be included as a potential explanatory variable; more sophisticated estimates of air pollution and climate were used, which allowed for

modelled estimates of these measures to be applied to all wards in England; both hospitalisation rates and mortality rates were used as outcome variables. The results of this thesis complement the results of phase two of the BRHS, and extend the interpretations in similar ways to how they extended the results of phase one. In addition, the results here were reported for men and women separately, and were sufficiently powered at the area-level to allow for inclusion of several environmental variables in the models simultaneously.

Validation of synthetic estimates

The development of the synthetic estimation technique is a fairly recent phenomenon (Heady et al., 2003), but it has proved popular. As shown in table 3.4 (chapter three) many different sets of synthetic estimates of the prevalence of risk factors for CHD at ward-level in England have been developed. In addition to the sets shown in table 3.4, synthetic estimates of smoking, binge drinking, fruit and vegetable consumption and obesity estimates have been developed by the National Centre for Social Research (Bajekal et al., 2004; Pickering et al., 2004), and synthetic estimates of the prevalence of obesity have been developed by researchers at the University of Portsmouth (Moon et al., 2007). The purpose of these sets of synthetic estimates is to help guide healthcare resource allocation at a small area level. Many have been published in Government documents aimed at identifying areas where healthcare resources should be directed, such as the *Health Profile of England* (Department of Health, 2006) and *The Smoking Epidemic in England* (Twigg et al., 2004). Many are published on a website funded by

the National Health Service and the Association of Public Health Observatories *(The Health Poverty Index)* designed to display health differentials between small areas (Dibben et al., 2004). The set of synthetic estimates of diabetes are published on the Yorkshire and Humber Public Health Observatory website, and are aimed at identifying areas which are likely to have high diagnosed and undiagnosed diabetes prevalence (YHPHO, 2005).

Despite their widespread use, many of these sets of synthetic estimates have not been through a formal process of validation. This is important, as it is acknowledged that the synthetic estimation technique is only likely to produce valid and accurate estimates when there is a well-mapped relationship between the risk factor under investigation and individual-level and area-level explanatory variables, and the model that supports the synthetic estimates is a reasonable representation of this relationship (EURAREA Consortium, 2004). The validation exercise conducted for this thesis (chapter six) showed that many of the sets of synthetic estimates that are in common use display substantial signs of invalidity and are therefore unlikely to be accurate. The analyses also showed that those sets of synthetic estimates that were considered to be valid are likely to be biased towards the mean (that is, they under-estimate for areas with high prevalence rates and over-estimate for areas with low prevalence rates) and therefore synthetic estimates for individual areas should be treated with caution.

Introduction of unhealthy lifestyle index

This thesis introduced a standardised index of unhealthy lifestyle based on the results of a principal components analysis of the age-standardised synthetic estimates of low fruit and vegetable consumption, smoking, obesity, raised blood pressure and raised cholesterol. In general, ecological analyses of chronic disease outcomes use a deprivation index as a proxy of differences in the behavioural risk factor profile of populations, since many risk factors for CHD are associated with socio-economic status. But this is likely to be a poor proxy as deprivation indices are measures of the economic and social environment, rather than of the health environment. The standardised index of unhealthy lifestyle introduced here aims to be a more direct measure of the differences in behavioural risk factor profiles of populations.

In analyses displayed in chapter eight, the standardised index of unhealthy lifestyle was shown to be highly significantly associated with CHD mortality and hospitalisation rates for both men and women, after adjustment for deprivation. This suggests that the unhealthy lifestyle variable adds explanatory power to the models beyond that provided by the deprivation index, which should be expected since the two indices are measuring different phenomena. This is clearly shown in table 8.2 (chapter eight), where the deprivation index and the unhealthy index are shown to have strongly opposing values in metropolitan wards, implying that heavily urbanised, inner-city wards are heavily economically deprived, but their populations follow a reasonably healthy lifestyle. This is likely to be due to the high residency of ethnic minority groups in inner-city wards

(around 26% non-white in metropolitan wards at the 2001 census), who tend to have a healthier diet, lower smoking levels (particularly for females), more modest alcohol intake, lower blood pressure and lower cholesterol levels than the general population (Allender et al., 2008).

This thesis has demonstrated that an index of unhealthy lifestyle can be constructed that explains geographic variations in CHD rates independently of indices of economic and material deprivation. As such, the unhealthy lifestyle index adds explanatory power to models exploring geographic variation in CHD rates. The behavioural risk factors that were included in the unhealthy lifestyle index are also risk factors for other cardiovascular diseases, such as stroke, so the unhealthy lifestyle index could also be applied to ecological studies of other cardiovascular diseases. Also, the methods used to develop the unhealthy lifestyle index could be adapted for ecological studies of other chronic diseases with different behavioural risk factors, such as cancer.

Strengths and limitations

This thesis is the first instance of a study of geographic variation in small area CHD rates that accounts for behavioural risk factor profiles of populations, deprivation, and environmental variables for a large geographical area (in this case, England) within the same set of analyses. The multi-level design of the analyses allowed for the explanation of large scale and small scale geographic variation in CHD rates simultaneously, which allowed for disentanglement of the influence of environmental variables that are effective

at the different scales. The spatial error models allowed for an assessment of whether the multi-level models were prone to spatial autocorrelation bias, which was shown not to be the case.

An important limitation of this study was the use of purely ecological data for the analyses. It has long been known that the use of ecological data to model individual-level results can produce severe bias. This is because the relationship at area-level can be vey different to the relationship at individual-level. The potential difference is displayed well by Oakes, who produces a figure of the relationship between income and probability to vote Republican in three American states using artificial data based on an example first described by Gelman et al. (Oakes, 2009; Gelman et al., 2007). Here, in each of the three states the individual-level relationship between income and probability of voting Republican is positive, but the average income of individuals in the states and the average probability of voting Republican are set so as to produce a negative association at state-level. This potential difference between associations at individual-level and area-level implies that there is also a danger when assuming area-level relationships using only individual-level data. Using area-level data to model individual-level associations can result in the 'ecological fallacy', whereas using individual-level data to model area-level associations can result in the 'individualistic fallacy'. These two fallacies are different sides of the same coin, and are sometimes both termed together as 'cross-level fallacies' (Subramanian et al., 2009). The most appropriate way to model the relationships between areas and individuals is to use a multi-level structure that incorporates both individual-level and area-level data, thereby allowing for an assessment of the size of the area-level

association after adjustment for the individual-level association (and vice versa). Although the analyses conducted for this thesis were multi-level, they did not incorporate individual-level data (the behavioural risk factor, CHD mortality and CHD hospitalisation data were all aggregate data), since adequate individual-level data that included all individuals within England in 2001 were not available. Therefore, the interpretation of the results of this thesis has been restricted to the area-level: that is, the results can only tell us about how differences between wards in England affect the geographic variation in CHD rates. The results can not provide any information about how the explanatory variables affect individuals (Subramanian et al., 2009). For example, the results imply that the average temperature of an area has an impact on CHD rates within that area, but they do not tell us anything directly about how the temperature of an area affects the individuals living in the area, or whether certain individuals within the area are more at risk than others. Where interpretations of this kind have been made in this chapter and elsewhere in the thesis they are based on supposition, drawing on evidence reported elsewhere in the scientific literature.

The ecological cross-sectional study design utilised for this thesis is useful for studying a number of different environmental variables simultaneously, but it has its limitations, such as the lack of individual-level data necessitating the use of aggregated measures of risk factor profiles that do not take account of the underlying distribution of the risk factor in question. Cross sectional studies collect data on explanatory variables and outcomes simultaneously – they are therefore only capable of displaying correlations between the nominated 'explanatory' and 'outcome' variables. One of the necessary

scientific requirements for establishing a causal relationship is that the explanatory variable can be shown to occur before the outcome variable. However, in instances where there is a strong theoretical justification for regarding a relationship between two variables as causal then cross-sectional studies are useful ways of measuring the strength of this relationship. This is the case for the relationship between behavioural risk factor profiles of populations and CHD rates. The causal relationship between behavioural risk factors and CHD has been well-established at the individual-level (Stamler, 2005), and it is therefore theoretically reasonable to assume that the causal relationship at the individual-level should translate to a similar causal relationship at the area-level. As is demonstrated in this thesis by the lack of association between the prevalence of raised cholesterol and CHD rates, it is not always the case that strong associations at the individual-level should translate to strong associations at the area-level.

The relationship between the environmental variables under investigation and CHD is not so well established and therefore a causal relationship cannot be assumed with certainty from the results of the analyses in this thesis. For example, the discussion sections of this thesis have suggested that the positive association between urbanicity and hospitalisations and negative association with mortality shown for men in this thesis is causal, in that greater access to health care in urban areas results in greater uptake of those resources (i.e. higher hospitalisation rates) which in turn leads to improved treatment, lower case-fatality rates and hence lower CHD mortality rates. An alternative explanation of these findings is that there may be a tendency for individuals who develop CHD to migrate to urban areas to have greater access to healthcare resources, then later migrate from urban

areas to rural areas at older ages for a more peaceful lifestyle. Under this explanation, the relationship between urbanicity and CHD mortality and hospitalisation rates is not causal, it is confounded by the flux of migration of individuals at different disease stages.

A further problem of cross-sectional analyses is that they only consider a snapshot of explanatory and outcome variables, and therefore they are poorly suited to consider conditions which have a prolonged latent period between exposure and disease, such as CHD. The life course approach to chronic disease epidemiology suggests that development of CHD in an individual is dependent upon an accumulation of risk (from both behavioural and environmental risk factors) over the entire course of earlier life, including before birth (Kuh and Davey Smith, 2004; Lawlor et al., 2004). Cross-sectional studies that collect data on individuals record only current risk factor status – sometimes accompanied with recall data of earlier risk factor status. The effect of dismissing or inaccurately measuring risk exposure throughout the life course is that analyses will be biased towards under-estimating the cumulative impact of risk factors on health outcomes.

The equivalent situation in the ecological analyses performed for this thesis is that the snapshot of the explanatory variables for a ward does not accurately reflect the life course exposure of the population currently residing in the ward to these risk factors, and hence measures of the association between the explanatory variables and CHD rates would be under-estimated. The analyses would be particularly vulnerable to this bias if either the ward-level explanatory variables were not stable over time (and therefore did not

accurately reflect the exposure of the ward population in earlier times), or if migration between wards was non-negligible (and therefore a sizeable proportion of the population within the ward may have been exposed to risk factors corresponding to other wards at earlier stages of their life). The environmental variables studied in this thesis are likely to have been stable over recent history, with the possible exception of air pollution which could drastically change in certain areas that experience sentinel events (e.g. opening or closing of a coal-powered power station). However, as the discussion in chapter six demonstrated, it is not clear that the behavioural risk factor profiles of populations of areas in England have remained stable over time. Similarly, migration between wards is clearly not insignificant (4% of people within the UK moved residence to a new location over 10km away in 2000 (Office for National Statistics, 2007)), although it is unclear to what degree individuals are likely to migrate between wards that share similar behavioural risk factor profiles or environmental variables. An investigation of the BRHS dataset showed that current residence was more strongly associated with CHD events than zone of birth, and migrants were found to have a similar risk of CHD as individuals who have always lived in the towns under investigation (Wannamethee et al., 2002), similarly current residence was a much strong predictor of blood pressure levels than residence at birth (Elford et al., 1990). This suggests that individuals who move tend to adopt similar risk levels to the population which they have migrated to, and lose the risk level of the population which they have migrated from. If this is the case, then the influence of migration on the findings reported here is not likely to be substantial.

The ecological study design requires explanatory and outcome data to be available for a set of small areas within the larger geographical area of interest. Ideally the choice of the set of small areas should be based on theoretical grounds, either to ensure the maximum possible homogeneity within the small areas (and hence the maximum variance in the explanatory variables) or to reflect the local population's understanding of 'neighbourhood'. In practice, data are usually only available for a set of administrative areas that were designed with very different purposes in mind, such as the set of wards used in this thesis. It is well known that the choice of the set of areas can affect the results of the ecological analysis – this is known as the modifiable areal unit problem (Stafford et al., 2008). The impact of using a set of areas that have not been devised specifically for the analysis is to increase the heterogeneity of the population resident in the areas, and hence to reduce the variance of both the aggregated explanatory and outcome variables. This tends to bias the results of an ecological analysis towards the null hypothesis. Researchers investigating geographic variations in smoking, obesity, alcohol intake, walking and self-rated health in the Camden and Islington district of London found that geographic inequalities measured using wards were not substantially different to those found using two other area measures based on maximum possible homogeneity of population, and local definitions of neighbourhood (Stafford et al., 2008) and conclude that '*we can have … confidence in the results of numerous studies which have used administrative boundaries to define the neighbourhood'*. This suggests that the choice of wards as units of analysis for this thesis should not have substantially biased the results.

A related problem regards the variation in the population size of the wards, which were designed for the purposes of local elections and not with ecological analyses in mind. The smallest wards consist of only around 1,000 people, whereas the largest contain over 30,000. This is problematic, since larger wards have more CHD events, and hence more accurate estimates of underlying CHD rates. It would be possible to address this by weighting the analyses on the basis of either population size or number of events. However, to do so only addresses the variance in the model residuals introduced by sampling errors, whereas the total variance also has an unexplained component (the variance that is due to missing explanatory variables, poorly measured explanatory variables, and random variance) – therefore to weight analyses purely to account for sampling errors could result in over-compensation for this problem. This issue has previously been addressed using the dataset from phase one of the British Regional Heart Study, where models were recalculated using weights based on population size and compared with the original model results (Pocock et al., 1982). The authors found little difference in the results and concluded that it was more appropriate to use unweighted regression techniques in a situation where the outcome variables are based on a substantial number of events. As shown in table 3.2 (chapter three), the six year data collection period for the outcome variables used in this thesis should have ensured a substantial number of events for the majority of wards.

There were also a number of limitations that were specific to the analyses conducted here that should be acknowledged. For instance, the analyses take no account of geographic variation in treatment rates for CHD. It is known that prescribing rates for CHD drugs

show significant geographic variation within the UK (Ward et al., 2004) which could influence both CHD mortality and hospitalisation rates. Similarly, referral rates for investigation and revascularisation rates for CHD also show significant geographic variation (Hippisley-Cox and Pringle, 2000; Black et al., 1995; Ben-Shlomo and Chaturvedi, 1995). Local data on prescription rates for the entire of England are unavailable so could not have been included in the analyses, but it would have been possible to include investigation and revascularisation rates as a potential explanatory variable in the mortality models. However, such rates are a product of factors including need, supply and local referral thresholds. The supply factor is already proxied in the models by the urbanicity variable, and the need factor is what the explanatory variables are trying to explain. In other words, to a large degree the investigation and revascularisation rate is on the causal pathway between the explanatory and outcome variables, and inclusion in the model would result in collinearity and double-counting. Local referral thresholds for both hospitalisation and for investigation and revascularisation would be valuable explanatory variables, but unfortunately suitable data were not available. It has been suggested that local prescribing, intervention and revascularisation rates may be negatively associated with need after appropriate adjustments for risk factor levels and deprivation, and therefore follow the inverse care law (Ward et al., 2004; Ben-Shlomo and Chaturvedi, 1995). If this were the case, then access to treatment may be an additional causal factor for the positive associations between behavioural risk factor profiles of populations and deprivation with CHD mortality rates that are reported here.

As was mentioned in chapter three, the explanatory variables for this thesis were all model-based (generated by either trend surface models, cluster models or synthetic estimation models), and as such are not direct estimates of the phenomena under investigation. It is inevitable, therefore, that they should be accompanied by some degree of error. If we assume that this error is analogous with measurement error (i.e. randomly generated and normally distributed with mean zero) then the impact is to increase the variance of the explanatory variables, and therefore to bias the assessments of their association with CHD rates towards the null hypothesis. Since the error accompanying the sets of synthetic estimates is larger than that accompanying the environmental variables, we would expect the bias accompanying the behavioural risk factor profile of populations effect estimates to be larger than for the environmental variables. This would result in an over-estimate of the amount of geographic variation in CHD rates that could be explained by environmental variables. However, we know that the error accompanying the synthetic estimates is *not* analogous with measurement error, as it has been observed that the total variance of the sets of synthetic estimates are *lower* than would be expected (see chapter six). Reduced variance in the synthetic estimates would bias the assessment of the association between behavioural risk factor profiles of populations and CHD rates away from the null hypothesis, as was demonstrated with the predictive validity assessment on the synthetic estimates of raised cholesterol in chapter six. This bias has been addressed somewhat in the final models reported in chapter eight by introducing the standardised index of unhealthy lifestyle. But it is unclear whether the error accompanying the environmental variables is analogous to measurement error, and hence whether the accompanying bias of the effect estimates would be towards or away

from the null hypothesis. It is likely, though, that the error accompanying the environmental variables is fairly small – similar trend surface modelling of environmental variables have shown that the resultant models produce very small residuals (e.g. a trend surface model that predicted UV radiation accounted for 96% of the variance in measures from the UV detectors (Scwartz and Hanchette, 2006)).

The hospitalisation rates were based on data derived only from NHS hospitals. Referrals to private practice account for around 10% of all general practice referrals in England (Mulvaney et al., 2005), although it is unclear whether this percentage is relevant to hospital admissions for CHD, which are often emergency admissions. Since private referrals are more common in more affluent areas (Mulvaney et al., 2005) this is likely to have resulted in an over-estimate of the association between hospitalisation rates and deprivation, although this bias is likely to be small because of the relatively small use of private hospitals in England.

The proposed mechanism for the impact of urbanicity on CHD rates is via increased access to healthcare resources in more urban areas. It would have been more appropriate to include a direct measure of access to healthcare resources, such as distance to nearest hospital, or density of GP surgeries. However, it was not possible to secure data on location of either GP surgeries or hospitals that covered the data collection period for this thesis. Use of the urbanicity variable as a proxy for access to healthcare is likely to have biased the association with CHD rates towards the null hypothesis.

Recommendations

One of the conclusions of this thesis is that the climate has an impact on CHD mortality and hospitalisation rates that is independent of both deprivation and behavioural risk factor profiles of populations. The temporal association between cold weather and CHD is well known (excess CHD mortality of around 2,000 in England and Wales in winter months of 2004/05 compared to summer mortality levels, for example (Allender et al., 2008)). The analyses here suggest that, on top of excess winter mortality, there is a general annual increase in the CHD mortality rate of 40 deaths per 100,000 in men and 25 deaths per 100,000 in women in the coldest parts of England compared to the warmest. Whilst this impact is small compared to differences in the lifestyle of populations, it could still be an area which could be targeted in order to reduce geographic inequalities in CHD. Analyses of excess winter mortality in different regions of Europe have shown that the excess mortality is generally greater in countries with milder climates and this has led researchers to suggest that the impact of a cold climate on cardiovascular health can be substantially reduced if the population were better prepared for cold weather by improving household heating and insulation and wearing more appropriate clothing during cold periods of the year (Keatinge et al., 1997; Mercer, 2003). Interventions such as these would be beneficial for reasons other than improving cardiovascular health. Cold weather has been implicated in the development of a number of conditions such as respiratory disease, particularly in elder people. Improvements in home heating have the potential to improve quality of life, and increased insulation of

homes would reduce fuel use thereby saving household finances and reducing greenhouse gas emissions.

Further work should be conducted on developing and validating the standardised index of unhealthy lifestyle introduced in this thesis. There are strong theoretical grounds for including an index that measures the prevalence of behavioural risk factors in ecological studies of chronic disease, rather than an index of deprivation which is generally used as a proxy for (amongst other things) the behavioural risk factor profile of a population. However, the unhealthy lifestyle index used in this thesis is only suitable for ecological studies set in England that use wards as the units of analysis, and it is best placed for the study of cardiovascular disease (since, for example, the index includes a measure of the prevalence of raised blood pressure which has little theoretical association with cancer rates). Possible areas of development of the index include an exploration of the impact of weighting the sets of synthetic estimates using proportion attributable fractions for the impact of individual-level risk factors on chronic disease developed for the Global Burden of Disease project (World Health Organization, 2002) as the basis for the weighting. Also, efforts to incorporate the synthetic estimates of unhealthy behaviour developed by the National Centre for Social Research (Bajekal et al., 2004) which were not used in this thesis but have been subjected to various validity assessments (Pickering et al., 2004), should be explored. This may allow for inclusion of synthetic estimates of binge drinking within the unhealthy lifestyle index. Finally, the unhealthy lifestyle index should be updated when new sets of synthetic estimates – that have been appropriately

validated – are developed once the data from the 2011 census become generally available.

The synthetic estimates for physical inactivity that are published on the Health Poverty Index website (Dibben et al., 2004), and the synthetic estimates of diabetes prevalence developed by the Yorkshire and Humberside Public Health Observatory (YHPHO, 2005) have been shown in this thesis to be invalid. For the physical inactivity synthetic estimates, this is because important interaction terms between gender and social class were not included in the modelling process. The diabetes synthetic estimates were based on studies of diabetes prevalence in Coventry and Brent in the early 1990s (Simmons et al., 1991; Chaturvedi et al., 1993) that do not seem to be representative of the situation in England in 2001. Further work should be conducted to develop valid and accurate synthetic estimates of physical inactivity and diabetes prevalence at ward-level in England. Physical inactivity is an important risk factor for many chronic diseases – development of valid and accurate synthetic estimates of physical inactivity would therefore aid the development of an index of unhealthy lifestyle. Valid and accurate synthetic estimates of the prevalence of diabetes would allow for a better targeting of health care resource allocation, and would also be beneficial to an index of unhealthy lifestyle.

Allender S, Peto V, Scarborough P, Kaur A, Rayner M. *Coronary heart disease statistics 2008.* British Heart Foundation: London, 2008.

Allender S, Scarborough P, Keegan T. Differences in deprivation between local authority is more closely associated with CHD mortality than relative deprivation within local authorities. Forthcoming.

Altman D. *Practical statistics for medical research.* Chapman & Hall: London, 1991.

Anselin L. *GeoDa version 0.9.* University of Illinois: Illinois, 2003.

Bajekal S, Scholes S, Pickering K, Purdon S. *Synthetic estimation of healthy lifestyle indicators: stage 1 report.* Department of Health: London, 2004.

Bartley M. *Health inequality.* Polity Press: Cambridge, 2004.

Ben-Shlomo Y, Chaturvedi N. Assessing equity in access to health care provision in the UK: does where you live affect your chances of getting a coronary artery bypass graft? *Journal of Epidemiology and Community Health*, 1995; 49: 200-204.

Black N, Langham S, Petticrew M. Coronary revascularisation: why do rates vary geographically in the UK? *Journal of Epidemiology and Community Health,* 1995; 49: 408-412.

Bush T, Tsagatakis I, King K, Passant N. *NAEI UK emission mapping methodology 2006.* DEFRA: London, 2008.

Capewell S, Allender S, Critchley J, Lloyd-Williams F, O'Flaherty M, Rayner M, Scarborough P. *Modelling the burden of cardiovascular disease to 2020.* Cardio & Vascular Coalition and the British Heart Foundation: London, 2009.

Carstairs V, Morris R. Deprivation and health in Scotland. *Health Bulletin,* 1990; 48: 162-175.

Chadwick E. *Report on the sanitary conditions of the labouring population of Great Britain.* London Poor Law Commission: London, 1842.

Chaturverdi N, McKeigue P, Marmot M. Resting and ambulatory blood pressure differences in Afro-Caribbeans and Europeans. *Hypertension,* 1993; 22(1): 90-96.

Cheng C-L, Van Ness J. *Statistical regression with measurement error.* Arnold: London, 1999.

Cliff A, Ord J. *Spatial autocorrelation.* Pion: London, 1973.

Congdon P, Shouls S, Curtis S. A multi-level perspective on small-area health and mortality: a case study of England and Wales. *International Journal of Population Geography,* 1997; 3: 243-263.

Cressie N. Geostatistical methods for mapping environmental exposures. In: Elliott P, Wakefield J, Best N, Briggs D (eds). *Spatial epidemiology.* Oxford University Press: Oxford, 2000.

Davey Smith G, Phillips A. Confounding in epidemiological studies: why "independent" effects may not be all they seem. *British Medical Journal,* 1992; 305: 757-759.

Davey Smith G, Hart C, Watt G, Hole D, Hawthorne V. Individual social class, area-based deprivation, cardiovascular disease risk factors, and mortality: the Renfrew and Paisley Study. *Journal of Epidemiology and Community Health,* 1998; 52: 399-405.

Davey Smith G, Hart C. Life-course socioeconomic and behavioural influences on cardiovascular disease mortality: the collaborative study. *American Journal of Public Health,* 2002; 92 (8): 1295-1298.

Department of the Environment, Transport and the Regions (DETR). *Measuring multiple deprivation at the small area level: the indices of deprivation 2000.* DETR: London, 2000.

Department of Health. *National Service Framework for coronary heart disease – modern standards and service models.* Department of Health: London, 2000a.

Department of Health. *Health Survey for England 1998: Cardiovascular disease.* The Stationery Office: London, 2000b.

Department of Health. *Health Survey for England 2001.* The Stationery Office: London, 2003.

Department of Health. *Health Survey for England 2003.* The Stationery Office: London, 2004.

Department of Health. *Health Profile for England.* The Stationery Office: London, 2006.

Department of Health. *Hospital Episode Statistics 2006/07.* www.hesonline.nhs.uk. Accessed October 2008.

Dibben C, Sims A, Watson J, Barnes H, Smith T, Sigala M, Hill A, Manley D. *The Health Poverty Index.* South East Public Health Observatory: Oxford, 2004. http://www.hpi.org.uk/index.php Accessed January 2007.

Doll R, Hill A. The mortality of doctors in relation to their smoking habits: a preliminary report. *British Medical Journal,* 1954; 4877: 1451-1455.

Dong W, Ben-Shlomo Y, Colhoun H, Chaturvedi N. Gender differences in accessing cardiac surgery across England: a cross-sectional analysis of the Health Survey for England. *Social Science and Medicine,* 1998; 47: 1773-1780.

Drever F, Whitehead M. Health inequalities: setting the scene. In: Drever F, Whitehead M (eds). *Health inequalities.* Office for National Statistics: London, 1997.

Duncan C, Jones K, Moon G. Heath-related behaviour in context: a multilevel modelling approach. *Social Science & Medicine,* 1996; 42 (6): 817-830.

Edina. UKBORDERS website, UK historical administrative borders for download page http://borders.edina.ac.uk/ukborders/action/restricted/index, accessed November 2008.

Elford J, Phillips A, Thomson A, Shaper A. Migration and geographic variations in blood pressure in England. *British Medical Journal,* 1990; 300: 291-295.

Elliott P, Wakefield J. Bias and confounding in spatial epidemiology. In: Elliott P, Wakefield J, Best N, Briggs D (eds). *Spatial epidemiology.* Oxford University Press: Oxford, 2000.

Epstein F. Contribution of epidemiology to understanding coronary heart disease. In: Marmot M, Elliott P (eds). *Coronary heart disease epidemiology: from aetiology to public health.* 2nd edition. Oxford University Press: Oxford, 2005.

EURAREA Consortium. *Enhancing small area estimation techniques to meet European needs. Final project report.* Office for National Statistics: London, 2004.

Farmer J, Iversen L, Campbell N, Guest C, Chesson R, Deans G, MacDonald J. Rural / urban differences in accounts of patients' initial decisions to consult primary care. *Health & Place,* 2006; 12: 210-221.

Foster C, Hillsdon M, Thorogood M. Environmental perceptions and walking in English adults. *Journal of Epidemiology and Community Health,* 2004; 58: 924-928.

Garg R, Madans J, Kleinman J. Regional variation in ischemic heart disease incidence. *Journal of Clinical Epidemiology*, 1992; 45 (2): 149-156.

Gartner A, Farewell D, Dunstan F, Gordon E. Differences in mortality between rural and urban areas in England and Wales, 2002-2004. *Health Statistics Quarterly,* 2008; 39: 6-13.

Gatrell A, Lancaster G, Chapple A, Horsley S, Smith M. Variations in use of tertiary cardiac services in part of North-West England. *Health & Place,* 2002; 8: 147-153.

Gelman A, Shor B, Bafumi J, Park D. Rich state, poor state, red state, blue state: what's the matter with Connecticut? *Quarterly Journal of Political Science,* 2007; 2: 345-367.

Ghosh M, Rao J. Small area estimation: an appraisal (with discussion). *Statistical Science* 1994; 9: 55-93.

Godfrey R, Julien M. Urbanisation and health. *Clinical Medicine,* 2005; 5 (2): 137-141.

Griffiths C, Brock A. Twentieth century mortality trends in England and Wales. *Health Statistics Quarterly,* 2003; 18: 5-17.

Griffiths C, Brock A, Rooney C. The impact of introducing ICD-10 on trends in mortality from circulatory diseases in England and Wales. *Health Statistics Quarterly,* 2004; 22: 14-20.

Grimes D, Hindle E, Dyer T. Sunlight, cholesterol and coronary heart disease. *Quarterly Journal of Medicine,* 1996; 89: 579-589.

Heady P, Clarke P, Brown G, Ellis K, Heasman D, Hennell S, Longhurst J, Mitchell B. *Model-based small area estimation series no.2 Small area estimation project report.* Office for National Statistics: London, 2003.

Hippisley-Cox J, Pringle M. Inequalities in access to coronary angiography and revascularisation: the association of deprivation and location of primary care services. *British Journal of General Practice,* 2000; 50: 449-454.

Hospital Episode Statistics. *How good is the HES ethnicity coding and where do the problems lie? Ethnicity coding in HES: 1997-98 to 2002-03.* The Information Centre: Leeds, 2004.

Janzon E, Engstrom G, Hedblad B, Berglund G, Janzon L. Smoking as a determinant of the geographical pattern of cardiac events among women in an urban population. *Scandanavian Journal of Public Health,* 2007; 35: 272-277.

Joint Health Surveys Unit. *Health Survey for England 2006: Cardiovascular disease.* The Information Centre: Leeds, 2008.

Jones K, Duncan C. Individuals and their ecologies: analysing the geography of chronic illness within a multilevel modelling framework. *Health & Place,* 1995; 1 (1): 27-40.

Kannel W, Dawber T, Kagan A, Revotskie N, Stokes III J. Factors of risk in the development of coronary heart disease: six-year follow-up experience. *Annals of Internal Medicine,* 1961; 55: 33-50.

Keatinge W, Donaldson G, Bucher K, Jendritsky G, Cordioli E, Martinelli M, Dardanoni L, Katsouyanni K, Kunst A, Mackenbach J, McDonald C, Vuori I (The Eurowinter Group). Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. *The Lancet,* 1997; 349: 1341-1346.

Kuh D, Davey Smith G. The life course and adult chronic disease: an historical perspective with particular reference to coronary heart disease. In: Kuh D, Ben-Shlomo Y (eds). *A life course approach to chronic disease epidemiology.* 2nd edition. Oxford Medical Publications: Oxford, 2004.

Last J. *A dictionary of epidemiology. 4th edition.* Oxford University Press: Oxford, 2001.

Law M, Wald N. An ecological study of serum cholesterol and ischaemic heart disease between 1950 and 1990. *European Journal of Clinical Nutrition,* 1994; 48: 305-325.

Law M, Morris J. Why is mortality higher in poorer areas and in more northern areas of England and Wales? *Journal of Epidemiology and Community Health,* 1998; 52: 344-352.

Lawlor D, Bedford C, Taylor M, Ebrahim S. Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study. *Journal of Epidemiology and Community Health,* 2003; 57: 134-140.

Lawlor D, Ben-Shlomo Y, Leon D. Pre-adult influences on cardiovascular disease. In: Kuh D, Ben-Shlomo Y (eds). *A life course approach to chronic disease epidemiology.* 2nd edition. Oxford Medical Publications: Oxford, 2004.

Lawlor D, Davey Smith G, Patel R, Ebrahim S. Life-course socioeconomic position, area deprivation, and coronary heart disease: findings from the British Women's Heart and Health Study. *American Journal of Public Health,* 2005; 95 (1): 91-97.

Lewes F. William Farr and Cholera. *Population trends,* 1983; 31: 8-12.

Loftin C, Ward S. A spatial autocorrelation model of the effects of population density on fertility. *American Sociological Review,* 1983; 48(1): 121-128.

Lynch J, Kaplan G, Cohen R, Tuomilehto J, Salonen J. Do cardiovascular risk factors explain the relation between socio-economic status, risk of all-cause mortality,

cardiovascular mortality, and acute myocardial infarction? *American Journal of Epidemiology*, 1996; 144: 934-942.

Macintyre S, Maciver S, Sooman A. Area, class and health: should we be focusing on places or people? *Journal of Social Policy*, 1993; 22 (2): 213-234.

Macintyre S, Ellaway A, Cummins S. Place effects on health: how can we conceptualise, operationalise and measure them? *Social Science & Medicine,* 2002; 55: 125-139.

MacLeod M, Finlayson A, Pell J, Findlay I. Geographic, demographic, and socioeconomic variations in the investigation and management of coronary heart disease in Scotland. *Heart*, 1999; 81: 252-256.

Maheswaran R, Morris S, Falconer S, Grossinho A, Perry I, Wakefield J, Elliott P. Magnesium in drinking water supplies and mortality from acute myocardial infarction in north west England. *Heart,* 1999; 82: 455-460.

Maheswaran R, Haining R, Brindley P, Law J, Pearson T, Fryers P, Wise S, Campbell M. Outdoor air pollution, mortality, and hospital admissions from coronary heart disease in Sheffield, UK: a small-area level ecological study. *European Heart Journal*, 2005; 26: 2543-2549.

Marmot M. Socioeconomic determinants of CHD mortality. *International Journal of Epidemiology*, 1989; 18: 196-202.

Marx A, Neutra R. Magnesium in drinking water and ischemic heart disease. *Epidemiologic Reviews*, 1997; 19(2): 258-272.

McIsaac S, Wilkinson R. Income distribution and cause-specific mortality. *European Journal of Public Health*, 1997; 7: 45-53.

Mercer J. Cold – an underrated risk factor for health. *Environmental Research,* 2003; 92: 8-13.

MIMAS. *Casweb. 2001 census data for England and Wales.* University of Manchester: Manchester. http://casweb.mimas.ac.uk/. Accessed January 2006.

Monarca S, Zerbini I, Donato F. *Drinking water hardness and cardiovascular diseases: a review of epidemiological studies 1979-2004.* World Health Organization: Geneva, 2004.

Moon G, Quarendon G, Barnard S, Twigg L, Blyth B. Fat nation: deciphering the distinctive geographies of obesity in England. *Social Science & Medicine,* 2007; 65: 25-31.

Morgan O, Baker A. Measuring deprivation in England and Wales using 2001 Carstairs scores. *Health Statistics Quarterly,* 2006: 31; 28-33.

Morris R, Whincup P, Lampe F, Walker M, Wannamethee S, Shaper A. Geographic variation in incidence of coronary heart disease in Britain: the contribution of established risk factors. *Heart,* 2001; 86: 277-283.

Morris R, Whincup P, Emberson J, Lampe F, Walker M, Shaper A. North-South gradients in Britain for stroke and CHD. Are they explained by the same factors? *Stroke,* 2003; 34: 2604-2611.

Morris R, Walker M, Lennon L, Shaper A, Whincup P. Hard drinking water does not protect against cardiovascular disease: new evidence from the British Regional Heart Study. *European Journal of Cardiovascular Prevention and Rehabilitation,* 2008; 15: 185-189.

Multiple Risk Factor Intervention Trial (MRFIT) Research Group. Mortality rates after 10.5 years for participants in the Multiple Risk Factor Intervention Trial. *Journal of the American Medical Association,* 1990; 263: 1795-1801.

Mulvaney C, Coupland C, Wilson A, Hammersley V, Dyas J, Carlisle R. Does increased use of private health care reduce the demand for NHS care? A prospective survey of General Practice referrals. *Journal of Public Health,* 2005; 27(2): 182-188.

Neighbourhood statistics. *Combined air quality indicator 2001*. www.neighbourhood.statistics.gov.uk. Accessed 28 January 2007.

North East Public Health Observatory (NEPHO). *County Durham and Darlington Health and Lifestyle Survey 2002*. NEPHO: Newcastle-upon-Tyne, 2002.

Oakes J. Commentary: Individual, ecological and multilevel fallacies. *International Journal of Epidemiology*, 2009. doi:10.1093/ije/dyn356.

Office for National Statistics (ONS). *Area classification for statistical wards*. ONS: London, 2001.

Office for National Statistics (ONS). *National Statistics Postcode Directory 2001*. ONS: London, 2002. © Crown copyright 2002.

Office for National Statistics (ONS). *Living in Britain. Results from the General Household Survey 2004/05*. The Stationery Office: London, 2006a.

Office for National Statistics (ONS). *Expenditure and Food Survey 2004/05*. The Stationery Office: London, 2006b.

Office for National Statistics (ONS). *Moves within the UK*. .

www.statistics.gov.uk/cci/nugget.asp?id=1310 Accessed 1st November 2007a.

Office for National Statistics (ONS). *Beginner's guide to UK geography.* http://www.statistics.gov.uk/geography/census_geog.asp. Accessed September 2009.

Office of the Deputy Prime Minister (ODPM). *The English indices of deprivation 2004 (revised).* ODPM: London, 2004.

Petticrew M, McKee M, Jones J. Coronary artery surgery: are women discriminated against? *British Medical Journal,* 1993; 306: 1164-1166.

Phillips A, Davey Smith G. How independent are "independent" effects? Relative risk estimation when correlated exposures are measured imprecisely. *Journal of Clinical Epidemiology,* 1991; 44 (11): 1223-1231.

Pickering K, Scholes S, Bajekal S. *Synthetic estimation of healthy lifestyle indicators: stage 2 report.* London: Department of Health, 2004.

Pickering K, Scholes S, Bajekal M. *Synthetic estimation of healthy lifestyle indicators: stage 3 report.* London: Department of Health, 2005.

Pocock S, Shaper A, Cook D, Packham R, Lacey R, Powell P, Russell P. British Regional Heart Study: geographic variations in cardiovascular mortality, and the role of water quality. *British Medical Journal,* 1980; 280: 1243-1249.

Pocock S, Cook D, Shaper A. Analysing geographic variation in cardiovascular mortality: methods and results. *Journal of the Royal Statistical Society,* 1982; 145(3): 313-341.

Pope C, Burnett R, Thurston G, Thun M, Calle E, Krewski D, Godleski J. Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. *Circulation,* 2004; 109: 71-77.

Pope C. Air pollution. In: Marmot M, Elliott P (eds) *Coronary heart disease epidemiology. From aetiology to public health.* 2nd edition. Oxford Medical Publications: Oxford, 2005.

Purslow N, Fraser J, Dayan Y, Burnett J. *Active people survey: small area estimates.* Sport England: London, 2007.

Rasbash J, Browne J, Goldstein H. *MLwiN v2.02*. Centre for Multilevel Modelling, Institute of Education: London, 2003.

Romeri E, Baker A, Griffiths C. Mortality by deprivation and cause of death in England and Wales, 1999-2003. *Health Statistics Quarterly,* 2006; 32: 19-34.

Sandstrom T, Kelly F. Traffic-related air pollution, genetics and asthma development in children. *Thorax,* 2009; 64: 98-99.

Scarborough P, Allender S, Peto V, Rayner M. *Regional and social differences in coronary heart disease.* British Heart Foundation: London, 2008.

Schwartz G, Hanchette C. UV, latitude, and spatial trends in prostate cancer mortality: All sunlight is not the same (United States). *Cancer Causes Control,* 2006; 17: 1091-1101.

Shohaimi S, Luben R, Wareham N, Day N, Bingham S, Welch A, Oakes S, Khaw K-T. Residential area deprivation predicts smoking habit independently of individual educational level and occupational social class. A cross sectional study in the Norfolk cohort of the European Investigation into Cancer (EPIC-Norfolk). *Journal of Epidemiology and Community Health,* 2003; 57: 270-276.

Simmons D, Williams D, Powell M. The Coventry diabetes study: prevalence of diabetes and impaired glucose tolerance in Europids and Asians. *Quarterly Journal of Medicine,* 1991; 81 (296): 1021-1030.

Sloggett A, Joshi H. Higher mortality in deprived areas: community or personal disadvantage? *British Medical Journal,* 1994; 309: 1470-1474.

Snowdon H, Cording H, Al-Durrah F, Martin P. *South Tyneside Health and Lifestyle Survey 2003*. Northumbria University: Newcastle upon Tyne, 2004.

Stafford M, Bartley M, Sacker A, Marmot M, Wilkinson R, Boreham R, Thomas R. Measuring the social environment: social cohesion and material deprivation in English and Scottish neighbourhoods. *Environment & Planning A*, 2003; 35: 1459-1475.

Stafford M, Duke-Williams O, Shelton N. Small area inequalities in health: are we underestimating them? *Social Science & Medicine*, 2008; 67(6): 891-899.

Stamler J. Established major coronary risk factors: historical overview. In: Marmot M, Elliott P (eds). *Coronary heart disease epidemiology: from aetiology to public health.* 2nd edition. Oxford University Press: Oxford, 2005.

StataCorp. *Stata statistical software: release 10.* StataCorp LP: College station, Texas, 2007.

Subramanian S, Jones K, Kaddour A, Krieger N. Revisiting Robinson: the perils of individualistic and ecologic fallacy. *International Journal of Epidemiology*, 2009; 38: 342-360.

Toledano M, Shaddick G, Elliott P. Seasonal variations in all-cause and cardiovascular mortality and the role of temperature. In: Marmot M, Elliott P (eds) *Coronary heart disease epidemiology. From aetiology to public health.* 2nd edition. Oxford Medical Publications: Oxford, 2005.

Townsend P, Davidson N, Whitehead M. *The Black Report and the health divide.* Penguin: Harmondsworth, 1986.

Twigg L, Moon G, Jones K. Predicting small area health-related behaviour: a comparison of smoking and drinking indicators. *Social Science & Medicine,* 2000; 50: 1109-1120.

Twigg L, Moon G. Predicting small area health-related behaviour: a comparison of multilevel synthetic estimation and local survey data. *Social Science & Medicine,* 2002; 54: 931-937.

Twigg L, Moon G, Walker S. *The smoking epidemic in England.* Health Development Agency: London, 2004.

Tyler D, Stephens C, Blackman T, Bellingham M. *South Tyneside Health and Lifestyles Survey.* South Tyneside PCT: Newcastle upon Tyne, 1995.

Vittinghoff E, Glidden D, Shiboski S, McCulloch C. *Regression methods in biostatistics.* Springer: New York, 2005.

Walker M, Whincup P, Shaper A. The British Regional Heart Study 1975-2004. *International Journal of Epidemiology,* 2004; 33: 1185-1192.

Wannamethee S, Shaper A, Whincup P, Walker M. Migration within Great Britain and cardiovascular disease: early life and adult environmental factors. *International Journal of Epidemiology,* 2002; 31: 1054-1060.

Ward P, Noyce P, St Leger A. Are GP practice prescribing rates for coronary heart disease drugs equitable? A cross sectional analysis in four primary care trusts in England. *Journal of Epidemiology and Community Health,* 2004; 58: 89-96.

Wennemo I. Infant mortality, public policy and inequality: a comparison of 18 industrialised countries 1950-85. *Sociology of Health and Illness,* 1993; 15: 429-446.

West Midlands Public Health Observatory (WMPHO). *European Standard Population.* http://www.wmpho.org.uk/localprofiles/metadata.aspx?id=META_EUROSTD Accessed January 2009.

West Midlands Public Health Observatory (WMPHO). *West Midlands Regional Lifestyle Survey 2005.* WMPHO: Birmingham, 2006.

West R, Lowe C. Mortality from ischaemic heart disease – inter-town variation and its association with climate in England and Wales. *International Journal of Epidemiology*, 1976; 5 (2): 195-201.

White M, Bunting J, Raybould S, Adamson A, Williams L, Mathers J. *Do 'food deserts' exist? A multi-level, geographical analysis of the relationship between retail food access, socio-economic position and dietary intake.* Food Standards Agency: London, 2004.

World Health Organization. *World Health Report 2002: Reducing risk, promoting healthy life.* World Health Organization: Geneva, 2002.

Yorkshire and Humberside Public Health Observatory (YHPHO). *PBS diabetes population prevalence model – phase 2.* YHPHO: Hull, 2005.

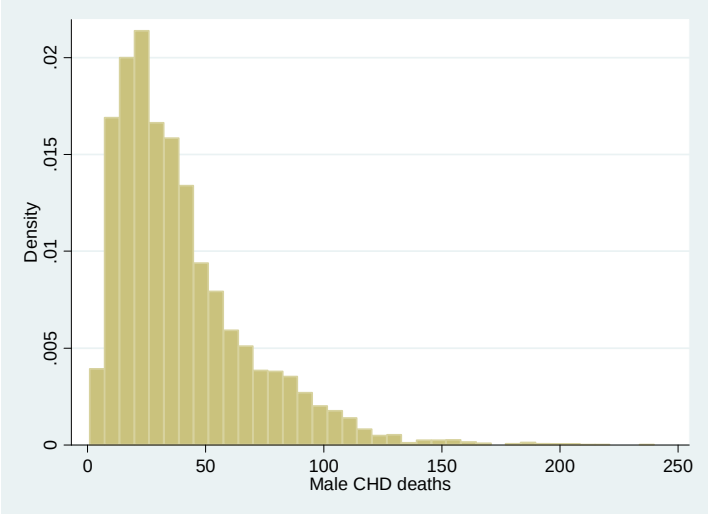
Yusuf S, Hawken S, Ounpuu S, Dans T, Avezum A, Lanas F, McQueen M, Budaj A, Pais P, Varigos J, Lisheng L, on behalf of the INTERHEART study investigators. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *The Lancet*, 2004; 364: 937-952.

Appendix 1: Distribution of outcome variables, evidence of spatial autocorrelation, details of principal components analysis

---

Figure 4.1 Distribution, mean and variance of number of CHD events in England 2001-06 (wards, n = 7,929)

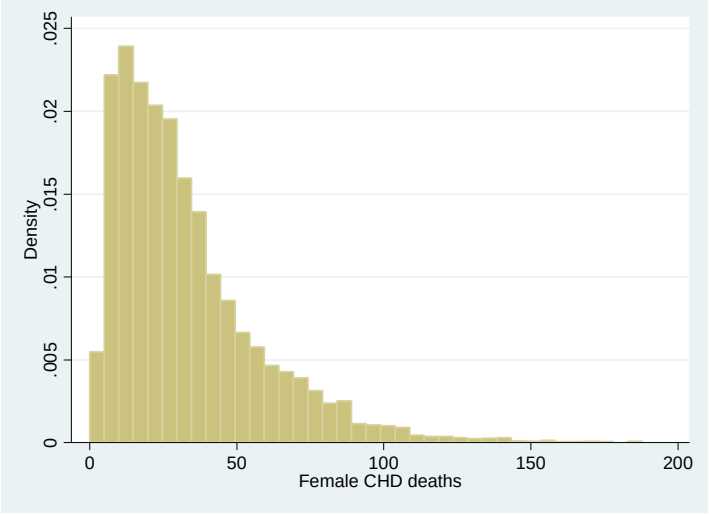
Mortalities, men



Mean:
39.3

Variance:
712.0

Mortalities, women

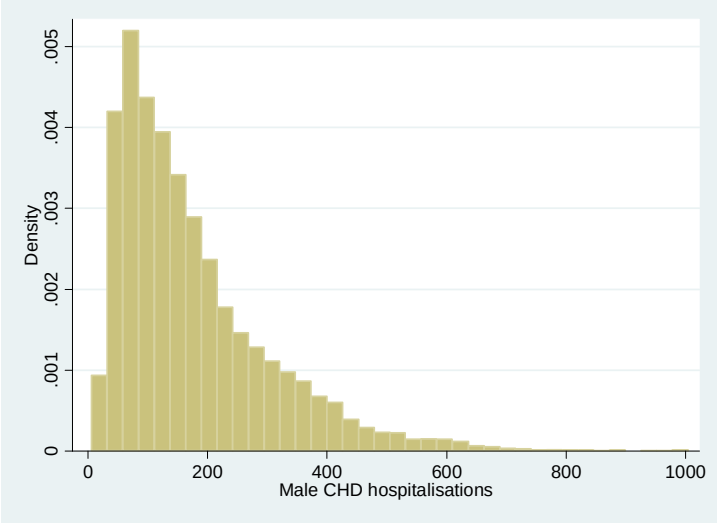


Mean:
32.8

Variance:
636.2

Figure 4.1 (cont.)

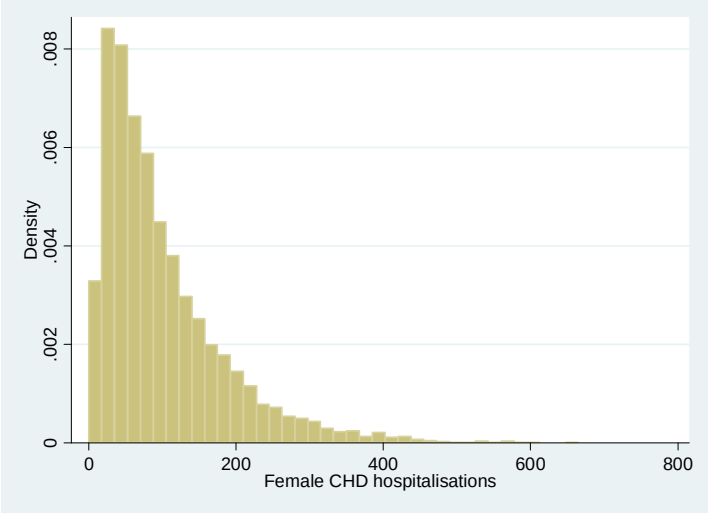
Hospitalisations, men



Mean:
176.3

Variance:
16,806.7

Hospitalisations, women

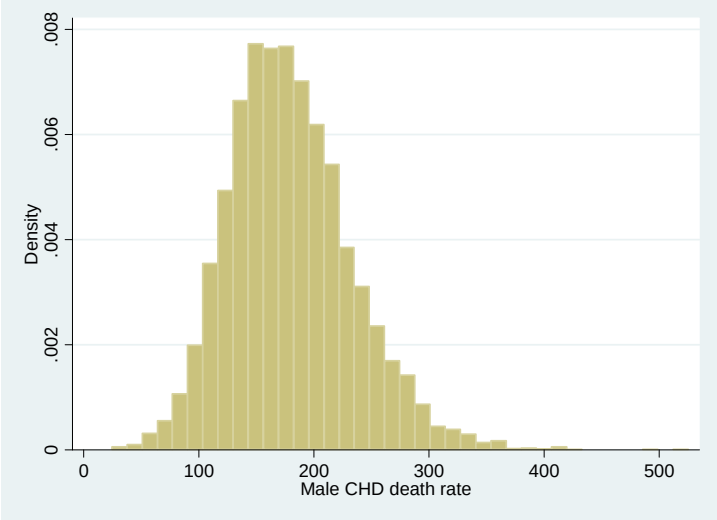


Mean:
100.4

Variance:
7,024.0

Figure 4.2 Distribution, mean, variance, skew and kurtosis of age-standardised CHD event rate in England, 2001-06 (wards, n = 7,929)

Mortalities, men



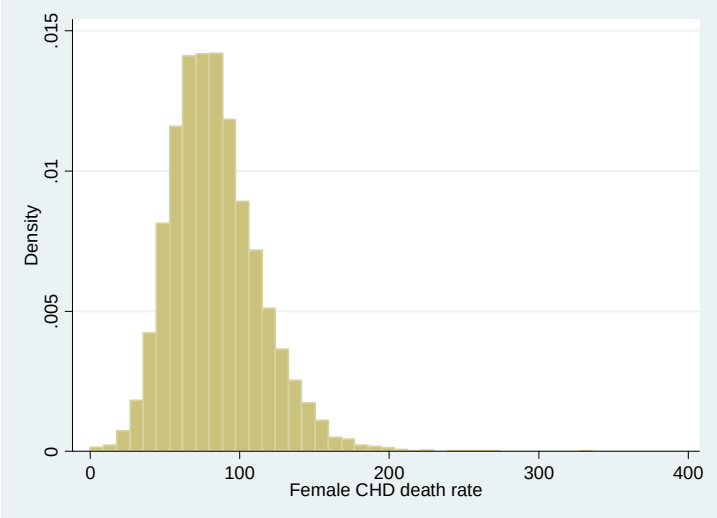
Mean:
179.9

Variance:
2,873.6

Skew:
0.6

Kurtosis:
3.9

Mortalities, women



Mean:
83.6

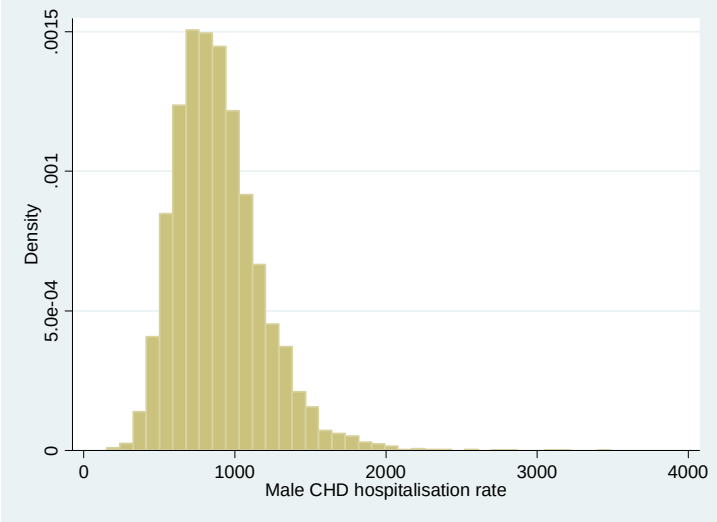
Variance:
884.9

Skew:
0.8

Kurtosis:
4.9

Figure 4.2 (cont.)

Hospitalisations, men



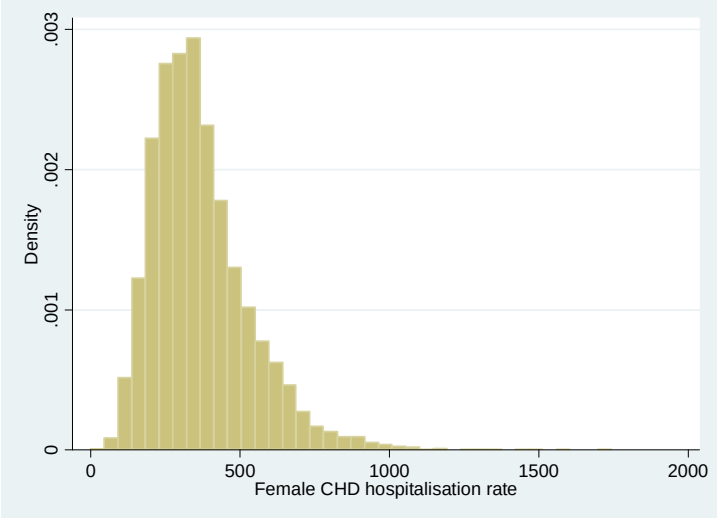
Mean:
892.1

Variance:
84,936.3

Skew:
1.1

Kurtosis:
6.5

Hospitalisations, women



Mean:
368.4

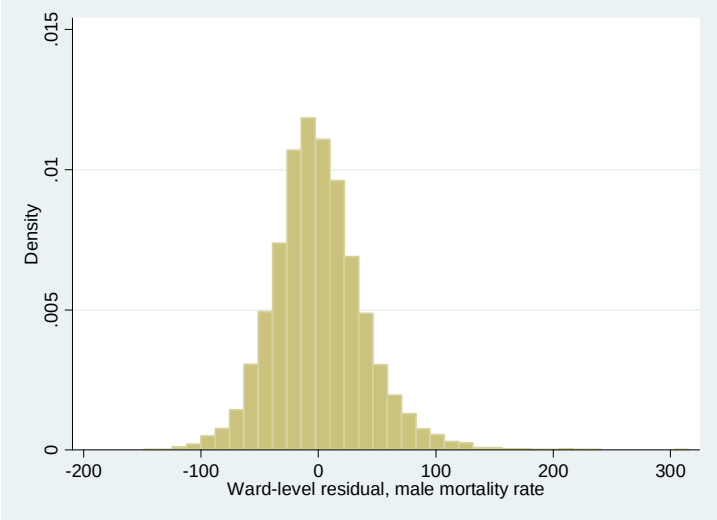
Variance:
27,164.2

Skew:
1.3

Kurtosis:
6.7

Figure 4.3 Distribution, mean, variance, skew and kurtosis of ward-level residuals of final multi-level models (table 8.3, chapter eight) (wards, n = 7,929)

Mortalities, men



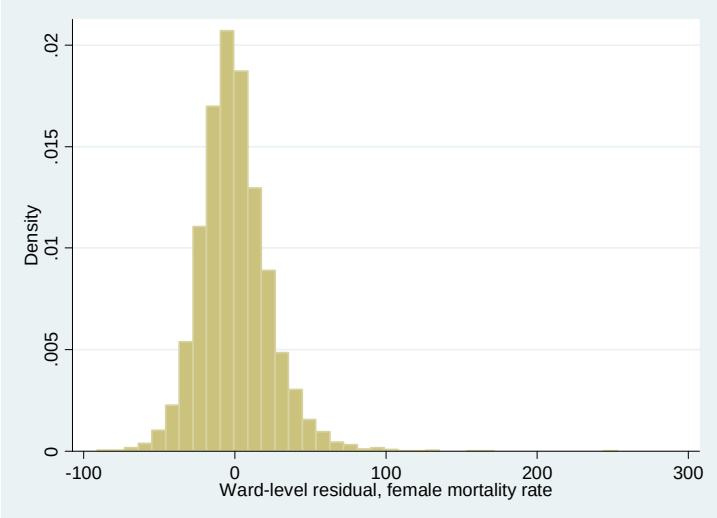
Mean:
-0.0

Variance:
1,509.5

Skew:
0.6

Kurtosis:
5.2

Mortalities, women



Mean:
0.0

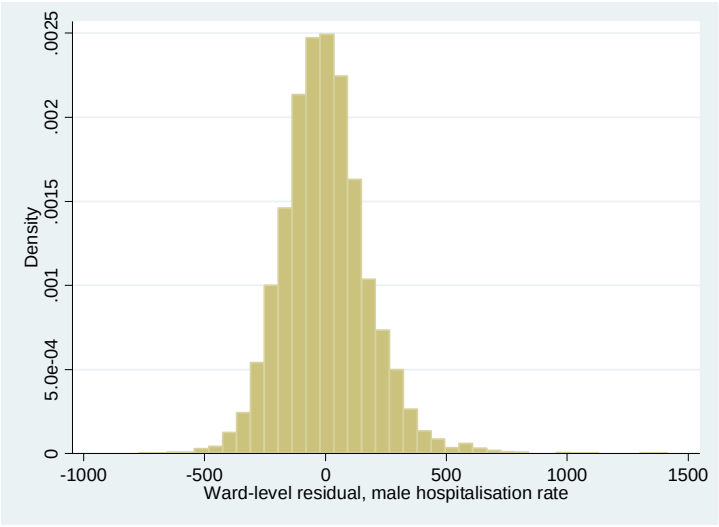
Variance:
523.0

Skew:
0.8

Kurtosis:
7.1

Figure 4.3 (cont.)

Hospitalisations, men



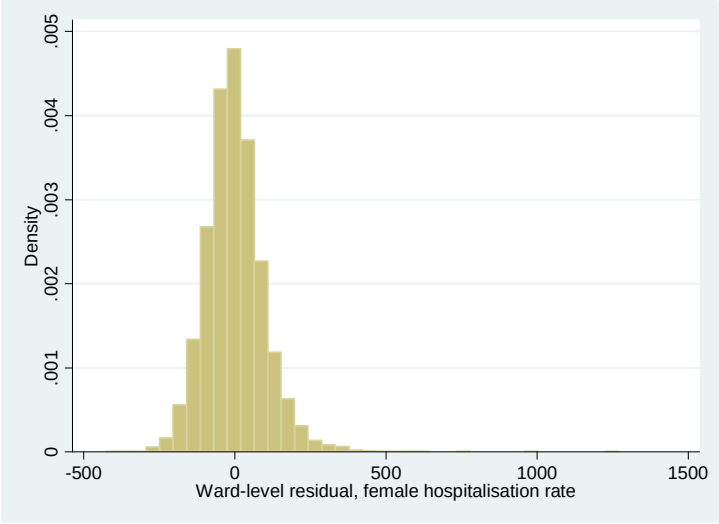
Mean:
-0.0

Variance:
31,115.6

Skew:
0.6

Kurtosis:
5.6

Hospitalisations, women



Mean:
-0.0

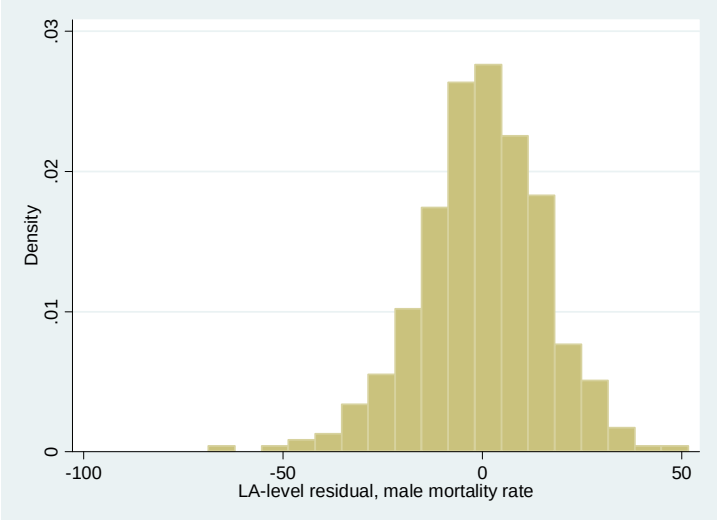
Variance:
9,487.0

Skew:
1.0

Kurtosis:
10.1

Figure 4.4 Distribution, mean, variance, skew and kurtosis of local authority-level residuals of final multi-level models (table 8.3, chapter eight) (local authorities, n = 354)

Mortalities, men



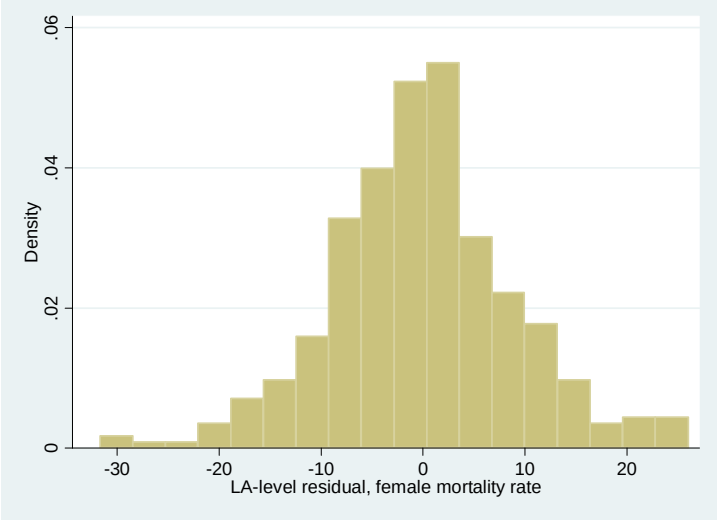
Mean:
-0.1

Variance:
247.8

Skew:
-0.3

Kurtosis:
4.1

Mortalities, women



Mean:
-0.1

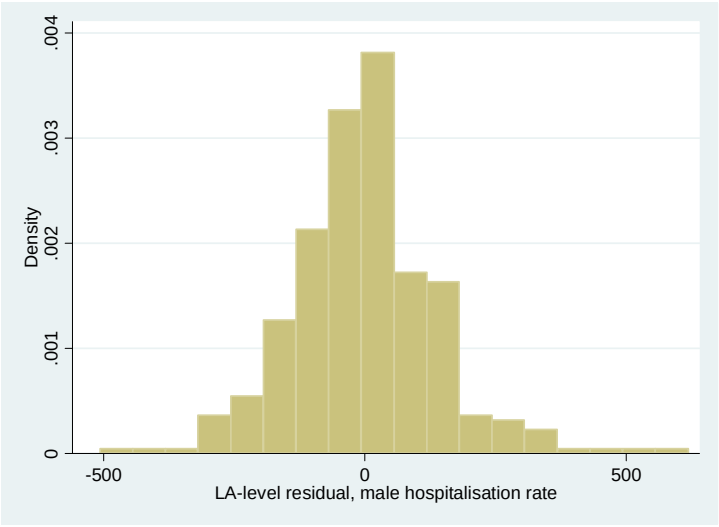
Variance:
82.4

Skew:
-0.0

Kurtosis:
3.8

Figure 4.4 (cont.)

Hospitalisations, men



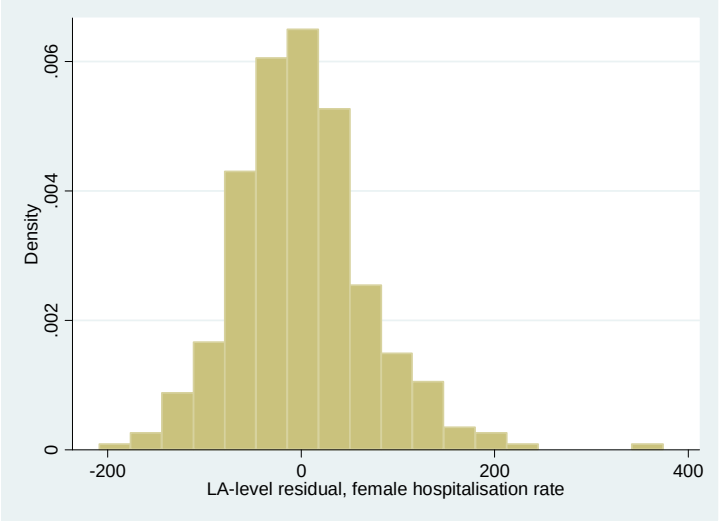
Mean:
0.2

Variance:
19,164.4

Skew:
0.4

Kurtosis:
5.2

Hospitalisations, women



Mean:
0.0

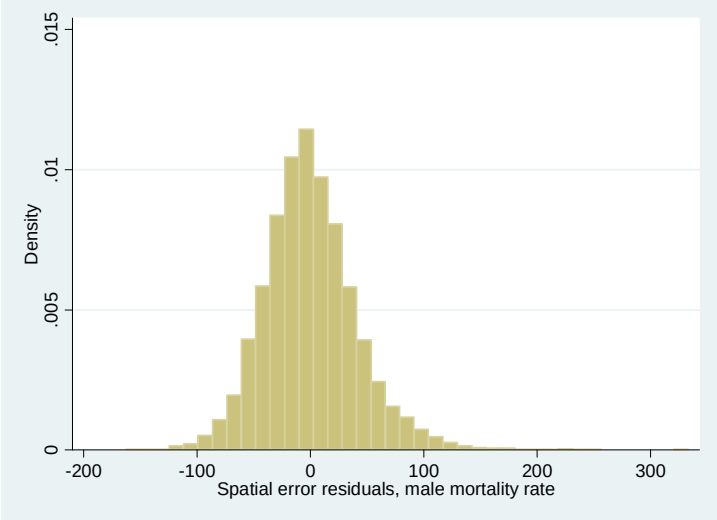
Variance:
4,875.4

Skew:
0.7

Kurtosis:
5.4

Figure 4.5 Distribution, mean, variance, skew and kurtosis of residuals of final spatial error models (table 8.4, chapter eight) (wards, n = 7,929)

Mortalities, men



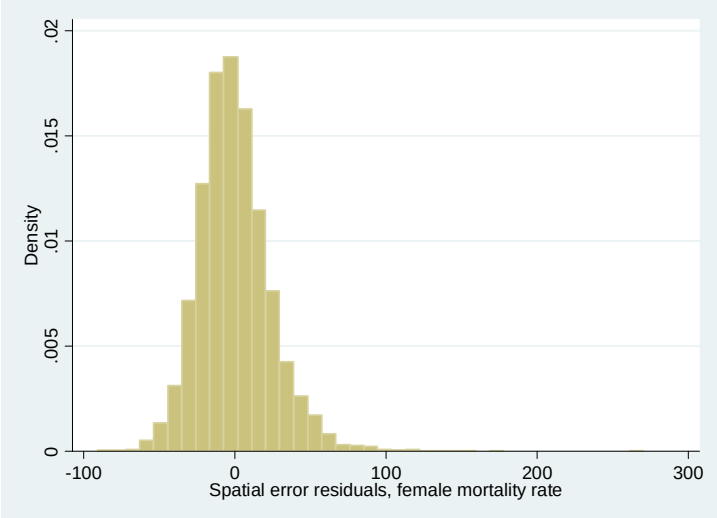
Mean:
0.0

Variance:
1,683.2

Skew:
0.7

Kurtosis:
5.2

Mortalities, women



Mean:
0.0

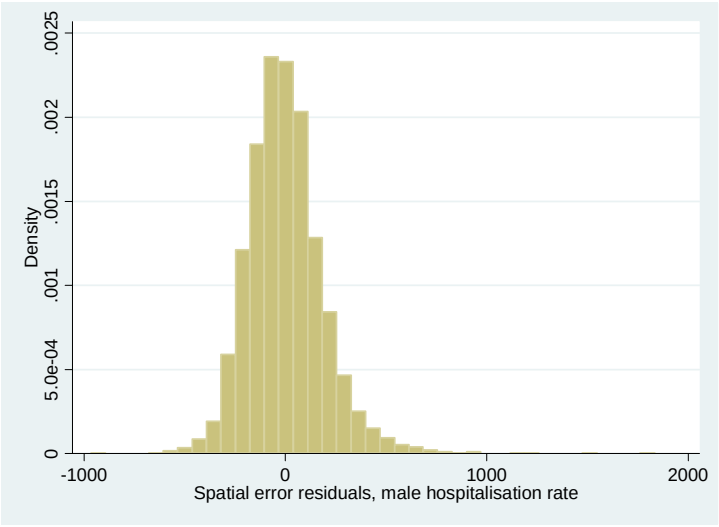
Variance:
579.5

Skew:
0.9

Kurtosis:
7.2

Figure 4.5 (cont.)

Hospitalisations, men



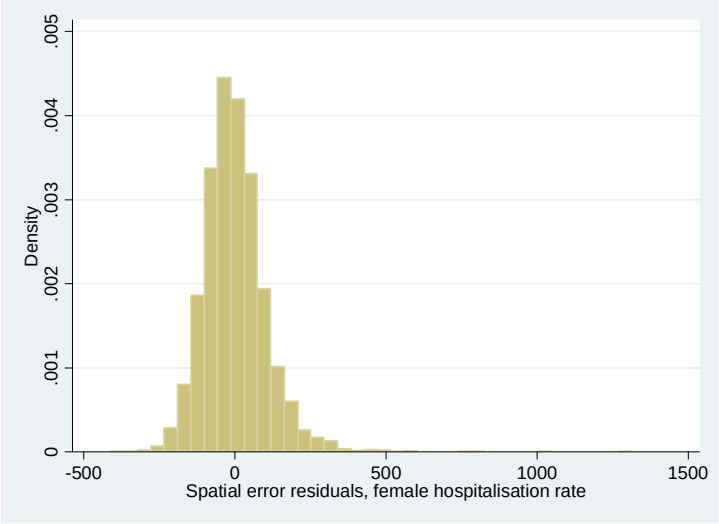
Mean:
-0.1

Variance:
34,766.1

Skew:
0.8

Kurtosis:
6.5

Hospitalisations, women



Mean:
0.1

Variance:
10,791.0

Skew:
1.2

Kurtosis:
10.5

Table 4.1 Global Moran's I (spatial autocorrelation) measure for all outcome and explanatory variables used in this thesis (wards, n = 7,929)

	Global Moran's I	p
Outcome variables		
CHD mortality rate, men	0.33	<0.001
CHD mortality rate, women	0.31	<0.001
CHD hospitalisation rate, men	0.51	<0.001
CHD hospitalisation rate, women	0.53	<0.001
Environmental variables		
Mean daily maximum temperature	0.99	<0.001
Mean daily minimum temperature	0.97	<0.001
Total annual sunshine	0.99	<0.001
Total annual rainfall	0.97	<0.001
Air quality index	0.95	<0.001
Behavioural risk factor profiles of populations		
Low fruit and vegetable consumption, men	0.53	<0.001
Low fruit and vegetable consumption, women	0.53	<0.001
Obesity, men	0.79	<0.001
Obesity, women	0.79	<0.001
Raised blood pressure, men	0.83	<0.001
Raised blood pressure, women	0.83	<0.001
Raised cholesterol, men	0.59	<0.001
Raised cholesterol, women	0.57	<0.001
Smoking, men	0.47	<0.001
Smoking, women	0.48	<0.001
PCA: Unhealthy lifestyle 1, men	0.74	<0.001
PCA: Unhealthy lifestyle 1, women	0.74	<0.001
PCA: Unhealthy lifestyle 2, men	0.84	<0.001
PCA: Unhealthy lifestyle 2, women	0.82	<0.001
Deprivation		
Carstairs deprivation index	0.62	<0.001

Table 4.2 Transformation matrices calculated by principal components analysis for the sets of synthetic estimates for (1) male prevalence of behavioural risk factors for CHD and (2) female prevalence of behavioural risk factors for CHD, and amount of original variance explained by the transformed variables (wards, n = 7,929).

(1) Synthetic estimates of male prevalence rates

	Fruit & Veg	Obesity	Blood pressure	Cholesterol	Smoking	*Proportion of original variance*
PCA 1	0.51	0.54	0.51	0.17	0.41	*0.63*
PCA 2	0.14	-0.08	-0.06	0.92	-0.36	*0.21*
PCA 3	0.17	-0.38	-0.45	0.22	0.76	*0.11*
PCA 4	-0.75	-0.10	0.50	0.26	0.33	*0.06*
PCA 5	-0.37	0.74	-0.54	0.13	0.10	*0.00*

(2) Synthetic estimates of female prevalence rates

	Fruit & Veg	Obesity	Blood pressure	Cholesterol	Smoking	*Proportion of original variance*
PCA 1	0.51	0.51	0.48	0.22	0.45	*0.66*
PCA 2	0.05	-0.31	-0.28	0.89	0.17	*0.19*
PCA 3	-0.26	0.27	0.46	0.39	-0.70	*0.11*
PCA 4	-0.70	-0.18	0.47	0.02	0.50	*0.04*
PCA 5	-0.42	0.73	-0.51	0.09	0.15	*0.00*

PCA1 to PCA5 are the transformed variables.

Face validity

Bias

Figure 6.1a-n Scatter plot of synthetic estimates versus Health Survey for England (HSfE) 2000-02 estimates of the prevalence of behavioural risk factors for CHD (local authorities, n ~ 355)

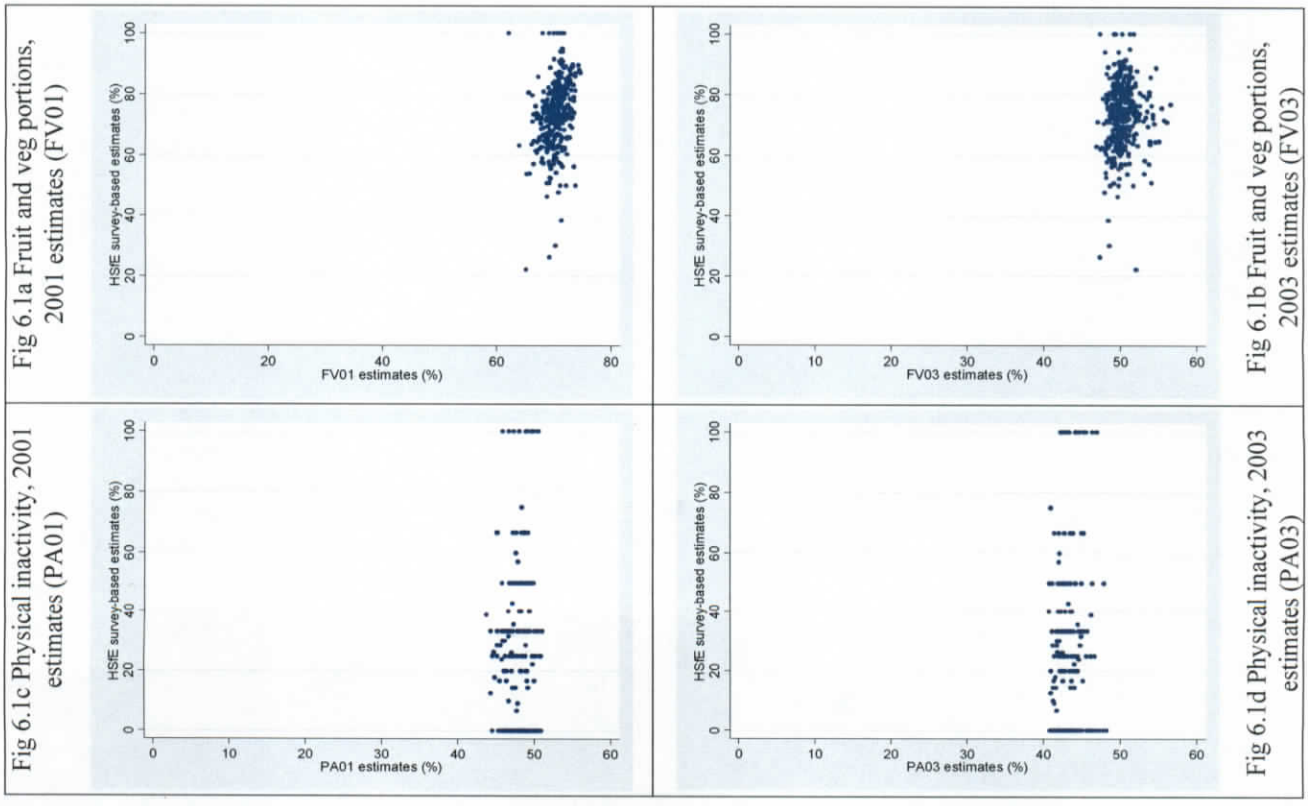


Figure 6.1 (cont.)

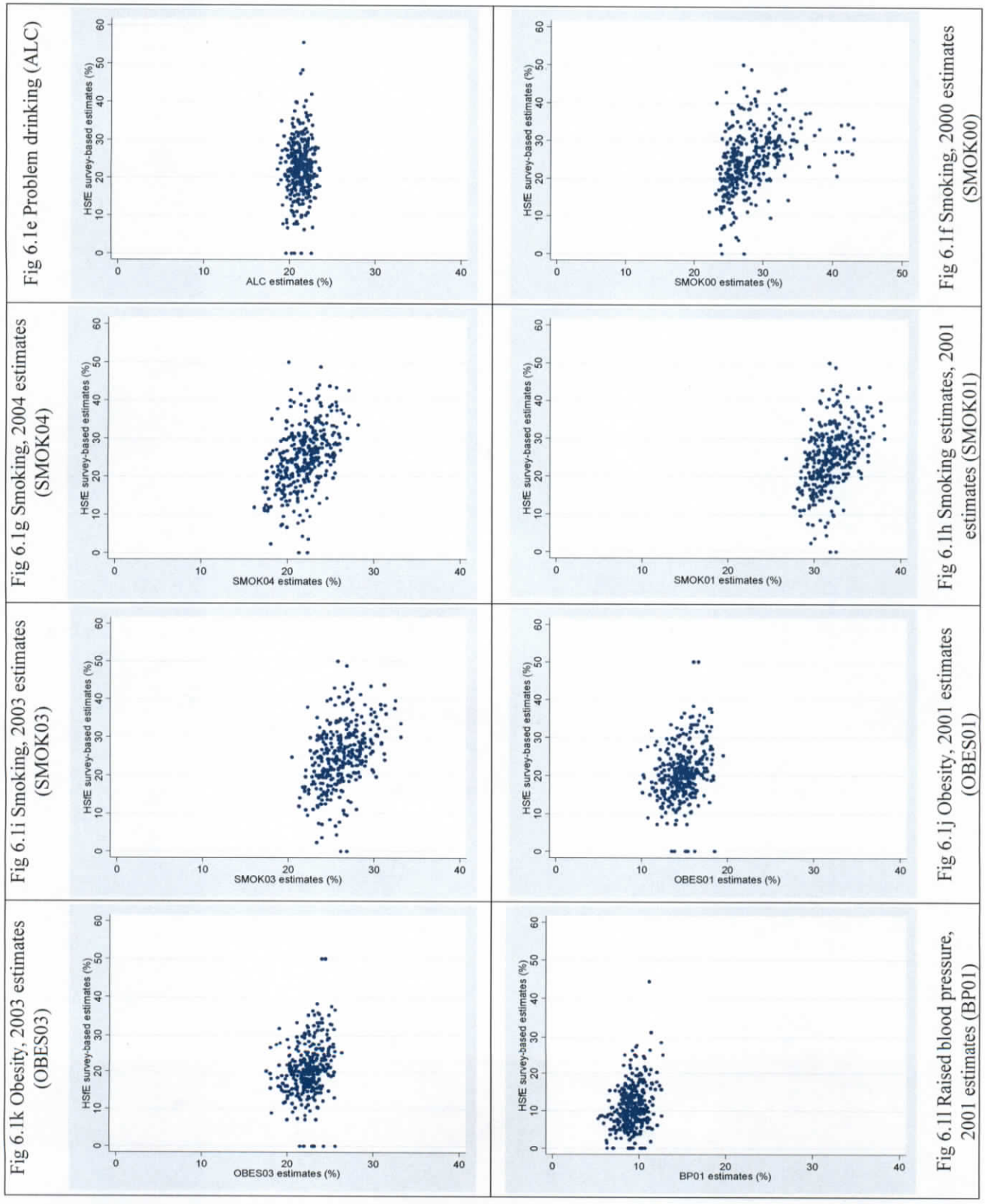


Fig 6.1e Problem drinking (ALC)

Fig 6.1f Smoking, 2000 estimates (SMOK00)

Fig 6.1g Smoking, 2004 estimates (SMOK04)

Fig 6.1h Smoking estimates, 2001 estimates (SMOK01)

Fig 6.1i Smoking, 2003 estimates (SMOK03)

Fig 6.1j Obesity, 2001 estimates (OBES01)

Fig 6.1k Obesity, 2003 estimates (OBES03)

Fig 6.1l Raised blood pressure, 2001 estimates (BP01)

Figure 6.1 (cont.)

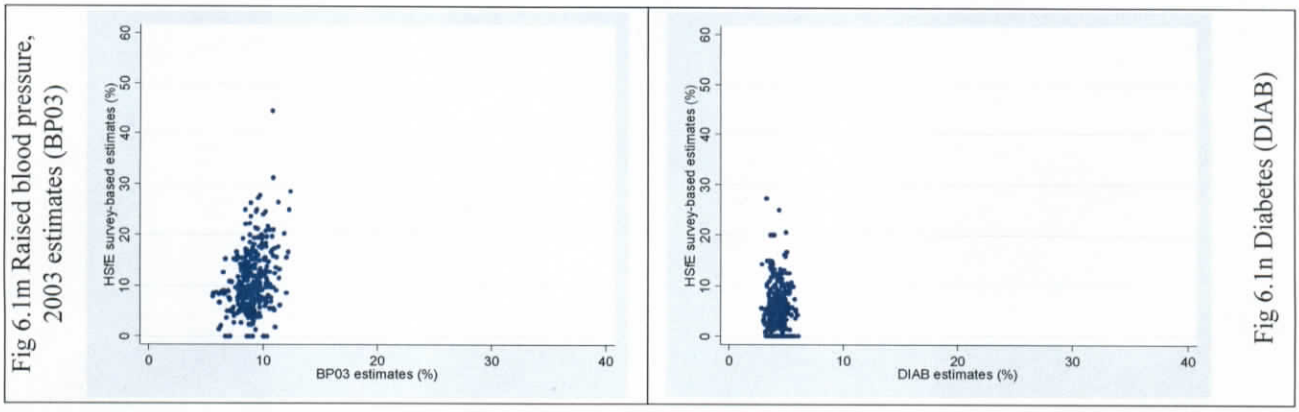


Fig 6.1m Raised blood pressure, 2003 estimates (BP03)

Fig 6.1n Diabetes (DIAB)

Heteroskedasticity

Figures 6.2a-n Scatter plot of synthetic estimates versus residuals (local authorities, n ~ 355)

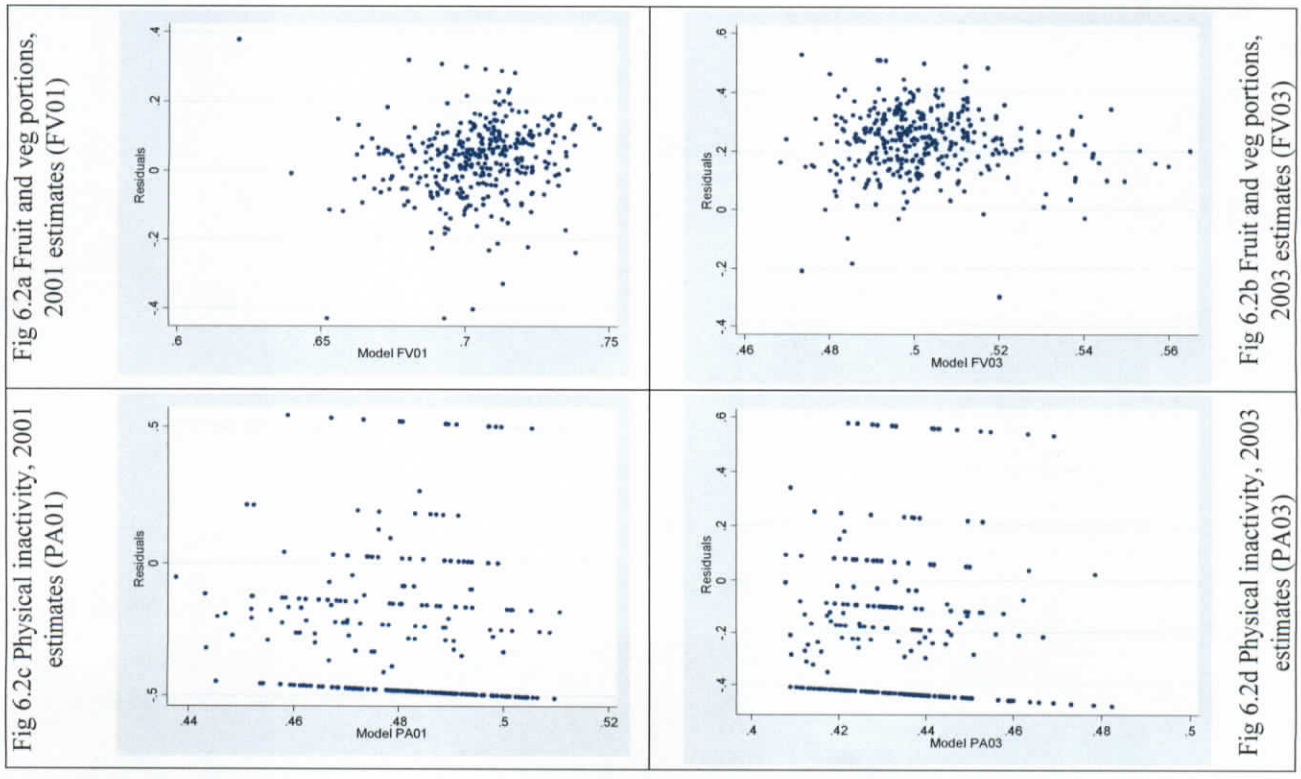


Fig 6.2a Fruit and veg portions, 2001 estimates (FV01)

Fig 6.2b Fruit and veg portions, 2003 estimates (FV03)

Fig 6.2c Physical inactivity, 2001 estimates (PA01)

Fig 6.2d Physical inactivity, 2003 estimates (PA03)

Figure 6.2 (cont.)

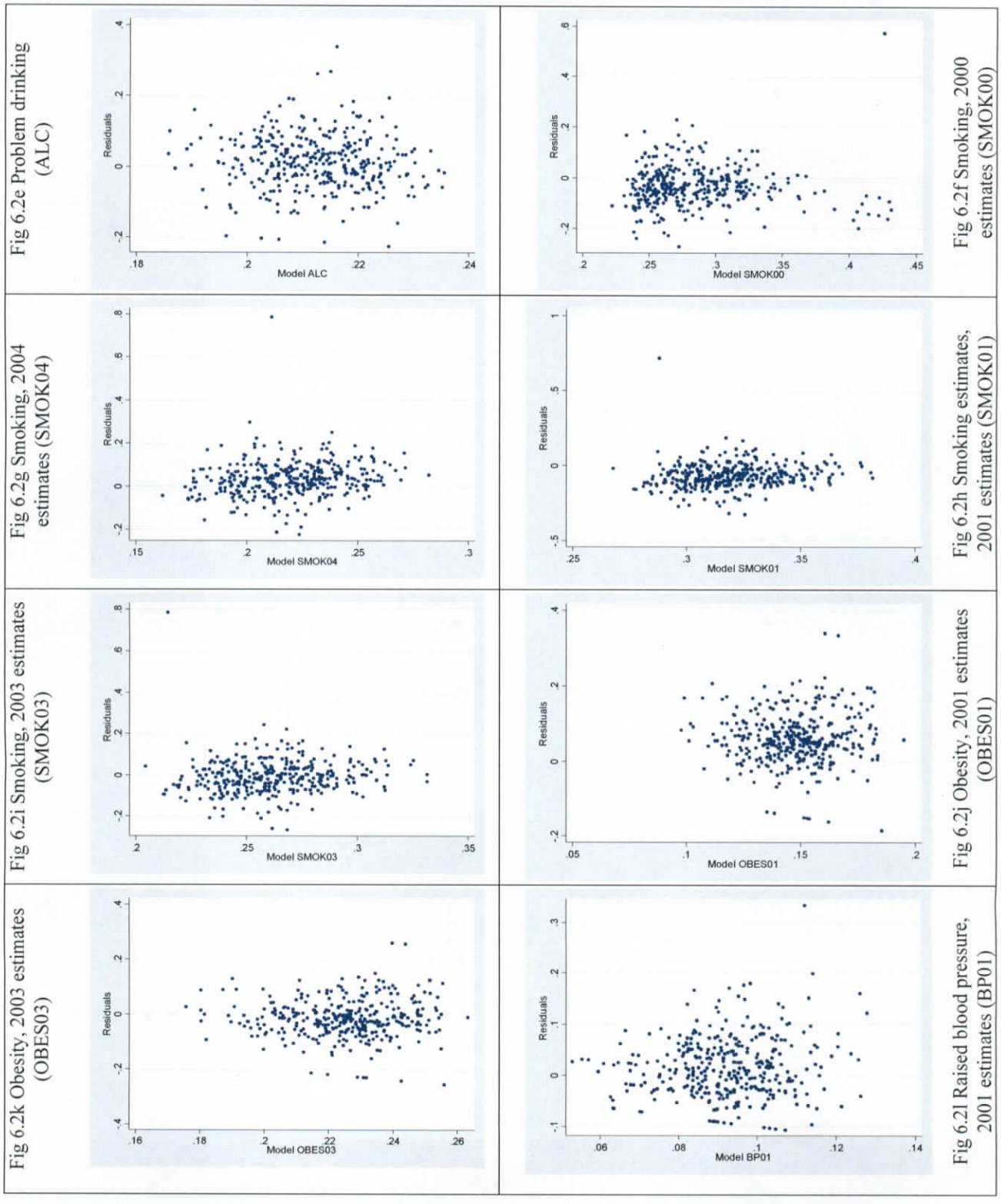


Fig 6.2e Problem drinking (ALC)

Fig 6.2f Smoking, 2000 estimates (SMOK00)

Fig 6.2g Smoking, 2004 estimates (SMOK04)

Fig 6.2h Smoking estimates, 2001 estimates (SMOK01)

Fig 6.2i Smoking, 2003 estimates (SMOK03)

Fig 6.2j Obesity, 2001 estimates (OBES01)

Fig 6.2k Obesity, 2003 estimates (OBES03)

Fig 6.2l Raised blood pressure, 2001 estimates (BP01)

Figure 6.2 (cont.)

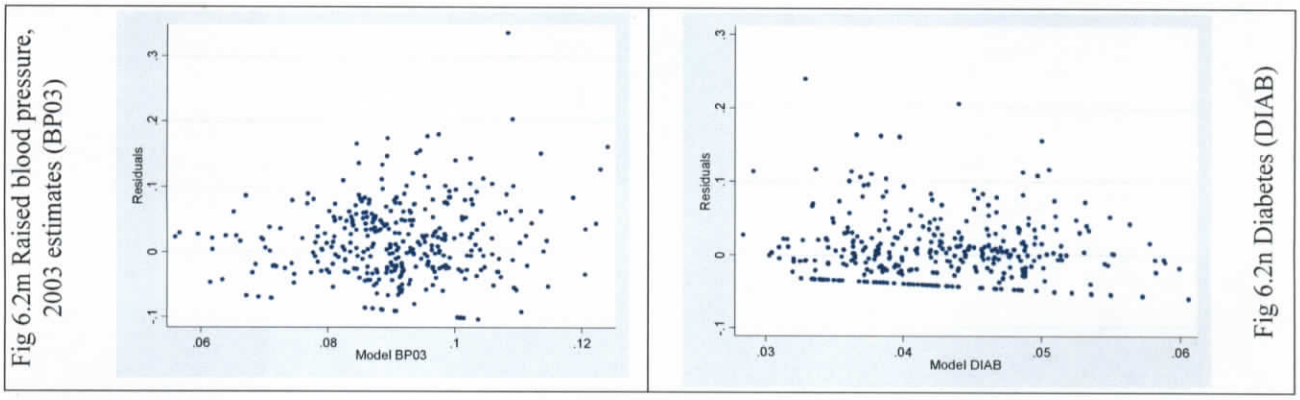


Fig 6.2m Raised blood pressure, 2003 estimates (BP03)

Fig 6.2n Diabetes (DIAB)

Spatial clustering of residuals

Figure 6.3a-n Maps of spatial clustering of synthetic estimate residuals (quartiles of residuals: white = largest underestimate of survey estimates; black = largest overestimate of survey estimates) (local authorities, n ~ 355)

Fig 6.3a Fruit and veg portions, 2001 estimates (FV01)

Fig 6.3b Fruit and veg portions, 2003 estimates (FV03)

Fig 6.3c Physical inactivity, 2001 estimates (PA01)

Fig 6.3d Physical inactivity, 2003 estimates (PA03)

Figure 6.3 (cont.)

Fig 6.3k Obesity, 2003 estimates (OBES03)

Fig 6.3i Smoking, 2003 estimates (SMOK03)

Fig 6.3g Smoking, 2004 estimates (SMOK04)

Fig 6.3e Problem drinking (ALC)

Fig 6.3l Raised blood pressure, 2001 estimates (BP01)

Fig 6.3j Obesity, 2001 estimates (OBES01)

Fig 6.3h Smoking estimates, 2001 estimates (SMOK01)

Fig 6.3f Smoking, 2000 estimates (SMOK00)

Figure 6.3 (cont.)

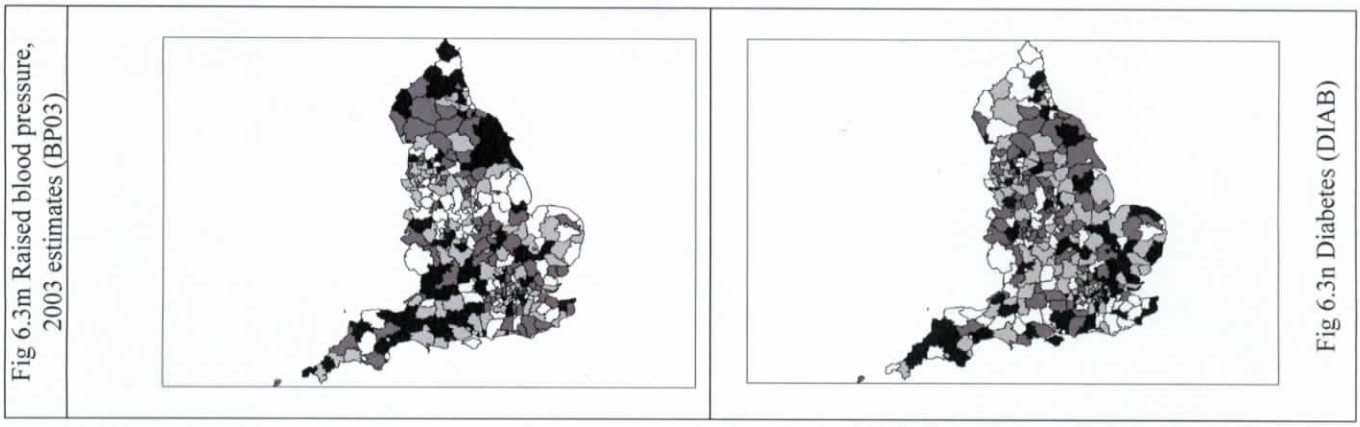


Fig 6.3m Raised blood pressure, 2003 estimates (BP03)

Fig 6.3n Diabetes (DIAB)

Convergent validity

Figure 6.4a-k Scatter plot of West Midlands Regional Lifestyle Survey (WMRLS) 2005 prevalence estimates for risk factors against synthetic estimates (local authorities, n = 34)

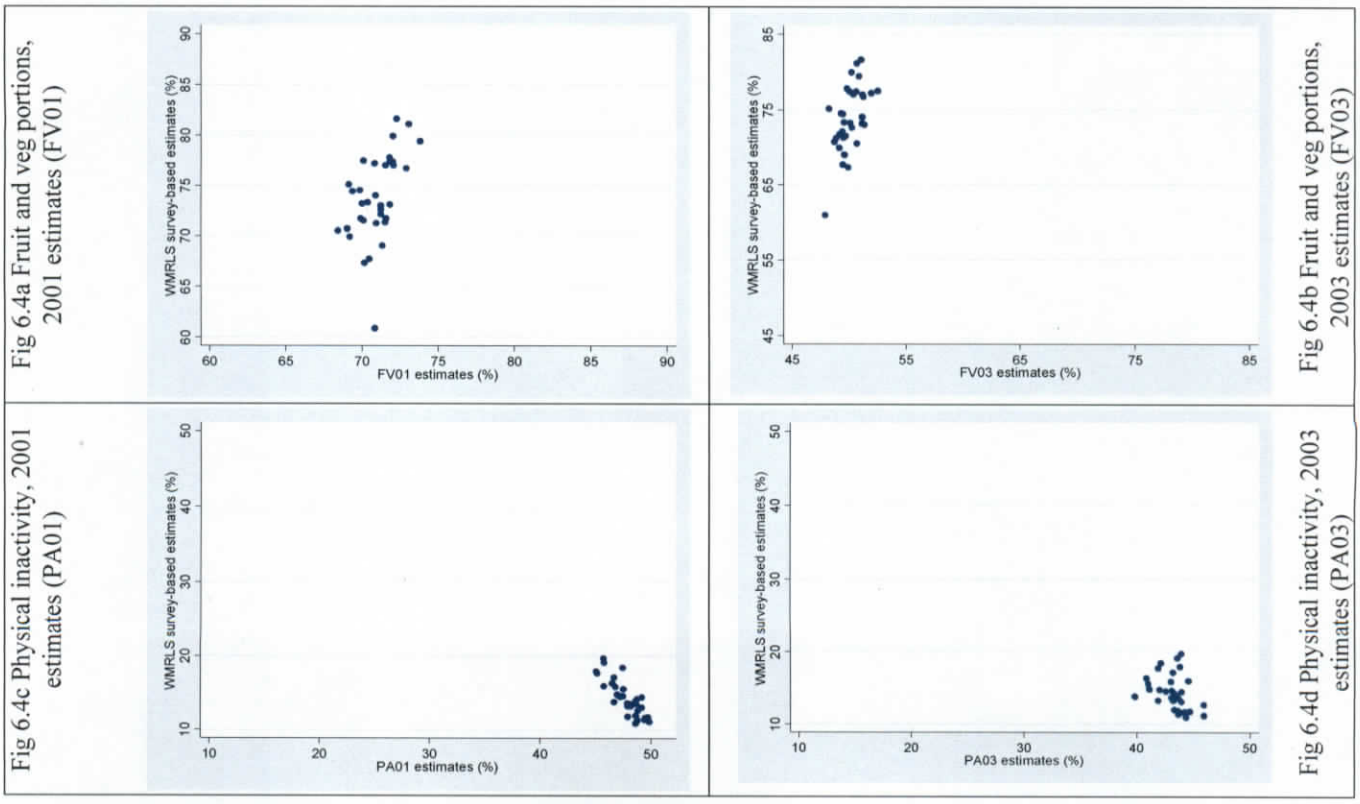


Fig 6.4a Fruit and veg portions, 2001 estimates (FV01)

Fig 6.4b Fruit and veg portions, 2003 estimates (FV03)

Fig 6.4c Physical inactivity, 2001 estimates (PA01)

Fig 6.4d Physical inactivity, 2003 estimates (PA03)

Figure 6.4 (cont.)

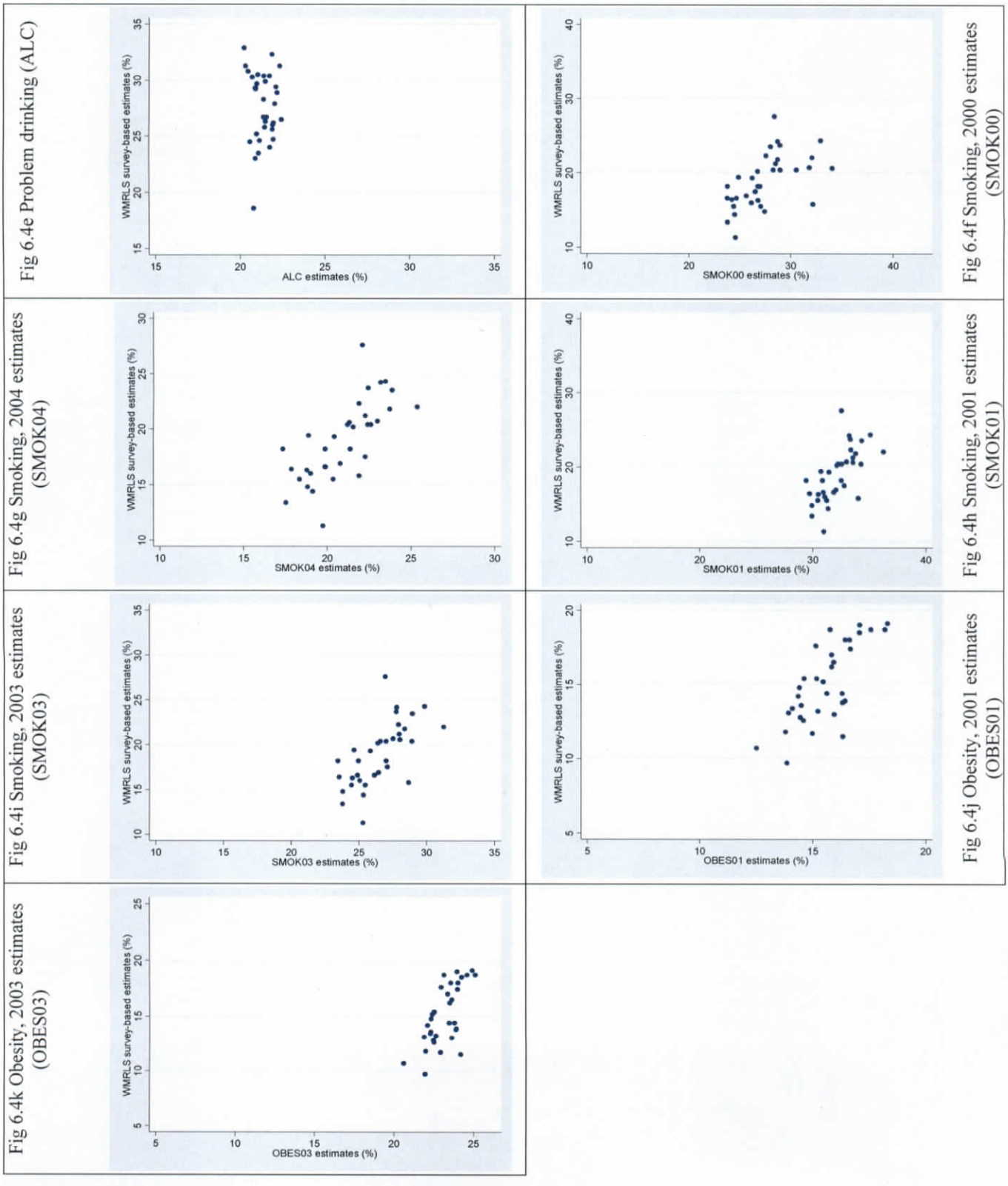


Figure 6.5a-n Scatter plot of Durham and Darlington Health and Lifestyle Survey 2002 (DDHLS) prevalence estimates for risk factors against synthetic estimates (wards, n = 116)

Fig 6.5a Fruit and veg portions, 2001 estimates (FV01)

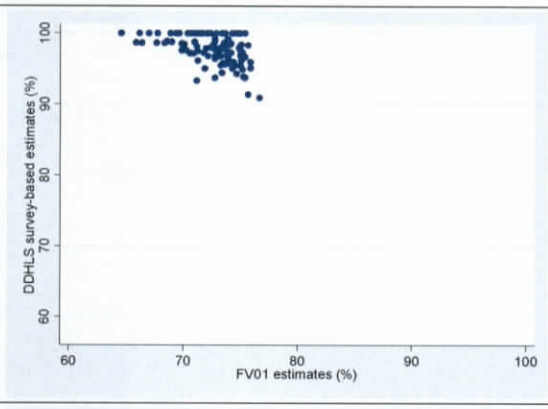


Fig 6.5b Fruit and veg portions, 2003 estimates (FV03)

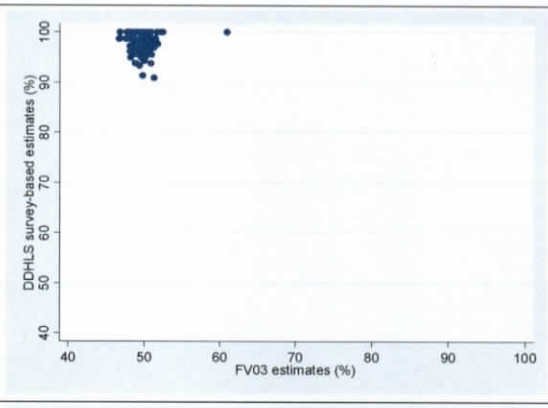


Fig 6.5c Physical inactivity, 2001 estimates (PA01)

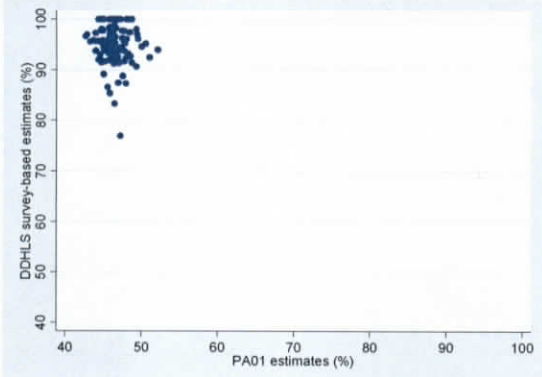


Fig 6.5d Physical inactivity, 2003 estimates (PA03)

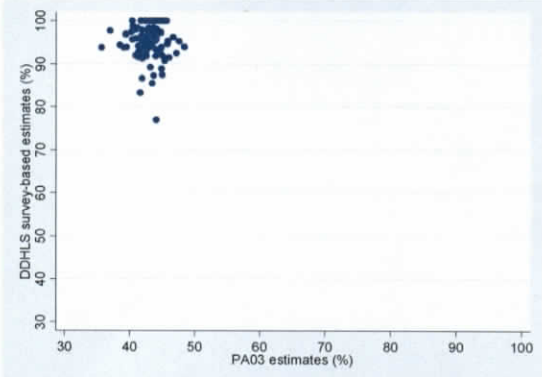


Figure 6.4e Model ALC

Figure 6.4f Model SMOK00

Fig 6.5e Problem drinking (ALC)

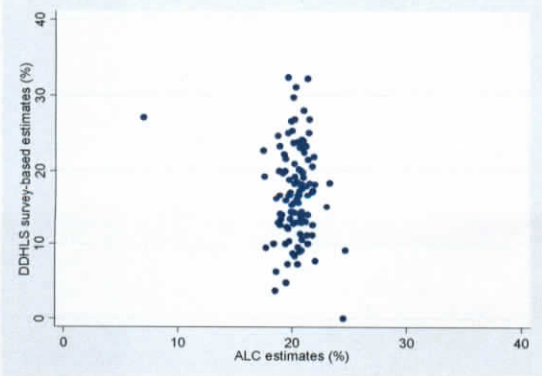


Fig 6.5f Smoking, 2000 estimates (SMOK00)

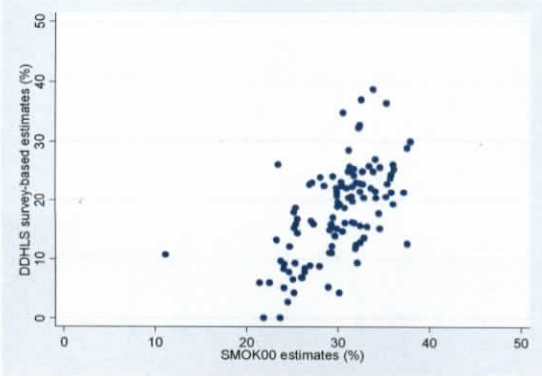


Figure 6.5 (cont.)

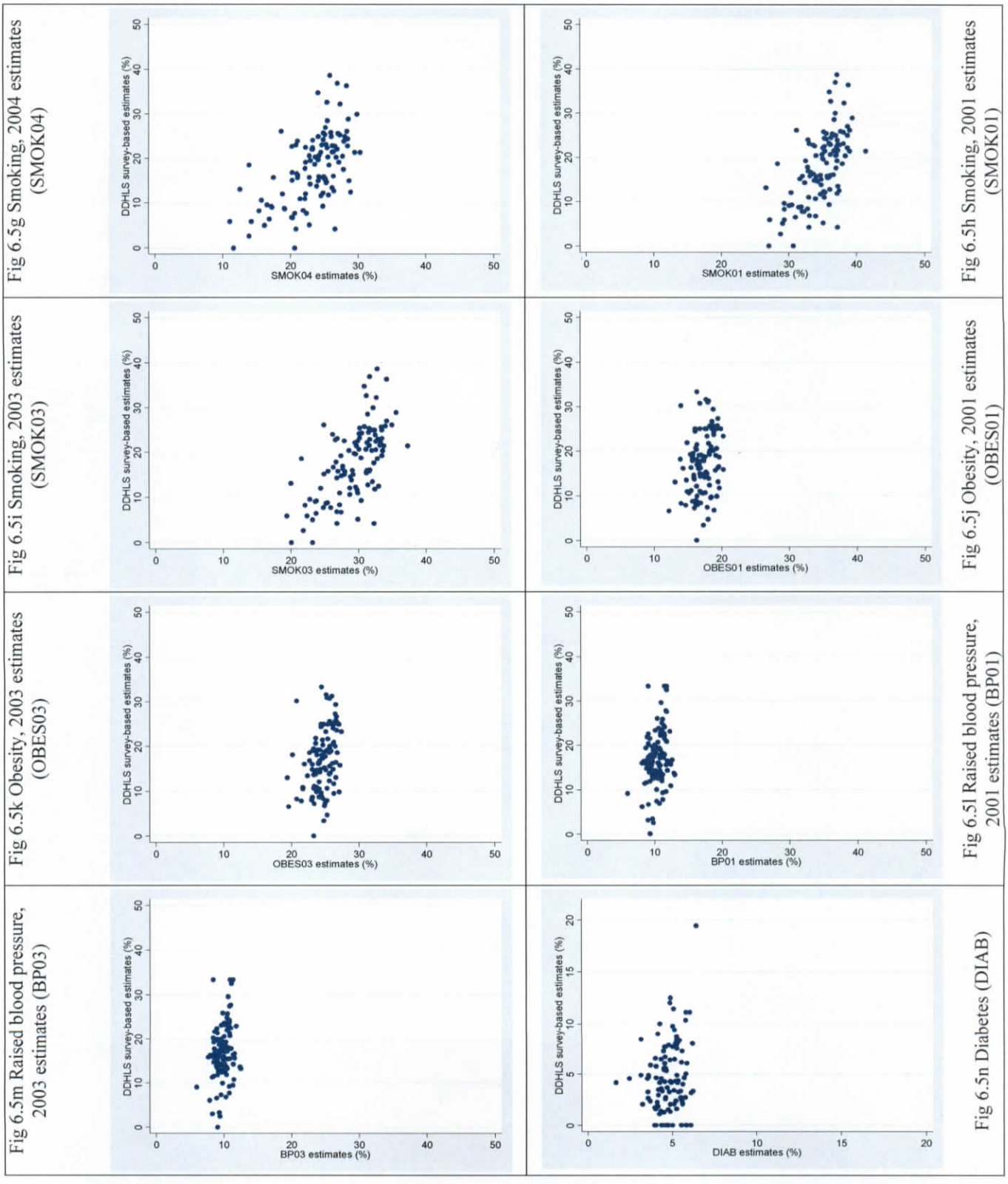


Fig 6.5g Smoking, 2004 estimates (SMOK04)

Fig 6.5h Smoking, 2001 estimates (SMOK01)

Fig 6.5i Smoking, 2003 estimates (SMOK03)

Fig 6.5j Obesity, 2001 estimates (OBES01)

Fig 6.5k Obesity, 2003 estimates (OBES03)

Fig 6.5l Raised blood pressure, 2001 estimates (BP01)

Fig 6.5m Raised blood pressure, 2003 estimates (BP03)

Fig 6.5n Diabetes (DIAB)

Construct validity

Table 6.11 Health Survey for England variables used to derive local authority and government office region level prevalence rates of behavioural risk factors

Smoking:	
Survey years:	1999-2003
Variable name:	cigst1
Variable definition:	Cigarette smoking status
Category indicating uptake of behaviour:	Current cigarette smoker
Individuals excluded if response is:	No answer / refused; item not applicable; missing data
Diet:	
Survey years:	2001-2003
Variable name:	porfv
Variable definition:	Total portions of fruit and vegetables consumed
Category indicating uptake of behaviour:	<5
Individuals excluded if response is:	Item not applicable; Missing data
Physical activity:	
Survey years:	1999; 2002-2003
Variable name:	adt30gp
Variable definition:	Summary activity level
Category indicating uptake of behaviour:	Group 1 – low physical activity level (less than 30 minutes activity per day)
Individuals excluded if response is:	No answer / refused; Don't know; item not applicable; missing data
Alcohol:	
Survey years:	1999-2002
Variable name:	overlim
Variable definition:	Drinking in relation to weekly limits
Category indicating uptake of behaviour:	Over weekly limits: male, 21 units; female, 14 units
Individuals excluded if response is:	No answer / refused; Don't know; Item not applicable; Missing data

Table 6.11 (cont.)

Blood pressure:	
Survey years:	1999-2002
Variable name:	diaval; sysval
Variable definition:	Valid mean diastolic blood pressure; valid mean systolic blood pressure
Category indicating uptake of behaviour:	diaval>=95 OR sysval>=160
Individuals excluded if response is:	Don't know; Schedule not applicable; Item not applicable; Missing data
Survey years:	2003
Variable name:	didiaval; disysval
Variable definition:	Dinamap valid mean diastolic blood pressure; dinamap valid mean systolic blood pressure
Category indicating uptake of behaviour:	didiaval>=95 OR dissysval>=160
Individuals excluded if response is:	Schedule not applicable; item not applicable; missing data
Blood cholesterol:	
Survey years:	1999; 2003
Variable name:	cholval
Variable definition:	Valid cholesterol result
Category indicating uptake of behaviour:	cholval>=6.5
Individuals excluded if response is:	Item not applicable; Missing data
Obesity:	
Survey years:	1999-2003
Variable name:	bmivg4
Variable definition:	Valid BMI (grouped: <20, 20-25, 25-30, 30+)
Category indicating uptake of behaviour:	Over 30
Individuals excluded if response is:	Item not applicable; Missing data
Diabetes:	
Survey years:	1999-2003
Variable name:	illsm1; illsm2; illsm3; illsm4; illsm5; illsm6
Variable definition:	Type of illness
Category indicating uptake of behaviour:	Diabetes
Individuals excluded if response is:	Item not available; missing data

Predictive validity

Figure 6.6a-af Gender stratified scatter graphs of age-standardised CHD mortality rates for 2001-06 versus synthetic estimates (wards, n= 7,929)

Fig 6.6a Fruit and veg portions, 2001 estimates (FVO1), men

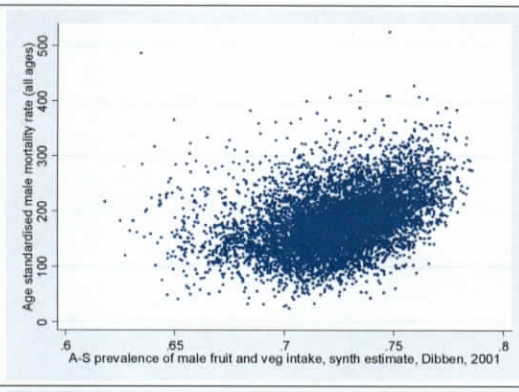


Fig 6.6b Fruit and veg portions, 2001 estimates (FV01), women

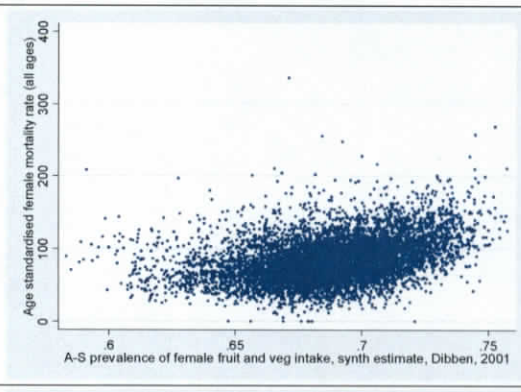


Fig 6.6c Fruit and veg portions, 2003 estimates (FV03), men

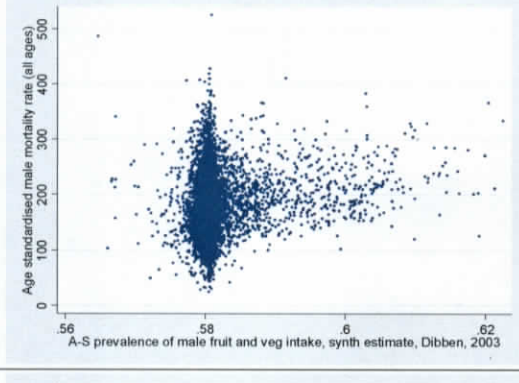


Fig 6.6d Fruit and veg portions, 2003 estimates (FV03), women

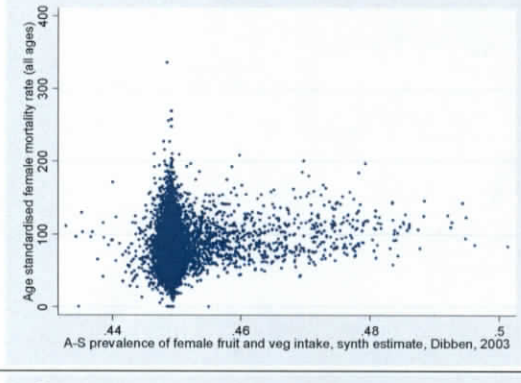


Fig 6.6e Physical inactivity, 2001 estimates (PA01), men

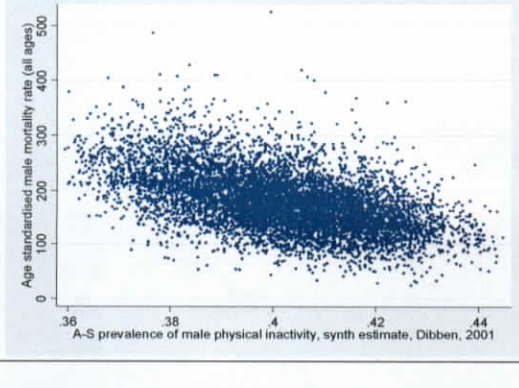


Fig 6.6f Physical inactivity, 2001 estimates (PA01), women

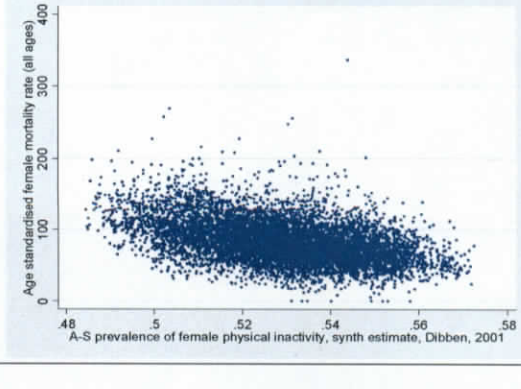


Figure 6.6 (cont.)

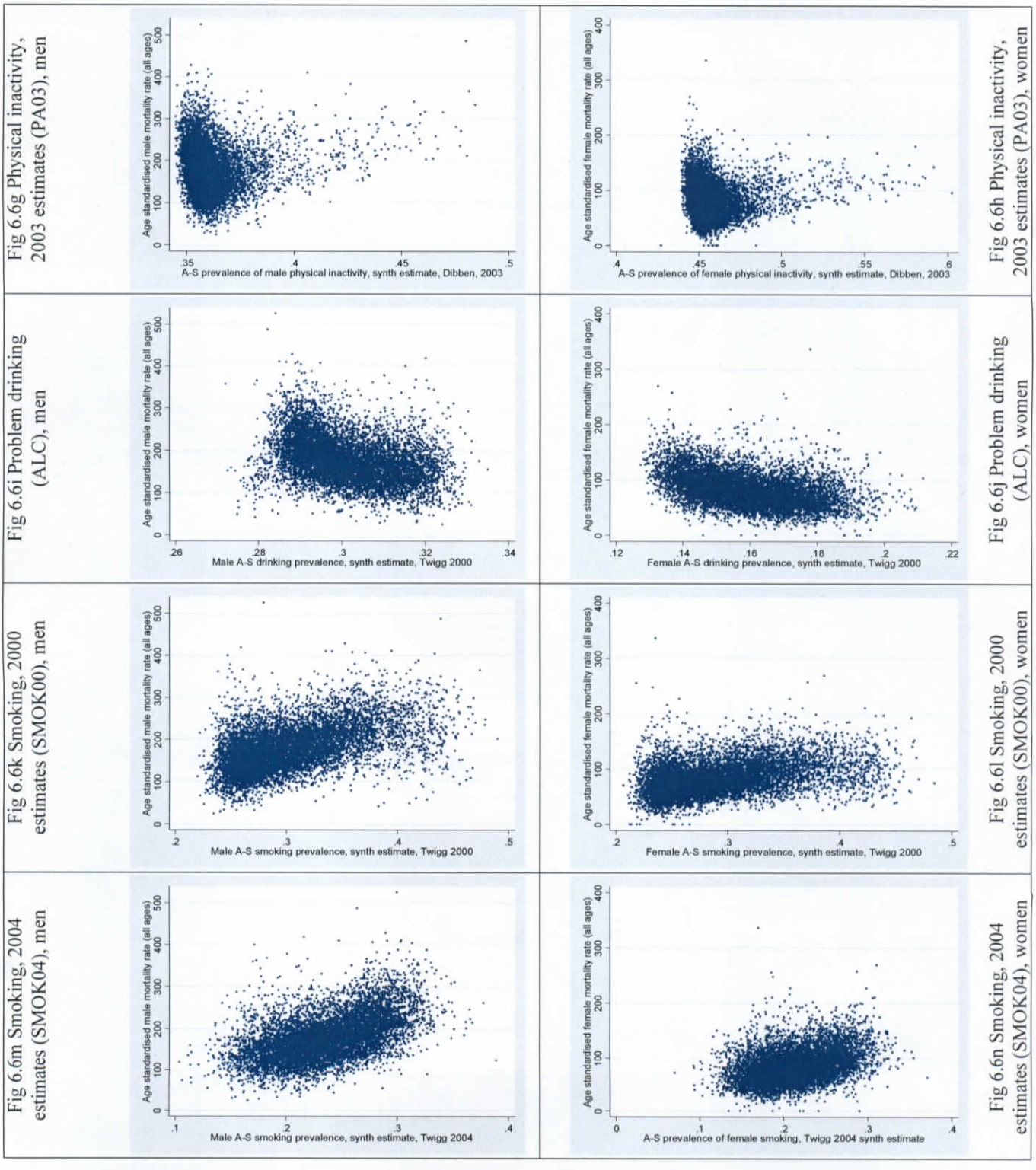


Figure 6.6 (cont.)

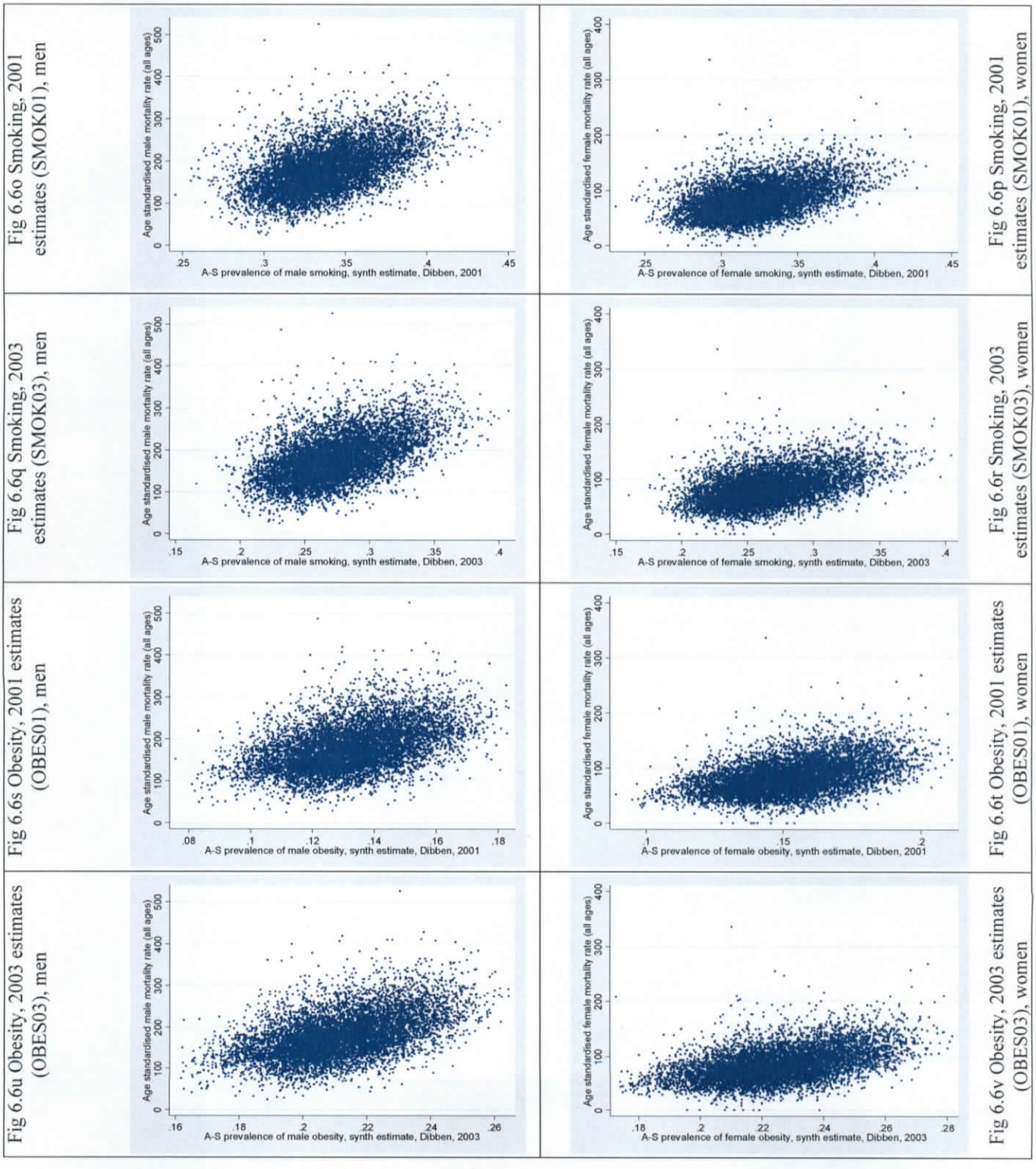


Figure 6.6 (cont.)

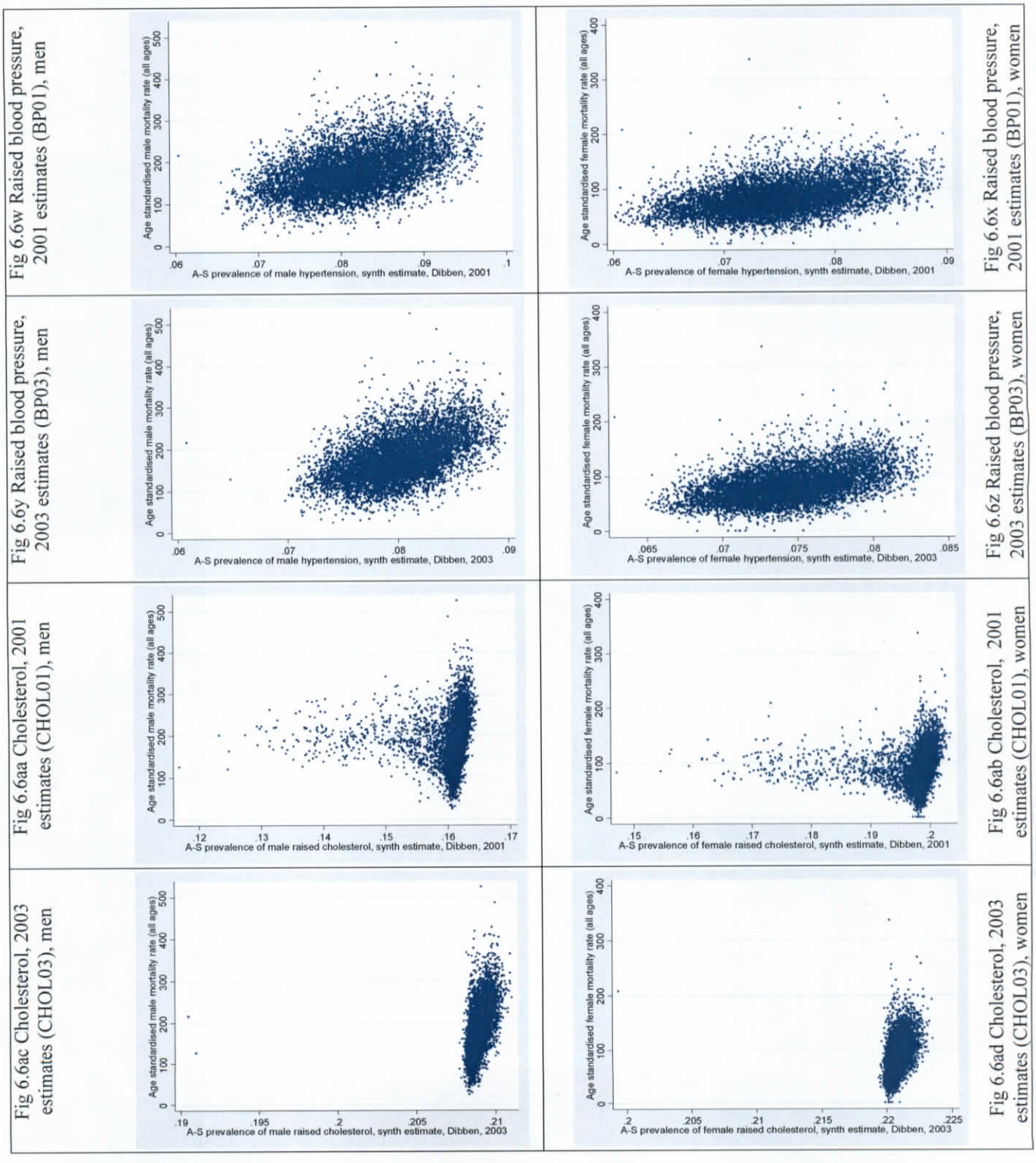
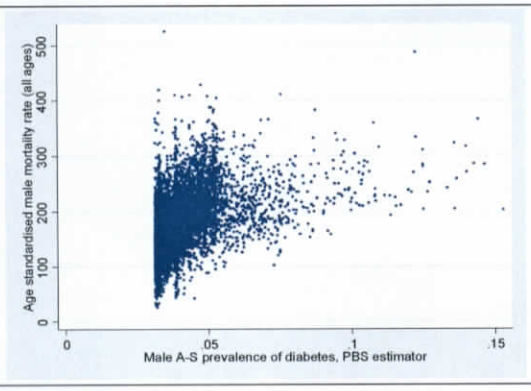


Figure 6.6 (cont.)

Fig 6.6ae Diabetes (DIAB), men



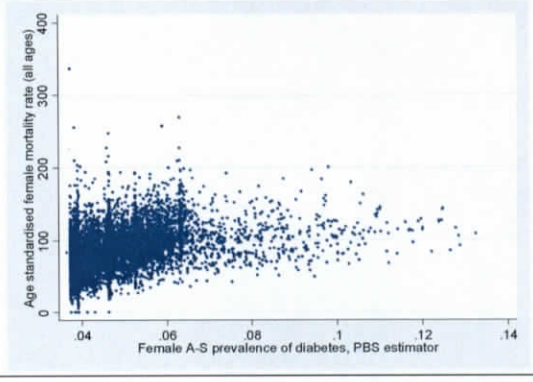


Fig 6.6af Diabetes (DIAB), women

Multi-level regression models

Table 8.8 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against environmental variables (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	605.5			319.0		
Mean max. temp (°C)	-31.0	2.3	<0.001	-16.5	1.2	<0.001
Sunshine (000s hrs / yr)	-33.3	15.3	0.029	-21.4	8.1	0.008
Air quality index (SDs)	40.4	3.7	<0.001	19.0	2.0	<0.001
Urban†	12.5	1.3	<0.001	5.3	0.7	<0.001
Metropolitan†	31.4	3.4	<0.001	15.1	1.9	<0.001
Ward-level variance explained:		4%			3%	
LA-level variance explained:		55%			59%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	2,539.8			1,559.8		
Mean max. temp (°C)	-156.4	18.7	<0.001	-100.0	9.9	<0.001
Sunshine (000s hrs / yr)	60.1	122.7	0.624	-32.3	64.9	0.617
Air quality index (SDs)	336.4	21.9	<0.001	196.1	12.0	<0.001
Urban†	74.3	6.2	<0.001	34.1	3.4	<0.001
Metropolitan†	236.5	17.5	<0.001	106.9	9.7	<0.001
Ward-level variance explained:		8%			7%	
LA-level variance explained:		20%			35%	

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations

Table 8.9 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against behavioural risk factor profiles of populations (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	180.4			83.8		
Unhealthy lifestyle (SDs)	17.9	0.4	<0.001	8.1	0.2	<0.001
Ward-level variance explained:		16%			11%	
LA-level variance explained:		49%			45%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	899.8			372.7		
Unhealthy lifestyle (SDs)	108.2	2.2	<0.001	55.4	1.1	<0.001
Ward-level variance explained:		23%			21%	
LA-level variance explained:		24%			33%	

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations

Table 8.10 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against deprivation (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	180.3			83.7		
Deprivation (SDs)	9.0	0.2	<0.001	4.2	0.1	<0.001
Ward-level variance explained:		24%			17%	
LA-level variance explained:		46%			40%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	898.9			371.4		
Deprivation (SDs)	50.2	0.8	<0.001	28.2	0.5	<0.001
Ward-level variance explained:		31%			32%	
LA-level variance explained:		45%			48%	

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations

Table 8.11 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against environmental variables and behavioural risk factor profiles of populations (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	330.4			188.4		
Mean max. temp (°C)	-16.0	1.8	<0.001	-9.4	1.0	<0.001
Sunshine (000s hrs / yr)	18.5	12.0	0.124	2.6	6.6	0.697
Air quality index (SDs)	32.7	3.1	<0.001	17.0	1.8	<0.001
Urban†	6.4	1.2	<0.001	2.3	0.7	<0.001
Metropolitan†	40.7	3.0	<0.001	23.6	1.8	<0.001
Unhealthy lifestyle (SDs)	16.4	0.4	<0.001	7.7	0.2	<0.001
Ward-level variance explained:		19%			14%	
LA-level variance explained:		77%			77%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	750.7			619.4		
Mean max. temp (°C)	-47.1	15.5	0.003	-42.8	7.9	<0.001
Sunshine (000s hrs / yr)	323.9	101.1	0.001	103.1	51.2	0.044
Air quality index (SDs)	242.7	19.0	<0.001	152.6	10.2	<0.001
Urban†	34.5	5.5	<0.001	12.3	3.1	<0.001
Metropolitan†	284.5	15.4	<0.001	160.7	8.6	<0.001
Unhealthy lifestyle (SDs)	106.2	2.2	<0.001	55.9	1.2	<0.001
Ward-level variance explained:		28%			26%	
LA-level variance explained:		48%			62%	

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations

Table 8.12 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against environmental variables and deprivation (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	437.3			241.1		
Mean max. temp (°C)	-13.4	1.8	<0.001	-8.6	1.1	<0.001
Sunshine (000s hrs / yr)	-46.8	11.8	<0.001	-26.4	6.8	<0.001
Air quality index (SDs)	-0.4	3.2	0.897	1.1	1.8	0.542
Urban†	1.6	1.2	0.171	0.3	0.7	0.697
Metropolitan†	-21.2	3.1	<0.001	-9.3	1.8	<0.001
Deprivation (SDs)	9.0	0.2	<0.001	4.2	0.1	<0.001
Ward-level variance explained:		24%			16%	
LA-level variance explained:		76%			73%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	1,619.1			1,038.3		
Mean max. temp (°C)	-48.3	15.2	0.001	-39.3	7.8	<0.001
Sunshine (000s hrs / yr)	-78.3	98.6	0.430	-108.2	50.9	0.033
Air quality index (SDs)	51.8	19.2	0.007	37.4	10.3	<0.001
Urban†	17.3	5.4	0.001	1.4	3.0	0.638
Metropolitan†	-26.7	15.8	0.091	-44.9	8.6	<0.001
Deprivation (SDs)	49.2	0.9	<0.001	28.2	0.5	<0.001
Ward-level variance explained:		31%			32%	
LA-level variance explained:		50%			61%	

† Compared to baseline of 'Coastal and Countryside'
SDs: Standard Deviations

Table 8.13 Multi-level regression models of (1) CHD mortality rates, and (2) CHD hospitalisation rates against behavioural risk factor profiles of populations and deprivation (wards nested in local authorities, n = 7,929)

(1) Mortality rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	180.3			83.7		
Unhealthy lifestyle (SDs)	7.2	0.5	<0.001	3.9	0.3	<0.001
Deprivation (SDs)	6.9	0.2	<0.001	3.2	0.1	<0.001
Ward-level variance explained:		25%			17%	
LA-level variance explained:		67%			62%	

(2) Hospitalisation rates models

	MEN			*WOMEN*		
Variable	*Beta*	*SE*	*p*	*Beta*	*SE*	*p*
Constant	898.8			371.7		
Unhealthy lifestyle (SDs)	38.3	2.8	<0.001	20.1	1.4	<0.001
Deprivation (SDs)	39.7	1.1	<0.001	23.0	0.6	<0.001
Ward-level variance explained:		33%			33%	
LA-level variance explained:		53%			60%	

SDs: Standard Deviations

Spatial error regression models

Table 8.14 Spatial error regression models of (1) CHD mortality rates and (2) CHD hospitalisation rates against environmental variables (wards, n = 7,929)

(1) Mortality rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	582.7				307.3			
Mean max. temp (°C)	-27.7	1.8	<0.001	✓	-15.4	1.0	<0.001	✓
Sunshine (000s hrs / yr)	-0.0	0.0	<0.001	x	-0.0	0.0	<0.001	✓
Air quality index (SDs)	33.9	3.4	<0.001	✓	17.7	1.9	<0.001	✓
Urban†	11.3	1.3	<0.001	✓	4.8	0.7	<0.001	✓
Metropolitan†	25.4	3.4	<0.001	✓	11.9	1.9	<0.001	✓
Spatial error	0.4	0.0	<0.001		0.4	0.0	<0.001	
Model r^2:	0.27				0.25			

(2) Hospitalisation rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	2,366.2				1,440.1			
Mean max. temp (°C)	-126.6	14.0	<0.001	✓	-83.0	7.4	<0.001	✓
Sunshine (000s hrs / yr)	-0.0	0.1	0.900	✓	-0.1	0.1	0.096	✓
Air quality index (SDs)	240.8	24.4	<0.001	✓	170.9	13.1	<0.001	✓
Urban†	62.2	6.0	<0.001	✓	27.2	3.4	<0.001	✓
Metropolitan†	210.4	18.9	<0.001	✓	82.8	10.5	<0.001	✓
Spatial error	0.7	0.0	<0.001		0.6	0.0	<0.001	
Model r^2:	0.47				0.48			

* ✓ = Direction of association and whether association is significant (p<0.01) agrees with equivalent multi-level regression models; x = otherwise.

† Compared to baseline of 'Coastal and Countryside'

SDs: Standard Deviations

Table 8.15 Spatial error regression models of (1) CHD mortality rates and (2) CHD hospitalisation rates against behavioural risk factor profiles of populations (wards, n = 7,929)

(1) Mortality rates models

Variable	*MEN* *Beta*	*SE*	*p*	*Agree with multi level model**	*WOMEN* *Beta*	*SE*	*p*	*Agree with multi level model**
Constant	179.6				83.3			
Unhealthy lifestyle (SDs)	15.9	0.4	<0.001	✓	7.3	0.2	<0.001	✓
Spatial error	0.4	0.0	<0.001		0.4	0.0	<0.001	
Model r^2:		0.34				0.29		

(2) Hospitalisation rates models

Variable	*MEN* *Beta*	*SE*	*p*	*Agree with multi level model**	*WOMEN* *Beta*	*SE*	*p*	*Agree with multi level model**
Constant	897.5				368.9			
Unhealthy lifestyle (SDs)	95.1	2.2	<0.001	✓	50.9	1.2	<0.001	✓
Spatial error	0.7	0.0	<0.001		0.7	0.0	<0.001	
Model r^2:		0.55				0.55		

* ✓ = Direction of association and whether association is significant (p<0.01) agrees with equivalent multi-level regression models; x = otherwise.

† Compared to baseline of 'Coastal and Countryside'

SDs: Standard Deviations

Table 8.16 Spatial error regression models of (1) CHD mortality rates and (2) CHD hospitalisation rates against deprivation (wards, n = 7,929)

(1) Mortality rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	180.1				83.6			
Deprivation (SDs)	8.7	0.2	<0.001	✓	4.1	0.1	<0.001	✓
Spatial error	0.4	0.0	<0.001		0.4	0.0	<0.001	
Model r^2:		0.39				0.32		

(2) Hospitalisation rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	899.2				370.3			
Deprivation (SDs)	49.7	0.9	<0.001	✓	28.3	0.5	<0.001	✓
Spatial error	0.6	0.0	<0.001		0.6	0.0	<0.001	
Model r^2:		0.59				0.60		

* ✓ = Direction of association and whether association is significant (p<0.01) agrees with equivalent multi-level regression models; x = otherwise.

† Compared to baseline of 'Coastal and Countryside'

SDs: Standard Deviations

Table 8.17 Spatial error regression models of (1) CHD mortality rates and (2) CHD hospitalisation rates against environmental variables and behavioural risk factor profiles of populations (wards, n = 7,929)

(1) Mortality rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	350.3				195.2			
Mean max. temp (°C)	-16.1	1.4	<0.001	✓	-9.8	0.8	<0.001	✓
Sunshine (000s hrs / yr)	7.5	8.9	0.400	✓	0.6	5.1	0.911	✓
Air quality index (SDs)	31.7	2.7	<0.001	✓	17.8	1.5	<0.001	✓
Urban†	6.8	1.9	<0.001	✓	2.6	0.7	<0.001	✓
Metropolitan†	35.3	2.9	<0.001	✓	20.1	1.7	<0.001	✓
Unhealthy lifestyle (SDs)	14.5	0.4	<0.001	✓	6.9	0.2	<0.001	✓
Spatial error	0.3	0.0	<0.001		0.2	0.0	<0.001	
Model r^2:		0.37				0.32		

(2) Hospitalisation rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	810.4				611.0			
Mean max. temp (°C)	-37.7	11.3	0.001	✓	-37.5	5.8	<0.001	✓
Sunshine (000s hrs / yr)	232.2	74.0	0.002	✓	59.6	37.6	0.113	✓
Air quality index (SDs)	203.2	19.9	<0.001	✓	157.8	10.3	<0.001	✓
Urban†	31.7	5.5	<0.001	✓	10.3	3.1	0.001	✓
Metropolitan†	252.4	16.7	<0.001	✓	131.7	9.2	<0.001	✓
Unhealthy lifestyle (SDs)	95.3	2.2	<0.001	✓	51.5	1.2	<0.001	✓
Spatial error	0.6	0.0	<0.001		0.6	0.0	<0.001	
Model r^2:		0.56				0.57		

* ✓ = Direction of association and whether association is significant (p<0.01) agrees with equivalent multi-level regression models; x = otherwise.

† Compared to baseline of 'Coastal and Countryside'

SDs: Standard Deviations

Table 8.18 Spatial error regression models of (1) CHD mortality rates and (2) CHD hospitalisation rates against environmental variables and deprivation (wards, n = 7,929)

(1) Mortality rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	425.0				234.5			
Mean max. temp (°C)	-12.5	1.3	<0.001	✓	-8.5	0.8	<0.001	✓
Sunshine (000s hrs / yr)	-48.3	8.5	<0.001	✓	-24.4	5.0	<0.001	✓
Air quality index (SDs)	1.7	2.7	0.517	✓	3.5	1.6	0.027	✓
Urban†	1.7	1.2	0.156	✓	0.4	0.7	0.541	✓
Metropolitan†	-26.7	3.0	<0.001	✓	-12.3	1.8	<0.001	✓
Deprivation (SDs)	8.8	0.2	<0.001	✓	4.0	0.1	<0.001	✓
Spatial error	0.3	0.0	<0.001		0.3	0.0	<0.001	
Model r^2:		0.41				0.34		

(2) Hospitalisation rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	1,477.9				945.7			
Mean max. temp (°C)	-25.8	11.2	0.021	x	-28.8	5.7	<0.001	✓
Sunshine (000s hrs / yr)	-154.2	73.4	0.036	✓	-143.5	36.9	<0.001	x
Air quality index (SDs)	2.3	20.2	0.911	x	38.8	10.4	<0.001	✓
Urban†	17.0	5.4	0.002	✓	1.1	3.0	0.714	✓
Metropolitan†	-25.2	16.9	0.137	✓	-57.0	9.2	<0.001	✓
Deprivation (SDs)	49.2	1.0	<0.001	✓	28.2	0.5	<0.001	✓
Spatial error	0.6	0.0	<0.001		0.6	0.0	<0.001	
Model r^2:		0.59				0.60		

* ✓ = Direction of association and whether association is significant (p<0.01) agrees with equivalent multi-level regression models; x = otherwise.

† Compared to baseline of 'Coastal and Countryside'

SDs: Standard Deviations

Table 8.19 Spatial error regression models of (1) CHD mortality rates and (2) CHD hospitalisation rates against behavioural risk factor profiles of populations and deprivation (wards, n = 7,929)

(1) Mortality rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	180.0				83.5			
Unhealthy lifestyle (SDs)	7.6	0.4	<0.001	✓	4.0	0.2	<0.001	✓
Deprivation (SDs)	6.4	0.2	<0.001	✓	3.0	0.1	<0.001	✓
Spatial error	0.3	0.0	<0.001		0.3	0.0	<0.001	
Model r^2:		0.40				0.33		

(2) Hospitalisation rates models

Variable	*MEN*				*WOMEN*			
	Beta	*SE*	*p*	*Agree with multi level model**	*Beta*	*SE*	*p*	*Agree with multi level model**
Constant	896.7				369.1			
Unhealthy lifestyle (SDs)	33.5	2.7	<0.001	✓	19.5	1.3	<0.001	✓
Deprivation (SDs)	39.6	1.2	<0.001	✓	22.7	0.6	<0.001	✓
Spatial error	0.6	0.0	<0.001		0.6	0.0	<0.001	
Model r^2:		0.59				0.60		

* ✓ = Direction of association and whether association is significant ($p<0.01$) agrees with equivalent multi-level regression models; x = otherwise.
SDs: Standard Deviations