ESSAYS ON THE LONGITUDINAL ANALYSIS OF HEALTH AND HEALTHCARE DATA

Daniel David Henry Howdon

PhD

University of York

Economics

September 2014

Abstract

The central theme of this thesis is the longitudinal analysis of health and healthcare data.

Chapter 2 uses the first wave of, and latest longitudinal follow-up to, the Health and Lifestyle Survey (HALS) to investigate the social gradient in cancer, considering both lifetime incidence and duration models of time-to-cancer – healthy time lived before developing cancer. Contrary to previous claims regarding the relationship between circumstances and the development of cancer, such as Deaton (2002) and Wilkinson and Pickett (2010), a social gradient in time-to-cancer is observed, with those in the lowest two social classes developing cancer approximately 15% sooner (significant at the 5% level) than individuals in the highest social class. This relationship holds after excluding smokers from the sample. No significant gradient is observed when only lifetime incidence of cancer is considered.

Chapter 3 investigates the relationship between smoking and ill-health, with a focus on cancer outcomes. A discrete latent factor model for smoking and health outcomes, allowing for these to be commonly affected by unobserved factors, is jointly estimated, using the British Health and Lifestyle Survey (HALS) dataset. Post-estimation predictions suggest the reduction in time-to-cancer to be 5.7 years for those with a smoking exposure of 30 pack-years, compared to never-smokers. Estimation of posterior probabilities for class membership show that individuals in certain classes exhibit similar observables but highly divergent health outcomes, suggesting that unobserved factors in this model substantially determine these outcomes. The use of a joint model changes the results substantially. The results show that failure to account for unobserved heterogeneity leads to differences in survival times between those in different social classes and with different smoking exposures to be overestimated by more than 50% (males, with 30 pack-years of exposure).

Chapter 4 uses Hospital Episode Statistics, English administrative data from the Department of Health, to further investigate the red herring thesis, as advanced by Zweifel et al. (1999). We use a sample of over 100,000 individuals who used healthcare in the financial year 2005/06 and had died by the end of the financial year 2012/13. We use a panel structure to follow individuals over seven years of this administrative data, containing estimates of inpatient healthcare expenditures (HCE), information regarding individuals' age, time-to-death (TTD), and morbidities at the time of their admission. We find that, while TTD might better explain HCE than does age, TTD itself merely proxies for individuals' morbidities, and no longer explains differences in HCE once we condition on morbidities. Our results point to an important role for including estimates of future changes in morbidity when estimating future HCE.

List of Contents

Abstract **2**

List of Tables **5**

List of Figures **7**

Acknowledgments **8**

Declaration **9**

1 Introduction **10**

2 Chapter 2 **15**

2.1 Introduction 17

2.2 Data 19

2.2.1 Clarifications and discussion of variables 20

2.3 Theoretical framework 25

2.4 Methods 27

2.4.1 Statistical models 27

2.5 Results 30

2.6 Conclusions 36

3 Chapter 3 **41**

3.1 Introduction 43

3.2 Background 44

3.3 Data 46

3.4 Theoretical model 48

3.5 Methods 50

3.5.1 The model . . . 50
3.5.2 Joint likelihood . . . 56
3.5.3 Key covariates and interpretation of parameters . . . 59
3.6 Results . . . 63
3.6.1 Posterior probabilities . . . 68
3.7 Counterfactual simulations . . . 69
3.8 Discussion . . . 74
3.A Appendix . . . 77

4 Chapter 4 **83**
4.1 Introduction . . . 85
4.1.1 Compression of morbidity . . . 87
4.1.2 Age and time-to-death . . . 89
4.2 Data . . . 93
4.2.1 Data sources . . . 93
4.3 Theoretical model . . . 103
4.4 Econometric model . . . 104
4.5 Results . . . 108
4.6 Conclusions . . . 112
4.A Appendix . . . 114

Conclusions **123**

Bibliography **127**

List of Tables

2.1 Descriptive statistics, full survey and cleaned full survey 21
2.2 Deaths where cancer is listed on an individual's death certificate, with the individual never registered as developing cancer (full survey) 22
2.3 Comparison of AIC and BIC scores 31
2.4 Lognormal and probit regressions, smoking variables excluded 33
2.5 Lognormal and probit regressions, smoking variables included 34
2.6 Lifetime incidence (lognormal) and healthy time before cancer (probit) models 35

3.1 Mass points: 4 points of support 58
3.2 Summary statistics (all 3784 observations) 59
3.3 DLFM results – main coefficients 65
3.4 DLFM results (2) – latent factor coefficients and class membership probabilities 66
3.5 Single equation - cancer registration 66
3.6 Descriptive statistics, by most probable latent class based on posterior probabilities. 69
3.7 Counterfactual estimates – median survival time to onset of cancer (years) . 70
A3.1 Comparison of baseline hazards 77
A3.2 Deaths where cancer is listed on an individual's death certificate, with the individual never registered as developing cancer 78

4.1 Summary statistics (Quarter 1, men. n=55039. 97
4.2 Summary statistics (Quarter 1, women. n=56097.) 97
4.3 Results 107
4.4 Diagnoses with highest estimated associated increase in costs, men 108
4.5 Diagnoses with highest estimated associated costs, women 108

A4.1 Clinical Classifications Software (CCS) groupings 114

List of Figures

2.1 Types of individual 28
2.2 Cox-Snell residuals plots 32

3.1 Types of observed outcomes 51
3.2 Estimated survival curves for cancer onset by smoking behaviour (males) . 71
3.3 Estimated survival curves for cancer onset by latent class (males) 71
3.4 Estimated survival curves for cancer onset (males) – comparison of single-equation and full model results 72
3.5 Estimated percentage reduction in time-to-cancer by smoking exposure – comparison of single-equation and full model results for current-smokers . . 72

4.1 Stylised change in survival curves 88
4.2 Stylised change in health profiles 89
4.3 Healthcare expenditures by age and proximity to death, males 100
4.4 Healthcare expenditures by age and proximity to death, females 101
4.5 Healthcare expenditures (y-axis, GBP) by time to death (x-axis, quarters), males by age 102
4.6 Average healthcare expenditures according to quarters to death (log scale for x- and y- axes) 103
4.7 Change in HCE according to time-to-death and age, hypothetical individual dying at 75 (top – men, bottom – women) 110

Acknowledgements

I extend my thanks to my supervisors, Nigel Rice and Andrew Jones, and the third member of my advisory panel, Hugh Gravelle, for providing invaluable help and support throughout the four years of my time spent writing this thesis. Eugenio Zucchelli and Pedro Rosa Dias, before flying the nest, also provided encouragement and practical support in the early stages of Chapters 2 and 3. John Hey made the teaching I carried out alongside my research as enjoyable a distraction as it could possibly have been. James Lomas was, for the last four years, the apotheosis of a everything a friend could be. There are many other people to thank: people who know who they are, still others whom I miss terribly. I hesitate to even attempt to thank my parents, to whom mere thanks is pitifully inadequate, to whom offering thanks comes close to doing violence.

> Few books today are forgivable. Black on canvas, silence on the screen, an empty white sheet of paper are perhaps feasible. There is little conjunction of truth and social "reality". Around us are pseudo-events, to which we adjust with a false consciousness adapted to see these events as true and real, and even as beautiful. In the society of men the truth resides now less in what things are than in what they are not. Our social realities are so ugly if seen in the light of exiled truth, and beauty is almost no longer possible if it is not a lie. What is to be done? We who are still half alive, living in the often fibrillating heartland of a senescent capitalism – can we do more than reflect the decay around and within us? Can we do more than sing our sad and bitter songs of disillusion and defeat? The requirement of the present, the failure of the past, is the same: to provide a thoroughly self-conscious and self-critical human account of man. Laing (1967).

I do not seek forgiveness; I merely offer my apologies.

Declaration

I confirm that the work presented in this thesis is my own, except where co-authorship is explicitly acknowledged. I received financial assistance from the Economic and Social Research Council (ESRC), attached to the funding for the Health, Econometrics and Data Group (HEDG) under the Large Grant Scheme (reference: RES-060-25-0045).

Chapter 2 is single-authored work and appears, in an earlier version, in the University of York's Health, Econometrics and Data Group (HEDG) Working Paper Series as WP12/06. A highly-abridged, non-technical version also appears in *Britain in 2014*, a magazine published annually by the Economic and Social Research Council. Discussions with Richard Cookson proved extremely fruitful in the final drafting stage of this chapter.

Chapter 3 is co-authored work with Andrew M. Jones. I am the lead author for this paper, having prepared the data, carried out the empirical analysis and written the first draft. An earlier version appears in the HEDG Working Paper Series as WP13/14. A preliminary version was presented in September 2012 at Erasmus University Rotterdam, as part of the Netspar Workshop, *Income, health, work and care across the life cycle*.

Chapter 4 is co-authored work with Nigel Rice. I am again the lead author for this paper, having prepared the data, carried out the empirical analysis and written the first draft. A preliminary version was presented in July 2014 at Trinity College, Dublin, as part of the International Health Economics Association's 10th World Congress, *Health Economics in the Age of Longevity*. Katja Grasic provided assistance with data preparation for this chapter.

Versions of Chapters 2, 3 and 4 have each been presented, at various stages of development, at the University of York's Health, Econometrics and Data Group Seminar Series between 2011 and 2014. Since the original submission of this thesis, Chapter 2 has appeared as Howdon, D and Jones, A. M. (2015), 'A discrete latent factor model for smoking, cancer and mortality', *Economics & Human Biology* **18**, 57–73.

Introduction

This collection of papers centres on the longitudinal analysis of health and healthcare data, considering topical issues in both policy and research: inequalities in the development of cancers, determinants of cancer, and healthcare expenditures in the terminal years of individuals' lives. Each chapter is underpinned by the theoretical approach outlined in Grossman (1972): this model is amended and augmented according to the research question employed. This collection of papers makes methodological contributions: featuring, as far as literature searches have revealed, the first papers to jointly model smoking behaviours, death and cancer outcomes from a pre-diagnosis starting point, while allowing individuals in heterogeneous groups to select into smoking based on expectations regarding health outcomes. These papers have important implications for policy, in areas from the prediction and allocation of healthcare costs to the relevance of socioeconomic inequality in explaining differences in health-related, particularly cancer-related, outcomes.

Some elaboration on the Grossman model (Grossman, 1972) seems appropriate at this stage. This has become the canonical theoretical model for research in topics relating to individuals' decision making, where their demand for health is relevant. In its original and full form, it presents an individual's multi-period optimisation (utility-maximising) decision when allocating time and income between healthcare inputs and consumption good inputs, in order to produce desiderata of health and consumption goods, where these individuals act as producer-consumers of each. Individuals face these economic (income and time) constraints, as well as biological constraints that cause the individual's 'health stock' to fall over time in the absence of investment, and to at some point fall to such a low level that precipitates the individual's death. While the model directly considers health as a desired good, predictions regarding consumption health inputs (primarily, healthcare with its associated expenditures) which produce this can also be drawn.

Chapter 2 directly examines the relationship between social class and individuals'

prospects for developing cancer. Existing research tends to find no significant social gradient, or social inequality in cancers. Deaton (2002) claims that data collected on UK civil servants (Marmot et al., 1978, 1991) show no social gradient in any cancer, once differential smoking behaviours are accounted for. While the authors of 2010's The Spirit Level (Wilkinson and Pickett, 2010) argue that sufficient evidence exists to prove socioeconomic inequality in health generally, and many diseases specifically, they find no social gradient in breast cancer and 'only small class differences' in prostate cancer.

This chapter augments the standard Grossman model with an additional health outcome: that of time to cancer. While Grossman (1972) features an individual with a finite lifespan that ends at some future time period $t = n$ (the time period in which he dies), I introduce a second health outcome: that an individual has a finite cancer-free lifespan that ends at some future time period $t = m$ (the time period in which he is registered as having developed cancer). This implies that the individual will be observed to develop cancer if and only if $m \leq n$: that is, if death does not intervene before the individual develops cancer.[1] If individuals in lower social classes disproportionately die before they develop cancer (that is, if they disproportionately belong to the set of individuals for whom $m > n$), a comparison of lifetime incidence by social class will provide an incomplete picture of the burden of disease faced by individuals in lower social classes. This potential problem is overcome by modelling cancer-free lifespan, rather than lifetime incidence: this produces a dependent variable that treats cancer registration at an earlier age as different from cancer registration in old age.

The chapter uses data gathered on all individuals aged 45 and over in the Health and Lifestyle Survey (HALS) dataset, consisting of socioeconomic and health-related information gathered on a cross-section of the population of Great Britain, as well as subsequently-collected information about the health status of those sampled, in order to consider this question. The chapter replicates existing findings that, when lifetime incidence of cancer alone is considered, evidence of a social gradient in cancer is weak. It is also shown that naive estimates, which fail to condition on smoking, would tend to result in higher estimates of social inequality in cancer. However, it is demonstrated that, while individuals in lower social classes develop cancer at roughly the same lifetime rate as those in higher social classes, they do so sooner in life – even after the sample is restricted to consider non-smokers only, suggesting a greater burden of disease for those in lower social classes.

[1] This weak inequality arises as the paper includes individuals whose cancer is detected only post-mortem. This information is derived from analysis of death certificate data, included in HALS.

Individuals in the lowest two social classes are found to develop cancer approximately 15% sooner than those in the highest social class. This implies an underestimate of the inequality in the burden of disease (in this case, cancer) faced by individuals in lower social classes, even after the effect of differential smoking behaviours is removed.

Chapter 3 examines the relationship between smoking, lifestyles and cancer, using data gathered on all individuals aged 45 and over in the Health and Lifestyle Survey. Building on work in Chapter 2, smokers are included in this sample, and smoking decisions (taking up smoking, time-to-taking up smoking and time-to-quitting smoking) and health outcomes (time-to-death, and time-to-registration as a cancer sufferer) are jointly modelled, using a discrete latent factor approach to allow these to be commonly affected by unobservable factors in the system of equations (Heckman and Singer, 1984; Mroz, 1999).

While randomised trials in this area are infeasible, often no attempt is made in existing literature to account for individuals' unobservable characteristics which may jointly affect smoking behaviours and health outcomes. One of the most important studies in initially establishing the link between smoking and ill-health – the British Doctors Study (Doll and Hill (1954) and subsequent papers) – focuses on one small stratum of society, and includes more limited information on smoking behaviour than that we are able to exploit. Further, we are able to rely upon almost 25 years of follow-up data, automatically collected through administrative records, of the individual's death and cancer registration.

This chapter is again based on the model of health demand proposed by Grossman (1972). As in the previous chapter, a second health outcome of cancer is introduced, and the individual is said to develop cancer at some time period m. The individual is assumed to maximise a utility function which is explicitly conceived of in terms of optimal choices around health-related goods, and two types of consumption goods: those not affecting health, and those affecting health. We assume that individuals' choices around consumption goods that affect health are affected by the individual's exogenously-determined, but unobserved to the researcher, health constraints: his initial (genetic, or early-life) stock of health, and the individual's discount rate at any future time period. This means that both health outcomes (cancer and lifespan) and decisions regarding consumption goods that affect health (in our model, decisions regarding smoking) are jointly affected by these unobserved factors. This motivates the use of a joint model which allows the impact of such unobservable factors to be estimated and recovered.

We find that jointly modelling these decisions and outcomes in this way alters results

substantially: unobservable factors are found to be responsible in large part for differentials between smokers and non-smokers. We find the reduction in time to cancer to be 5.7 (5.8) years for men (women) who were smokers at the time of HALS, with a total observed exposure of 30 pack-years, compared to never-smokers at the time of HALS, approximately 50% lower than when cancer outcomes are modelled in single-equation form. This implies that previous work which has examined the effect of smoking on cancer may overestimate causal effects: that individuals who select into cancer are different in relevant unobserved characteristics which jointly affect smoking behaviours and cancer outcomes.

Chapter 4 seeks to establish the causal relationship between ageing and healthcare expenditures (HCE). Observed changes in life expectancy and morbidity in the mid-to-late 20th century culminated in fears of an 'expansion of morbidity', or 'failures of success' (Gruenberg, 2005). An ageing population would, according to this thesis, lead to – on average – a greater burden of morbidity among elderly groups: that is, that benefits of better and more widespread healthcare provision would, by preventing or curing previously quickly-fatal diseases, lead to a greater proportion of the population suffering from costly chronic conditions. Research papers released by bodies such as the International Monetary Fund (Heller et al., 1986) pointed to a positive relationship between age and HCE. These views found an echo, over 25 years later in remarks made by Andrew Lansley, the UK's then-Secretary of State for Health, who identified 'the number of people aged over 85 in this country will double in the next 20 years' as one of two factors in 'costs... rising at an unaffordable rate' (Lansley, 2012).

In contrast to this, the emergence of a 'compression of morbidity' strand of literature, beginning with Fries (1980), gave reason to suspect that the developed world had entered an era in which individuals age more heathily than previously suspected and that, consequently, the implications for future HCE are moot. Although individuals born in recent decades see a longer expected lifespan, they will, *contra* the 'failures of success' thesis, not have an increased number of years living with chronic conditions. More recently, this has been empirically evidenced by, *inter alia*, Freedman et al. (2002), Romeu Gordo (2011) and Cutler et al. (2013). Due to improvements in living standards and medical treatment in the late 20th century and beyond, individuals may be able to entirely avoid the onset of costly chronic conditions. Complementing this, a 'red herring' strand of research, starting with Zweifel et al. (1999), identified the true driver of HCE as being closeness to death (usually termed time-to-death, or TTD, in existing literature) rather than ageing: that

once empirical analysis conditioned on TTD as well as age, the impact of age was muted. A Grossman framework for this chapter conceives of TTD in existing research as a proxy for health stock, and a proxy that is better replaced by observations of individuals' actual morbidity status.

The chapter investigates this further, using a sample of over 100,000 inpatient users of healthcare in England in 2005/6, who subsequently died within the following seven years, taken from in an administrative data from the UK's Department of Health. A reduced form Grossman model is used to underlie the analysis carried out. Morbidity markers are assumed to proxy for health and to causally affect HCE in a given time period. Because HCE is assumed to affect morbidity in future time periods (and thus affect TTD), a retrospectively-defined TTD variable becomes endogenous in a given time period, according to the Grossman framework employed. Small area measures of mortality (years of potential life lost in the UK government's defined lower super output areas) are used to instrument for a contemporaneous TTD variable.

The results of research in this chapter suggest that TTD is itself a red herring: that once we condition on individuals' observed morbidity, there is no economically-significant role for age and TTD in explaining HCE. This has both implications for projections of future HCE, as well as the allocation of healthcare budgets in a more short-term setting. While TTD may be a useful proxy for morbidity – which more directly explains HCE – the nature of TTD in explaining HCE will not necessarily be constant in future. Combined with the compression of morbidity thesis – suggesting that individuals age, and approach death, in greater health – these results give reason to treat the use of TTD rather than expected morbidity, in these settings, with some caution.

Chapter 2

Time and chance happen to them all? Duration modelling versus lifetime incidence of cancer.

Daniel Howdon

Department of Economics and Related Studies & Centre for Health Economics, University of York, YO10 5DD, UK

Tel.: +44 1904 32 1411.
E-mail address: daniel.howdon@york.ac.uk

Abstract

While much current work on socioeconomic inequality in cancer looks at lifetime incidence of cancer, it is more informative to consider survival times: healthy time lived without cancer. This paper uses the first wave of, and latest longitudinal follow-up to, the Health and Lifestyle Survey (HALS) to investigate the social gradient in cancer, considering both lifetime incidence and duration models of time-to-cancer. Contrary to previous claims regarding the relationship between circumstances and the development of cancer, such as Deaton (2002) and Wilkinson and Pickett (2010), a social gradient in time-to-cancer is observed, with those in the lowest two social classes developing cancer approximately 15% sooner (significant at the 5% level) than individuals in the highest social class. This relationship holds after excluding smokers from the sample. No significant gradient is observed when only lifetime incidence of cancer is considered.

JEL classification: C41; I14

Keywords: health; duration analysis; smoking; mortality; inequality of opportunity; determinants of health; lifestyles.

2.1 Introduction

While strong evidence exists regarding a social gradient in lifespan overall and illnesses such as cardiovascular disease, the existence of a social gradient in cancer is more controversial. Deaton (2002) argues that the Whitehall Studies (Marmot et al., 1978, 1991) show no social gradient in any cancer apart from lung cancer, the gradient in which is entirely explained by differences in smoking between the occupational grades. Studies, defining socioeconomic status variously (by occupation, income, levels of education), find evidence of socioeconomic inequality in cancer incidence, failing to adjust for differential smoking behaviours in the case of Mackillop et al. (2000), and adjusting for differential smoking behaviours in the case of (Mao et al., 2001). In a meta-analysis of 64 studies of the link between lung cancer and socioeconomic outcomes, Sidorchuk et al. (2009) find evidence of socioeconomic inequality (according to occupation and educational level) in cancer incidence which exists when no control is made for smoking behaviours, and persists even when smoking behaviours are adjusted for.

Using Canadian data drawn from individuals with lung cancer and a constructed control group, Nkosi et al. (2012) argue, however, that such studies may take only an incomplete account of smoking outcomes and report no relationship between socioeconomic status (variously defined) and incidence of the disease once appropriate and full account is taken of differential smoking behaviours across socioeconomic groups, adjusting for different functional forms of the relationship between smoking behaviours and health outcomes. Despite finding social gradients in health overall and in many diseases, Wilkinson and Pickett (2010) find no social gradient in breast cancer, and 'only small class differences' in prostate cancer. Some studies (Lyratzopoulos et al., 2012) into the stage at diagnosis find socioeconomic inequalities in this regard and, notably, inequality in stage at diagnosis – as far as studies find evidence of this in the direction of those in lower socioeconomic groups – would tend to mitigate against finding evidence of social inequality in the age of registration, which cannot precede diagnosis. Attention in this area has focused on incidence of cancer rather than survival time to cancer (such as, additionally, Singh et al. (2003); Banks et al. (2006); Dalstra et al. (2005)): the differences in implications of this, and the subsequent results obtained by the use of such approaches are highlighted in this

paper.

No study has, as yet, exploited the Health and Lifestyle Survey (HALS) dataset and the subsequent cancer follow-ups for the purposes of carrying out such an investigation. Hiatt and Breen (2008) identify the question of why social determinants are correlated with the development of cancer – because they are correlated with already-understood risk factors for cancer or biological factors, or because they are inherently causative of cancer – as 'a key question in cancer research'. Further, the authors call for the investigation of such factors to encourage 'a more complete understanding of the causes of cancer'. While the possibility of causal analysis is limited by the available dataset, this paper provides evidence of a link between social class and cancer, independent of differences in smoking behaviours, and goes some way towards addressing these concerns.

The importance of smoking in this analysis is clear: a link between smoking and ill-health in general, and cancer specifically, is uncontroversial. The risks of smoking have been well-explored since the link between smoking and lung cancer was made by Doll and Hill (1954). Smoking has been associated with a greater propensity to develop various cancers and other diseases (for instance, deaths from lung cancer are estimated to occur with between 10.8 and 24.9 times the frequency in smokers as in non-smokers (Doll, 1998)) and is estimated to be responsible for approximately 30% of all cancer deaths in developed countries, as well as causing deaths from respiratory, circulatory and other problems (Department of Health and Human Services, 1989; Jones et al., 2007; Peto et al., 2006; Vineis et al., 2004). Using Health Survey for England data from 1998 to 2006, Vallejo-Torres and Morris (2010) estimate that 2.3% of all socioeconomic (income-related) inequality in health observed was due to smoking. Successive reports by the US Surgeon General (Department of Health and Human Services, 1989, 2004, 2010) have examined the evidence linking smoking with mortality and diseases including cancer, increasingly making stronger links over time, with 30 diseases listed in the 2004 report for which evidence was 'sufficient to infer a causal relationship'. Doll (1998) provides a useful summary of the history of mounting evidence regarding the links between smoking and ill-health. Given this, along with the disproportionate levels of smoking among those in lower social classes and the possibility of this acting as a confounding factor when estimating relationships between social class and ill-health, three different models are estimated, each treating

smoking variables differently. A first model includes smokers in the sample, but excludes variables for smoking behaviours in the regression; a second model includes smokers in the sample and includes variables for smoking behaviours in the regression; a third and final model excludes smokers completely, in order to strip out (as far as possible) any error in the specification of the functional form relating smoking behaviours and cancer outcomes. Further, only individuals aged 45 or over at the time of HALS1 are included in this model, to reduce the confounding of cancer registrations with genetic factors unrelated to the covariates used in the health outcome models. An association, significant at the 5% level, between social class and accelerated time to cancer is observed.

2.2 Data

This paper uses the Health and Lifestyle Survey (HALS), conducted between 1984 and 1985, together with the most recent longitudinal follow-up (that of July 2009). HALS contains data on lifestyle, behaviours (such as smoking and alcohol consumption) and circumstances of a large cross-section of a representative sample of individuals in Great Britain (Cox et al., 1993)[1]. Data collection consisted of a one-hour face-to-face interview to collect information on individuals' health, lifestyles, and socio-demographic data. It involved a visit from a nurse to collect information on physiological and cognitive function, and a self-completed questionnaire to gather information regarding psychiatric health and personality (Cox et al., 1993; Jones et al., 2007). Details of individuals' diagnoses with cancer and information relating to individuals' deaths (such as date and cause of death) were subsequently provided to the HALS team. Such data, including details from death certificates and cancer diagnoses, are used, correct to the beginning of July 2009 – the seventh deaths revision and fourth cancer revision (University of Cambridge Clinical School, 2009). 9,003 individuals were initially entered into the study of whom, as of this revision, the status of 97.8% had been flagged on the NHS's Central Register at the Office for National Statistics: 2,883 individuals have been flagged as dead and 1,468 coded for cancer.

The original HALS dataset contains 9,003 observations: after cleaning, this is reduced

[1] That is the United Kingdom, not including Northern Ireland.

to 8,213. This cleaning procedure was particularly assiduous for smoking-related variables, and required the removing of individuals who did not report their smoking status, individuals who were reported to be current smokers but did not report age at starting, individuals who were reported to have started smoking at a date later than their date of interview, ex-smokers who did not declare when they quit or the age at which they began to smoke, current and ex-smokers who did not provide an average number of cigarettes smoked. Data was also cleaned, and records for individuals deleted, where they were reported to have died on or before their date of interview, registered for cancer on or before their date of interview, or where any other variable included in the regression was missing for the individual. The exclusion of individuals aged under 45 at the time of HALS further reduces the sample size to 3,800, and the exclusion of smokers (in the final model) reduces the sample to 1,397 individuals.

Descriptive statistics for the full survey, and the full survey after cleaning, are presented in Table 2.1[2].

Given that this survey took a cross-section of individuals living in Great Britain at one point in time, there is the possibility for, and such cases exist where, individuals had been diagnosed with cancer prior to their being interviewed for HALS. While the exclusion of those living with cancer in 1985/6 does mean that the sample is necessarily less representative of the population, this avoids the problem of the inclusion of such individuals with apparently negative survival times, which cannot be modelled using the distributions employed here.

2.2.1 Clarifications and discussion of variables

It must be borne in mind there are delays involved in the registration of deaths and developing cancer, and that delays are not uniform in all cases. The latest HALS follow-up manual (University of Cambridge Clinical School, 2009) suggests that cancer registrations tend to be slower to reach the NHS Central Register than death notifications (although such registrations are 'probably' complete up to the end of 2007), and that missing cases will exist due to 'patchy' returns from regional registries. Spikes are recorded in 2008

[2]It should be noted that some of these descriptive statistics – particularly those smoking-related – are potentially unreliable pre-cleaning.

		Mean			
Description	Label	n=9003	n=8213	n=3800	n=1397
HALS smoker	start	0.346	0.340	0.304	-
Pack-years (HALS1 smoker)	packyrss	7.082	7.414	9.845	-
Pack-years squared (HALS1 smoker)	packyrs2s	0.026	0.027	0.043	-
HALS ex-smoker	quit	0.248	0.254	0.329	-
Pack-years (HALS1 ex-smoker)	packyrsq	4.881	5.098	8.861	-
Pack-years squared (HALS1 ex-smoker)	packyrs2q	0.024	0.026	0.051	-
Non-prudent drinker of alcohol	NPAD	0.121	0.124	0.087	0.039
Eats meat 3+ times per week	redmeat3	0.452	0.451	0.527	0.510
Highest educational qualification: degree	lhqdeg	0.046	0.047	0.027	0.037
Highest educational qualification: other	lhqoth	0.006	0.006	0.006	0.006
Highest educational qualification: A-level or equivalent	lhqA	0.044	0.041	0.031	0.029
Highest educational qualification: O-level or equivalent	lhqO	0.143	0.148	0.072	0.080
Long-term unemployed (one year or more)	ltunemp	0.03	0.029	0.020	0.007
Not working due to sickness	sick	0.021	0.020	0.038	0.020
Retired	retd	0.210	0.209	0.451	0.481
Single	single	0.17	0.166	0.066	0.077
Separated or divorced	sepdiv	0.06	0.061	0.050	0.048
Widowed	widowed	0.087	0.080	0.166	0.204
Male	male	0.434	0.443	0.451	0.282
Social class 2 or 3	sc23	0.655	0.681	0.652	0.674
Social class 4 or 5	sc45	0.277	0.289	0.319	0.293
1920s birth	bc20	0.156	0.155	0.335	0.268
1930s birth	bc30	0.153	0.157	0.307	0.324
1940s birth	bc40	0.193	0.201	0.000	0.001
1930s or 1940s birth	bc3040	0.347	0.358	0.307	0.325
1950s birth	bc50	0.189	0.195	-	-
1960s birth	bc60	0.13	0.126	-	-

Table 2.1: Descriptive statistics, full survey and cleaned full survey

and 2009 for individuals who died with cancer present without ever being registered as developing such a disease (Table 2.2), suggesting that some late returns may exist for this revision[3]. Comparison of the previous HALS follow-up (to April 2005) with data held in this latest follow-up shows, however, that no cancer registrations were late – i.e. were included in the July 2009 follow-up with a date of April 2005 or earlier[4]. Furthermore, the age at the time of an individual's first cancer registration is not the same as the age of the individual first developing cancer. Diagnosis of cancer does not immediately take place upon the individual developing the disease, nor does it occur at the same stage of development of the cancer across individuals, or over time. In particular, the stage at diagnosis has varied over time, with US National Cancer Institute (2006) showing declines in the rates of late-stage diagnoses of cases of cancers of the cervix, colon, prostate and rectum between 1980 and 2006.

Year of death	No. of deaths	Percentage
1985	5	3.42
1986	17	11.64
1987	14	9.59
1988	18	12.33
1989	27	18.49
1990	23	15.75
1991	5	3.42
1992	1	0.68
1993	4	2.74
1994	4	2.74
1995	1	0.68
1996	1	0.68
1997	2	1.37
2000	2	1.37
2001	1	0.68
2002	2	1.37
2006	2	1.37
2007	1	0.68
2008	10	6.85
2009	4	2.74
Total	146	

Table 2.2: Deaths where cancer is listed on an individual's death certificate, with the individual never registered as developing cancer (full survey)

[3]This data is obtained using the Stata icd9 command to search for individuals whose death certificate shows any cancer (codes in the range 140 to 239.99).

[4]Seven death registrations were, however, late by this measure.

Following Balia and Jones (2011), the measure of social class is this paper is the registrar general's measure of social class (RGSC) for the individual's occupation or most recent occupation. Social class 1 is made up of individuals classified as being in 'professional occupations', social class 2 'managerial and technical occupations', social class 3 'skilled non-manual and manual occupations, social class 4 'partly-skilled occupations' and social class 5 'unskilled occupations'. While this measure is an imperfect proxy for an individual's status within a pattern of inequality, its use is dictated by partly practical and partly theoretical reasons. While income (both measured at the personal and household level) is included in the HALS dataset and may be argued to be a preferable measure of the individual's relative social position, it is characterised by substantial amounts of missing data, with 23% and 71% of the sample with a missing value for this variable on income measured at the household and individual level respectively. This compares with an over 99% reporting rate for RGSC.

If a social gradient exists in smoking behaviours, it may well be the case that excluding those who claimed to have ever smoked in 1985/6 does not fully remove the effect of any potential correlation between smoking and cancer: if individuals in lower social classes are more likely to smoke, they are also more likely to take up smoking after HALS was conducted, and consequently some of the apparent correlation with social class may be explained by differential smoking behaviours post-HALS. However, the probability of smoking take-up drops off precipitously after adolescence, and is almost zero by the time an individual reaches the age of 30 (Douglas and Hariharan, 1994). Further, dropping individuals who had ever smoked before 1985/6 may cause a masking of the initial selection effect into smoking: if individuals in lower social classes have some prior information that they are likely to die or develop cancer sooner and make decisions regarding smoking behaviours on this basis, some of the apparent effect of social class may be due to selection effects.

All right-hand-side variables other than social class are included merely as controls. Only social class is considered as a treatment variable: a variable on which (given the acceptance of assumptions discussed below) a causal effect is even postulated. There is clearly a complex web of interactions between education, social class and employment status and untangling these in a model such as this is no simple task, particularly when the

data available is a single cross-section (with subsequent follow-up only for cancer registration and death) such as this. At all times, models are estimated and conclusions are drawn under assumptions most likely to mitigate against a type I error on the treatment variables for social class. If the model was estimated, for instance, omitting measures of education, part of any relationship established between social class and health outcomes may be said to be due to social class merely proxying for education early in life. By including education in the model, any relationship between social class and cancer outcomes is made conditional on educational outcomes, and thus the relationship between social class and cancer outcomes must be considered necessarily a lower limit on the true relationship[5].

In addition to education (with a large reference case of individuals with no previous qualifications), controls are included for work status, age (through birth cohorts), gender, red meat consumption (based on individuals' self-report), and alcohol consumption (based on individuals' self-report for the last week's consumption). All of these RHS variables must be considered to be exogenous. Variables in the model measured at HALS are effectively assumed to be time-invariant: there is no way to establish how these variables subsequently (and, in most cases, previously) changed. In the case of social class, individuals' experience of being in a particular class, and the effect on their health of being in that class, is not constant over time, even if they remain in that same class, due to changes in, for instance, relative and absolute incomes over time. Being in social class V in 1985, for instance, does not necessarily have the same impact on health as being in social class V in 2010.

A variable for long-term unemployment (those unemployed for a period of one year or more) is included in the model to control only for the effect on individuals who may have been experiencing a substantial period of worklessness. While correlation between long-term unemployment and ill-health is well-established, evidence differs regarding the direction of causality. Gordo (2006) claims, accounting for endogeneity, that long-term unemployment has a significant and negative effect on the health of individuals (using German data), while Böckerman and Ilmakunnas (2009) (using Finnish data) conversely

[5]This paper takes no a priori conception of 'fairness', but it is perhaps worth noting that adopting such a method makes results robust to any claim that individuals should be held responsible for decisions regarding education taken early in life, and that social class is a 'fair' source of inequality.

suggest that individuals with poor health prospects are sorted into unemployment[6].

2.3 Theoretical framework

The model employed for this paper augments Grossman's original framework, removing the time constraint (with no loss of generality) and replacing the concept of individuals as consumer-producers of goods with the concept of them as mere consumers of two types of goods: those affecting health, and those not affecting health. While Grossman (1972) features an individual with a finite lifespan that ends (i.e, that the individual is registered as having developed cancer) at some future time period $t = n$ (the time period in which he dies), I introduce a second health outcome: that an individual has a finite cancer-free lifespan that ends at some future time period $t = m$ (the time period in which he is registered as having developed cancer). This implies that the individual will be observed to develop cancer if and only if $m \leq n$: that is, if death does not intervene before the individual develops cancer.[7] If individuals in lower social classes disproportionately die before they develop cancer (that is, if they disproportionately belong to the set of individuals for whom $m > n$), a comparison of lifetime incidence by social class will provide an incomplete picture of the burden of disease faced by individuals in lower social classes. This potential problem is overcome by modelling cancer-free lifespan, rather than lifetime incidence: this produces a dependent variable that treats cancer registration at an earlier age as different from cancer registration in old age.

The individual's utility optimisation problem is characterised as

$\max U = \sum_1^n U_t \left(h_t, Z_t, L_t\right) \quad t = 1...n$

where h_t is the individual's flow of health ('healthy days', in Grossman's original paper) in time period t, Z_t is a vector of consumption that does not affect health, and L_t is a vector of consumption that affects health.

The individual maximises this utility function subject to a number of constraints, as detailed below.

Individuals are assumed to invest in health (I_t) – take decisions that, positively or

[6]See Mathers and Schofield (1998) and Böckerman and Ilmakunnas (2009) for a review of the evidence on the relationship and possible direction of causation between unemployment and health.

[7]This weak inequality arises as the paper includes individuals whose cancer is detected only post-mortem. This information is derived from analysis of death certificate data, included in HALS.

negatively, affect their health – where this investment is a function of medical care (M_t), a vector of consumption goods affecting health (L_t), and a vector of characteristics which are assumed exogenous, such as the individual's level of education, his social class, and employment status (E_t).

$I_t = I_t\left(M_t, L_t; E_t\right)$

An individual's health in future time periods is a function of this investment, and health from the current time period (H_t), with some depreciation of this due to natural processes, where the depreciation rate is δ_t.

$H_{t+1} = I_t + \left(1 - \delta_t\right) H_t$

In some final time period, ($t = n$), the individual's health stock will fall below some level, $H_{\min}$, a minimum level of health stock required to survive.

$H_n < H_{\min}$

Further, these decisions are made subject to a constraint that discounted lifetime expenditure cannot exceed discounted lifetime income (where income in time period t is Y_t), plus some level of wealth held at time $t = 0$, termed A_0.

$\sum_{t=1}^{n} \frac{p_t M_t + q_t Z_t + s_t L_t}{(1+r)^t} = \sum_{t=1}^{n} \frac{Y_t}{(1+r)^t} + A_0$

where p_t is the price of medical treatment, q_t is a vector of prices of consumption goods that do not affect health, s_t is a vector of prices of consumption goods that do affect health, y_t is the individual's income and A_0 is an initial level of wealth held by the individual.

We also assume a positive correlation between n and m: that an earlier time of death is also associated with an earlier cancer registration. The result of this is that if social class is associated with lower levels of H_t, a pure measure of inequality in cancer could only be observed either among infinitely lived individuals, or those for whom $m \leq n$. If individuals in lower social classes see a reduction in both m and n, the effect on cancer lifetime incidence is ambiguous. At an extreme, assume that all individuals in lower social classes had a reduced time-to-cancer (should they live long enough) compared to individuals in higher social classes, but $m > n$ for individuals in lower social classes and $m \leq n$ for those in higher social classes. This would yield an incidence of 1 for cancer among individuals in higher social classes and an incidence of 0 among individuals in lower social classes, despite the reduced time-to-cancer had individuals in lower social

classes lived long enough to develop the disease. This motivates the use of a final model, detailed below, that treats a cancer registration at a young age as different from a cancer registration in old age.

2.4 Methods

2.4.1 Statistical models

This paper assesses the correlation between social class, and other lifestyle and socioeconomic variables, and life years from birth without cancer[8].

Figure 2.1 illustrates the basic possibilities for different types of individuals in the data. The horizontal axis represents time, with events to the left occurring before events to the right, and examples of subject types appearing on the vertical axis. Date of birth and dates of starting and quitting smoking were collected in the initial HALS survey, and date of death in subsequent follow-ups. Using this information, a solid line denotes years alive (survival time in the lifespan models), with a solid circle denoting birth and a solid cross denoting death (failure in the lifespan models). A non-filled square denotes cancer registration[9]. The dashed line beyond July 2009 represents the fact that these observations are right-censored at this point as such individuals' status as alive or dead (or registered cancer sufferers or not) is not known beyond this. Further discussion of this diagram follows, with a key provided within Figure 2.1.

[8] Analysis is carried out using Stata version 11.0.

[9] Note also, however, that some individuals are registered as cancer sufferers at the time of their death, and consequently some solid crosses also denote cancer registrations.

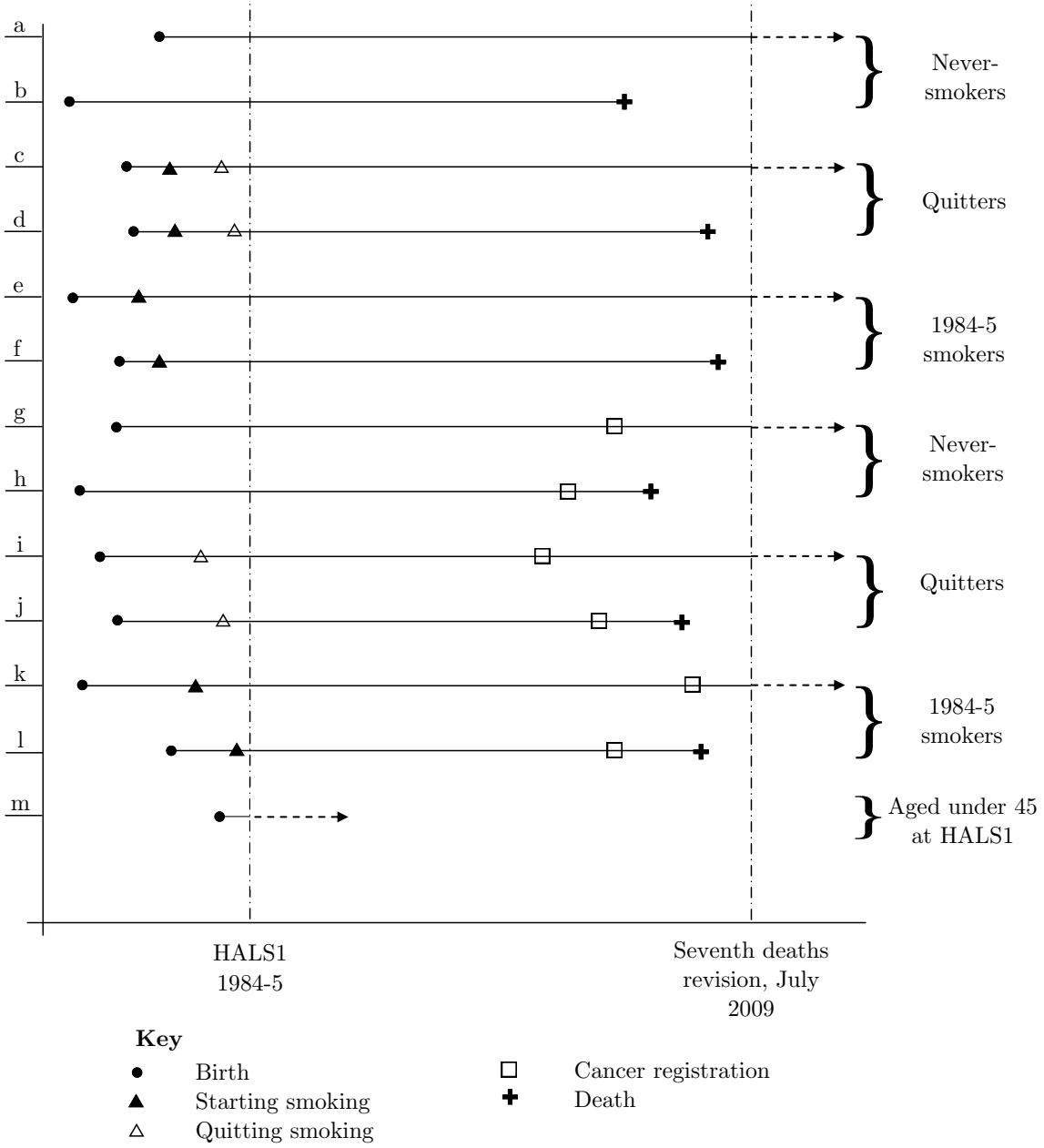

Figure 2.1: Types of individual

Cancer registration

Non-smokers only are included in this model (a, b, g, h in Figure 2.1), and are entered into the model conditional on survival at the time of HALS. While individuals can be, and indeed are, observed to have developed cancer before the survey began, individuals who had developed cancer before HALS are much more likely to have died before the survey took place. Those individuals with pre-existing cancer registrations are dropped from the sample[10]. Individuals who are registered as dead at the time of the most recent follow-up are checked for any appearance of a cancer on their death certificate. Such individuals are treated as failures in this model, with a failure time of their age at death. The dependent variable here is healthy time observed (*cancerage*): i.e. time before an individual is observed to have developed cancer. Individuals who have been registered as developing cancer at the time of the July 2009 HALS follow-up (g and h, of those here included) have a complete spell observed for this model while individuals who have never been registered as developing cancer at this time (a and b) are censored. Maximum likelihood estimation is used to estimate the link between various factors and the associated acceleration of time to failure.

This cancer registration model is clearly problematic in terms of interpretation. While cancer registration, if it occurs, must clearly precede death, death cannot precede cancer registration[11]. Consequently, individuals can be censored in this model for two reasons: that they are not registered as having developed cancer at the time of the follow-up (a), or that they have died without being registered as cancer or having cancer on their death certificate (b). These two types of censorings clearly differ. While survival (i.e., being alive and not registered as a cancer sufferer) at HALS is plausibly non-informative, death (particularly from certain causes) is potentially informative: for instance, cardiovascular disease and some cancers (such as lung cancer) share risk factors. Death from such diseases is therefore likely to be correlated with cancer registration: those dying from, for instance, CVD are more likely to, absent such a death, have developed cancer. The example of CVD is particularly pertinent given that smoking causes CVD with a relatively short lag

[10] Further, the inclusion of such individuals would lead to some individuals effectively having negative survival times in the left-truncated survival model used here.

[11] However, individuals can have a cancer registration age equal to their age at death, where cancer appears on the death certificate without the disease ever being previously diagnosed.

and lung cancer with a much longer lag (Cutler et al., 2006). As such, deaths are not accurately characterised as non-informative censorings but, where the cause of death is etiologically similar to cancers or the individual has innate susceptibilities to both the cause of death and cancers (Estève et al., 1994), death is likely to be correlated with the potential for cancer registration absent death. While a formal specification of the joint distributions of survival times for cancers and deaths is required to entirely eliminate any biases, such information is inherently unavailable (Estève et al., 1994; Honoré and Lleras-Muney, 2006). However, if individuals who die sooner are also disproportionately likely to develop cancer absent their death, and if being in a low socioeconomic group is correlated with both accelerated time to cancer and accelerated time to death, the coefficients for social class in this model provide (in absolute terms) lower bounds for the true coefficients.

2.5 Results

Distributions for the hazard of cancer registration (with all smokers excluded) are compared on Akaike Information Criteria and Bayesian Information Criteria scores, which take the loglikelihood generated through maximum likelihood estimation and impose a penalty for the introduction of additional parameters which more flexible distributions include[12]. Details for the final model are given in Table 2.3. A lognormal distribution is selected for this model – on both BIC and AIC score, it outperforms all other competing distributions [13]. Cox-Snell residuals are plotted in 2.2, and suggest that all these distributions fit the data well for most of the sample.

[12]For instance, the most flexible distribution compared here, the generalised gamma, includes three extra parameters than the least flexible, the exponential distribution. See, for instance, Cox et al. (2007) for more details on these nested distributions.

[13]Results are robust to the use of the more flexible generalised gamma distribution. Use of a semiparametric competing risks model (implemented through `stcrreg` in Stata), which seeks to explicitly model death as a competing risk in cancer registration, produces results that are qualitatively similar.

Model	Loglikelihood	degrees of freedom	AIC	BIC
Generalised gamma	-273.931	20	587.8621	692.7037
Lognormal	-274.4143	19	**586.8285**	**686.4281**
Loglogistic	-275.1758	19	588.3515	687.9511
Gompertz	-279.2357	19	596.4715	696.0711
Weibull	-276.6025	19	591.205	690.8046
Exponential	-299.1227	18	634.2455	728.603

Table 2.3: Comparison of AIC and BIC scores

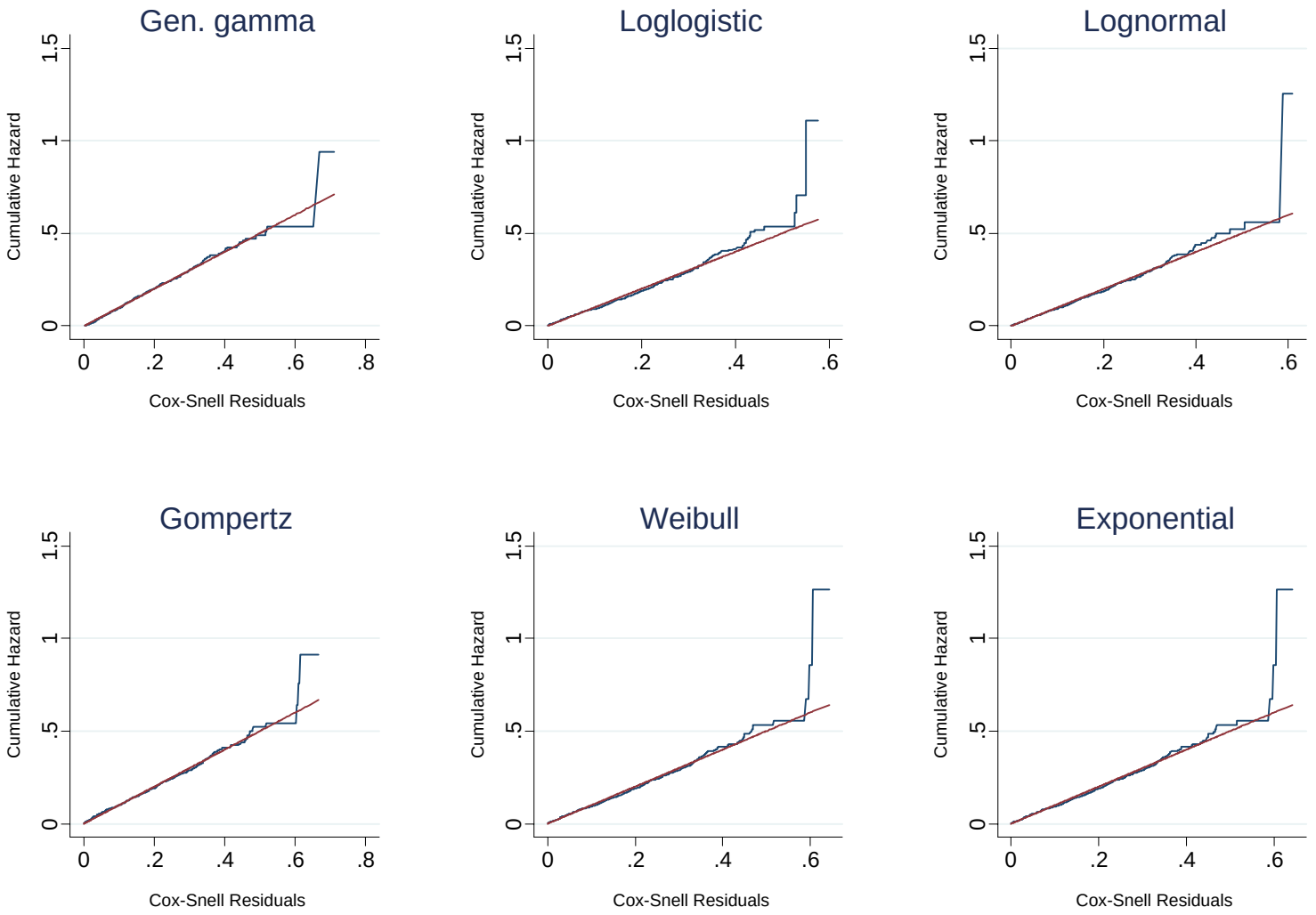

Figure 2.2: Cox-Snell residuals plots

Two preliminary models include smokers in the sample, for which results are shown in Tables 2.4 and 2.5. A social gradient – specifically, a significant difference in time to failure between class 1 and all other classes – in observed in cancer registrations when no account is taken of smoking behaviours (Table 2.4), but no such significant gradient is observed when these variables (dummy variables for smoking status, and variables for pack-years and pack-years squared) are included in the estimation, reflecting the disproportionate smoking exhibited among those in lower social classes[14]. This would seem to support the hypothesis that the social gradient in cancer is explained by differences in smoking behaviours. However, to exclude the possibilities (subject to the earlier clarification and assumption regarding smoking behaviours post-HALS) of differential smoking behaviours driving these results, or of selection into smoking based on an individual's own private information regarding his health expectations (as in, for instance, Balia and Jones (2011)), a final model – dropping all smokers – is estimated, and results presented in Table 2.6.

	Lognormal		Probit	
	Coef.	Std. err.	Coef.	Std. err.
NPAD	-0.071***	0.027	0.182**	0.079
redmeat3	0.01	0.015	-0.034	0.044
lhqdeg	0.097**	0.048	-0.270*	0.15
lhqoth	0.06	0.099	-0.271	0.288
lhqA	0.06	0.044	-0.11	0.131
lhqO	0.016	0.028	0.034	0.085
ltunemp	-0.134**	0.052	0.300**	0.153
sick	-0.047	0.041	-0.07	0.118
retd	0.028	0.024	-0.057	0.07
single	0.004	0.032	-0.085	0.09
sepdiv	0.003	0.034	-0.039	0.104
widowed	0.03	0.025	-0.140**	0.066
male	-0.052***	0.017	0.081*	0.048
sc23	-0.085*	0.047	0.162	0.141
sc45	-0.115**	0.049	0.182	0.145
bc20	0.034	0.026	0.006	0.067
bc3040	0.052	0.035	-0.282***	0.086
Constant	4.555***	0.055	-0.659***	0.164
ln(sigma)	-1.267***	0.053		
N. of cases	3800		3800	

* $p < 0.10, **p < 0.05, ***p < 0.01$

Table 2.4: Lognormal and probit regressions, smoking variables excluded

[14]For instance, 48% of those in the original HALS sample in unskilled occupations were regular smokers, compared to 22% in the highest social class.

	Lognormal		Probit	
	Coef.	Std. err.	Coef.	Std. err.
start	0.019	0.038	-0.025	0.120
packyrss	-0.006***	0.002	0.011**	0.005
packyrs2s	0.412**	0.206	-0.867	0.586
quit	-0.025	0.025	0.056	0.075
packyrsq	-0.001	0.001	0.000	0.003
packyrs2q	-0.023	0.083	0.134	0.215
NPAD	-0.046*	0.026	0.144*	0.080
redmeat3	0.015	0.015	-0.043	0.044
lhqdeg	0.076*	0.046	-0.235	0.15
lhqoth	0.057	0.095	-0.280	0.290
lhqA	0.057	0.042	-0.108	0.132
lhqO	0.010	0.027	0.045	0.086
ltunemp	-0.107**	0.050	0.251	0.154
sick	-0.028	0.040	-0.104	0.119
retd	0.025	0.023	-0.056	0.07
single	-0.004	0.030	-0.072	0.09
sepdiv	0.011	0.032	-0.056	0.104
widowed	0.029	0.024	-0.141**	0.066
male	-0.033**	0.017	0.049	0.051
sc23	-0.064	0.045	0.13	0.142
sc45	-0.083*	0.047	0.133	0.146
bc20	0.036	0.025	-0.016	0.068
bc3040	0.040	0.033	-0.293***	0.087
Constant	4.570***	0.053	-0.682***	0.165
ln(sigma)	-1.303***	0.052		
N. of cases	3800		3800	

* $p < 0.10, * * p < 0.05, * * *p < 0.01$

Table 2.5: Lognormal and probit regressions, smoking variables included

As we assume that smoking decisions may be affected by social class, all smokers are dropped from the sample in this final model and, in such a way, ignoring the effect of smoking altogether, a persistent gradient in social class is observed, as shown in Table 2.6, below.[15,16]

Socioeconomic inequalities appear to exist in time to cancer registration, even after the exclusion of all smokers from the sample. This gradient appears to be less obvious when the

[15]Note that, while the sign attached to many variables here is reversed between probit and lognormal distributions, this is to be expected. A positive coefficient in the probit model indicates an increased probability of the individual being registered as a cancer suffered over his or her lifespan, while a positive coefficient in the lognormal model indicates an *increase* in the expected time before the individual develops cancer.

[16]Only one individual was born in the 1940s in this model, and consequently bc40 is excluded from the probit model.

	Lognormal		Probit	
	Coef.	Std. err.	Coef.	Std. err.
NPAD	-0.078	0.060	0.328*	0.191
redmeat3	0.039	0.024	-0.113	0.076
lhqdeg	0.010	0.064	-0.091	0.213
lhqoth	0.011	0.148	-0.096	0.468
lhqA	0.009	0.069	0.024	0.221
lhqO	-0.003	0.043	-0.001	0.141
ltunemp	-0.035	0.133	-0.151	0.468
sick	-0.012	0.089	-0.137	0.28
retd	-0.026	0.04	0.081	0.125
single	-0.043	0.047	0.090	0.139
sepdiv	0.009	0.055	-0.092	0.181
widowed	0.030	0.037	-0.137	0.105
male	-0.044	0.028	0.113	0.088
sc23	-0.155**	0.074	0.428*	0.24
sc45	-0.157**	0.077	0.358	0.25
bc20	0.011	0.040	0.000	0.116
bc3040	-0.029	0.051	-0.170	0.148
Constant	4.709***	0.086	-1.066***	0.277
ln(sigma)	-1.301***	0.086		
N. of cases	1397		1397	

* $p < 0.10, * * p < 0.05, * * *p < 0.01$

Table 2.6: Lifetime incidence (lognormal) and healthy time before cancer (probit) models

model is estimated on the basis of lifetime incidence only, with differences appearing only significant at the 10% level when estimated in probit form, and no significant difference being shown to exist between social classes 4 or 5 and 1. Being in social class 2, 3, 4 or 5 is significant at the 5% level in reducing time to cancer by approximately 15%, with a slightly greater effect for social classes 4 and 5 than 2 and 3.

2.6 Conclusions

Many existing studies of cancer incidence find weak evidence of socieconomic gradients in cancer, considering lifetime incidence, once differential smoking behaviours between social classes are accounted for. Here, such incidence is compared with time to cancer, an innovation made possible by the format of HALS, incorporating a single cross-section in 1985 followed by future follow-up waves to provide data on cancer registrations. Individuals in lower social classes are found to develop cancer around 15% earlier than those in the highest social class, an inequality which is ill-evidenced when only lifetime incidence is considered. This points to the importance of considering age at diagnosis, in order to get a full picture of the differences in lifetime burden of disease posed by cancer across different social classes, a difference that is obscured by concentrating solely on lifetime incidence. Interpretation of coefficients in the cancer registration model is complicated by the way in which those who do not develop cancer are censored: (at least some) deaths are not non-informative censorings, but are symptomatic of the tendency of the individual to develop cancer in the absence of death. Under the assumptions discussed earlier, however, these coefficients can be treated as (in absolute terms) lower bounds on their true values in this regard.

Further, if the analysis of Link and Phelan (1995) is correct, the total effect of circumstantial factors such as social class, where correlated with particular lifestyles, is likely to be underestimated. According to this thesis, were people in lower social classes not to disproportionately adopt certain unhealthy lifestyles, for instance smoking, they may adopt others, and so the inherent link between the experience of being in a certain class and poor health outcomes may not be broken. Such dynamic effects are inevitably excluded from this model, and could only be identified using a dynamic analysis which looked at changing

behaviours over time, and potential substitutions out of, for instance, smoking into other risky behaviours. Furthermore, caution should be attached to interpreting these relationships as causal. While these results point to the importance of modelling pre-diagnosis cancer-free lifespan rather than simply lifetime incidence, limitations of the dataset – a single cross-section from almost 25 years prior to the most recent cancer follow-up – employed mean that caution should be attached to directly interpreting the magnitude of the estimated coefficients. A panel dataset with a long follow-up for cancer registration, which contained information regarding social class (or a more complete coverage of individuals' income) as well as information about early life circumstances would enable results to be drawn with more confidence in both their accuracy and causal nature.

References

Balia, S. and Jones, A. M. (2011), 'Catching the habit: a study of inequality of opportunity in smoking-related mortality', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**(1), 175–194.

Banks, J., Marmot, M., Oldfield, Z. and Smith, J. P. (2006), 'Disease and disadvantage in the United States and in England', *JAMA: The Journal of the American Medical Association* **295**(17), 2037–2045.

Böckerman, P. and Ilmakunnas, P. (2009), 'Unemployment and self-assessed health: evidence from panel data', *Health Economics* **18**(2), 161–179.

Cox, B., Huppert, F. and Whichelow, M., eds (1993), *The Health and Lifestyle Survey: Seven Years on: A Longitudinal Study of a Nationwide Sample, Measuring Changes in Physical and Mental Health, Attitudes and Lifestyle*, Dartmouth Pub. Co, Aldershot, Hants., England.

Cox, C., Chu, H., Schneider, M. F. and Muñoz, A. (2007), 'Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution', *Statistics in Medicine* **26**(23), 4352–4374.

Cutler, D., Deaton, A. and Lleras-Muney, A. (2006), 'The determinants of mortality', *The Journal of Economic Perspectives* **20**(3), 97–120.

Dalstra, J., Kunst, A., Borrell, C., Breeze, E., Cambois, E., Costa, G., Geurts, J., Lahelma, E., Van Oyen, H., Rasmussen, N., Regidor, E., Spadea, T. and Mackenbach, J. (2005), 'Socioeconomic differences in the prevalence of common chronic diseases: an overview of eight European countries', *International Journal of Epidemiology* **34**(2), 316 –326.

Deaton, A. (2002), 'Policy implications of the gradient of health and wealth', *Health Affairs* **21**(2), 13–30.

Department of Health and Human Services (1989), *Reducing the health consequences of smoking. 25 years of progress*, US Government Printing Office, Washington, DC.

Department of Health and Human Services (2004), 'Smoking and Tobacco Use; 2004 Surgeon General's Report', http://www.cdc.gov/tobacco/data_statistics/sgr/2004/index.htm.

Department of Health and Human Services (2010), 'A Report of the Surgeon General: How Tobacco Smoke Causes Disease', http://www.surgeongeneral.gov/library/tobaccosmoke/report/index.html.

Doll, R. (1998), 'Uncovering the effects of smoking: historical perspective', *Statistical Methods in Medical Research* **7**(2), 87–117.

Doll, R. and Hill, A. B. (1954), 'The mortality of doctors in relation to their smoking habits: a preliminary report', *BMJ* **328**(7455), 1529–1533.

Douglas, S. and Hariharan, G. (1994), 'The hazard of starting smoking: Estimates from a split population duration model', *Journal of Health Economics* **13**(2), 213–230.

Estève, J., Benhamou, E. and Raymond, L. (1994), 'Statistical methods in cancer research: descriptive epidemiology.', *IARC scientific publications* **IV**(128), 1.

Gordo, L. R. (2006), 'Effects of short- and long-term unemployment on health satisfaction: evidence from German data', *Applied Economics* **38**(20), 2335.

Grossman, M. (1972), 'On the concept of health capital and the demand for health', *The Journal of Political Economy* **80**(2), 223–255.

Hiatt, R. A. and Breen, N. (2008), 'The social determinants of cancer: A challenge for transdisciplinary science', *American Journal of Preventive Medicine* **35**(2, Supplement 1), S141–S150.

Honoré, B. E. and Lleras-Muney, A. (2006), 'Bounds in competing risks models and the war on cancer', *Econometrica* **74**(6), 1675–1698.

Jones, A. M., Rice, N., Bago d'Uva, T. and Balia, S. (2007), *Applied Health Economics*, Routledge advanced texts in economics and finance, Routledge, London.

Link, B. G. and Phelan, J. (1995), 'Social conditions as fundamental causes of disease', *Journal of Health and Social Behavior* **35**, 80–94.

Lyratzopoulos, G., Abel, G. A., Brown, C. H., Rous, B. A., Vernon, S. A., Roland, M. and Greenberg, D. C. (2012), 'Socio-demographic inequalities in stage of cancer diagnosis: evidence from patients with female breast, lung, colon, rectal, prostate, renal, bladder, melanoma, ovarian and endometrial cancer', *Annals of Oncology* p. mds526.

Mackillop, W. J., Zhang-Salomons, J., Boyd, C. J. and Groome, P. A. (2000), 'Associations between community income and cancer incidence in Canada and the United States', *Cancer* **89**(4), 901–912.

Mao, Y., Hu, J., Ugnat, A. M., Semenciw, R., Fincham, S. and Canadian Cancer Registries Epidemiology Research Group (2001), 'Socioeconomic status and lung cancer risk in Canada', *International Journal of Epidemiology* **30**(4), 809–817.

Marmot, M., Rose, G., Shipley, M. and Hamilton, P. J. (1978), 'Employment grade and coronary heart disease in British civil servants', *Journal of Epidemiology and Community Health* **32**(4), 244–249.

Marmot, M., Smith, G. D., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E. and Feeney, A. (1991), 'Health inequalities among British civil servants: the Whitehall II study', *The Lancet* **337**(8754), 1387–1393.

Mathers, C. D. and Schofield, D. J. (1998), 'The health consequences of unemployment: the evidence.', *The Medical Journal of Australia* **168**(4), 178.

Nkosi, T. M., Parent, M.-l., Siemiatycki, J. and Rousseau, M.-C. (2012), 'Socioeconomic position and lung cancer risk: how important is the modeling of smoking?', *Epidemiology (Cambridge, Mass.)* **23**(3), 377–385.

Peto, R., Lopez, A., Boreham, J. and Thun, M. (2006), 'Mortality from smoking in developed countries, 1950-2000'.

Sidorchuk, A., Agardh, E. E., Aremu, O., Hallqvist, J., Allebeck, P. and Moradi, T. (2009), 'Socioeconomic differences in lung cancer incidence: a systematic review and meta-analysis', *Cancer Causes & Control* **20**(4), 459–471.

Singh, G. K., Miller, B. A., Hankey, B. F., Edwards, B. K., Singh, G. K., Miller, B. A., Hankey, B. F. and Edwards, B. K. (2003), 'Area socioeconomic variations in US cancer incidence, mortality, stage, treatment, and survival, 1975-1999'.

University of Cambridge Clinical School (2009), 'HALS deaths and cancer working manual, june 2009'.

US National Cancer Institute (2006), 'Cancer trends progress report - stage at diagnosis'.

Vallejo-Torres, L. and Morris, S. (2010), 'The contribution of smoking and obesity to income-related inequalities in health in England', *Social Science & Medicine* **71**(6).

Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y. T., Gupta, P. C., Hackshaw, A., Matos, E., Samet, J., Sitas, F., Smith, J., Stayner, L., Straif, K., Thun, M. J., Wichmann, H. E., Wu, A. H., Zaridze, D., Peto, R. and Doll, R. (2004), 'Tobacco and cancer: Recent epidemiological evidence', *Journal of the National Cancer Institute* **96**(2), 99 –106.

Wilkinson, R. and Pickett, K. (2010), *The Spirit Level: Why Equality Is Better for Everyone*, Penguin, London.

Chapter 3

A Discrete Latent Factor Model for Smoking, Cancer and Mortality.

Daniel Howdon* [1,2] & Andrew M. Jones[1]

1. Department of Economics and Related Studies, University of York, YO10 5DD, UK
2. Centre for Health Economics, University of York, YO10 5DD, UK

Tel.: +44 1904 32 1411.
**E-mail address:* daniel.howdon@york.ac.uk

Abstract

This paper investigates the relationship between smoking and ill-health, with a focus on cancer outcomes. A discrete latent factor model for smoking and health outcomes, allowing for these to be commonly affected by unobserved factors, is jointly estimated, using the British Health and Lifestyle Survey (HALS) dataset. Post-estimation predictions suggest the reduction in time-to-cancer to be 5.7 years for those with a smoking exposure of 30 pack-years, compared to never-smokers. Estimation of posterior probabilities for class membership show that individuals in certain classes exhibit similar observables but highly divergent health outcomes, suggesting that unobserved factors in this model substantially determine these outcomes. The use of a joint model changes the results substantially. The results show that failure to account for unobserved heterogeneity leads to differences in survival times between those with different smoking exposures to be overestimated by more than 50% (males, with 30 pack-years of exposure).

JEL codes: C41; I14.

Keywords: health; health inequality; duration analysis; smoking; cancer; mortality; determinants of health; lifestyles.

Acknowledgements: Data from the HALS were supplied by the Economic and Social Data Service. Information on date of interview was supplied by Brian Cox of the University of Cambridge Clinical School. The authors thank Nigel Rice for helpful comments. This research was funded by the Economic and Social Research Council under the Large Grant Scheme (RES-060-25-0045).

3.1 Introduction

This paper develops a joint model of smoking, mortality and cancer, with a particular focus on the timing of the onset of cancer. The model is estimated with data from the British Health and Lifestyle Survey (HALS) from 1984-85, linked to the most recent follow-ups on mortality and cancer registration from July 2009. It features joint estimation of the decisions of individuals to start smoking, their age of starting, the pack-years of smoking exposure, time-to-cancer registration and age of death to analyse the relationship between individual lifestyles, socioeconomic circumstances and cancer. The model accounts for the possibility of common observable and unobservable factors that influence both smoking and the health outcomes.

The model brings together two approaches to modelling health and lifestyles, using the HALS dataset. In the first approach, Contoyannis and Jones (2004) specified an economic model of health production and lifestyle choices from which they derived an empirical specification that is estimated as a recursive model for a set of binary measures of health outcomes and health-related behaviours, including smoking. Common unobservable factors are assumed to have a multivariate normal distribution and the model is estimated as a multivariate probit. There is evidence from this model of a statistically significant correlation between unobservables that influence smoking and that influence the health outcomes, indicating selection bias. Estimates from the multivariate model show that being a non-smoker in 1984, along with sleeping well and taking exercise, are associated with a higher probability of reporting excellent or good self-assessed health in 1991, with non-smoking increasing the probability by 0.15. Contoyannis and Jones (2004) also find that a large proportion of the impact of lifestyles on socioeconomic inequality in health is masked if the unobserved heterogeneity is ignored. Balia and Jones (2008) extended the multivariate model by adding a binary indicator for deaths that had occurred by the time of the May 2003 longitudinal follow-up of the HALS deaths data. They find that being a non-smoker in 1984 is associated with a 0.22 lower probability of dying by 2003. Their decomposition analysis of a Gini coefficient for mortality suggests that lifestyle factors contribute strongly to inequality in mortality, reducing the direct role of socioeconomic status. They also reinforce the finding that ignoring unobserved heterogeneity leads to an

under-estimate of the contribution of lifestyle to socioeconomic inequality, showing that this applies to mortality as well as self-assessed health.

A second strand of models, initiated in Forster and Jones (2001), focuses on richer measures of the timing of decisions about smoking and estimates hazard functions for starting and quitting smoking. Balia and Jones (2011) developed this approach by estimating a recursive system of equations for starting smoking, the age of starting, the number of years smoked and age of death, with data from the April 2005 deaths follow-up. The equations in their model are tied together and estimated as a system by allowing for common unobservables that are modelled as discrete latent factors, following the approaches of Heckman and Singer (1984) and Mroz (1999). In line with the epidemiological literature such as Doll et al. (2004), they find a difference of about 12 years in median survival between current and never smokers and about 3.6 years between current and former smokers.

This paper takes the analysis of HALS a step further. By adding new cancer registration data and deaths data, from July 2009, we extend the model to add a duration model for the onset of cancer. In addition, intensity of smoking is captured by a measure of pack-years that augments data on the number of years smoked with a measure of the quantity of cigarettes consumed. Results derived using the joint modelling approach employed in this paper exhibit differences in the implied predicted survival function for cancer, suggesting a role for unobserved heterogeneity in explaining cancer outcomes. This is further illustrated by the estimation of posterior probabilities for each individual's class membership: large differences in health outcomes are exhibited between individuals in different latent classes, despite similar observable characteristics. Post-estimation prediction of median survival times shows the reduction in time to cancer to be 5.7 (5.8) years for men (women) who were smokers at the time of HALS, with a total observed exposure of 30 pack-years, compared to never-smokers at the time of HALS.

3.2 Background

The link between smoking and ill-health in general, and many specific diseases, is well-established. It is estimated that men born in the first 30 years of the 20th Century who took up smoking cigarettes, and did not stop, suffered a reduction of 10 years in

their lifespan, with smoking cessation at the age of 40 associated with an increased life expectancy of 9 years over those who continued to smoke (Doll et al., 2004). The risks of smoking have been well-explored since the link between smoking and lung cancer was made by Doll and Hill (1954). Smoking has been associated with a greater propensity to develop various cancers and other diseases (for example, deaths from lung cancer are estimated to occur with between 10.8 and 24.9 times the frequency in smokers as in non-smokers (Doll, 1998)) and is estimated to be responsible for approximately 30% of all cancer deaths in developed countries, as well as causing deaths from respiratory, circulatory and other problems (Department of Health and Human Services, 1989; Jones et al., 2007; Peto et al., 2006; Vineis et al., 2004). Vallejo-Torres and Morris (2010) estimate that 2.3% of all socioeconomic inequality in health between 1998 and 2006 was due to smoking. Successive reports by the US Surgeon General (Department of Health and Human Services, 1989, 2004, 2010) have examined the evidence linking smoking with mortality and diseases including cancer, making stronger causal links over time, with 30 diseases listed in the 2004 report for which evidence was 'sufficient to infer a causal relationship'. Doll (1998) provides a useful summary of the history of evidence regarding the (causal) links between smoking and ill-health.

One of the most influential studies into the effects of smoking on health is the British Doctors Study (see Doll and Hill (1954) and subsequent papers), a prospective cohort study with longitudinal follow-ups. Although vital in establishing the link between smoking and ill-health, studies based on this dataset necessarily focused solely on one small stratum of society – 34,494 male doctors working in Britain – and, as such, cannot inform research into the existence or otherwise of social gradients in health. Questions regarding smoking status sought to establish whether the doctor had ever smoked (one cigarette per day, for one year or more), whether he was a current smoker, the age at which he began to smoke and the amount that he was currently smoking[1]. While this is not an area where evidence from randomised trials is available, other, much smaller-scale, studies have since been carried out using innovative methods to confirm the causal relationship, such as following pairs of smoking and non-smoking twins to track health outcomes in order to control for

[1] In contrast to, for instance, the HALS dataset, which asked for an average number of cigarettes smoked over the period during which the individual (had) smoked.

possible genetic factors that predispose individuals to both smoking and disease (Kaprio and Koskenvuo, 1989).

3.3 Data

This paper uses baseline data from the British Health and Lifestyle Survey 1 (HALS1), conducted between 1984 and 1985, which sought to examine the relationships of lifestyle, behaviours (such as smoking and alcohol consumption) and circumstances of a large cross-section of a representative sample of individuals in Great Britain (Cox et al., 1993). Data collection consisted of a one-hour face-to-face interview to collect information on individuals' lifestyles, a visit from a nurse to collect information on physiological and cognitive function, and a self-completed questionnaire to gather information regarding psychiatric health and personality (Cox et al., 1993; Jones et al., 2007). Details of individuals' diagnoses with cancer and information relating to individuals' deaths (such as date and cause of death) were subsequently provided to the HALS team. Such data, including details from death certificates and cancer diagnoses are available to the beginning of July 2009 – the Seventh Deaths Revision and Fourth Cancer Revision (University of Cambridge Clinical School, 2009). 9,003 individuals were initially entered into the study of whom, as of this revision, the statuses of 97.8% have been flagged on the NHS's Central Register at the Office for National Statistics. As of this revision, 2,883 individuals have been flagged as dead and 1,468 coded for cancer.

Data was cleaned up to remove inconsistencies, and missing values for those variables included in the model. Further, individuals were excluded where they had been diagnosed with cancer prior to the initial HALS1 survey. While the exclusion of those living with cancer in 1985 does mean that the sample is necessarily less representative of the population, this avoids the problem of the inclusion of such individuals with a negative time-to-cancer.

It must be borne in mind that there were delays involved in this registration of deaths and developing cancer, and that these delays were not uniform in all cases. The latest HALS follow-up manual suggests that cancer registrations tend to be slower to reach the Central Register than death notifications (although such registrations are probably com-

plete up to the end of 2007), and that missing cases will exist due to patchy returns from regional registries (University of Cambridge Clinical School, 2009). A spike is recorded in more recent years (with 14 such cases in 2008 and 2009, more than in the previous 13 years combined) for individuals who died with cancer present without ever being registered as developing such a disease (Table A3.2 (Appendix)), suggesting that some late returns may exist for this revision[2]. Comparison of the previous HALS follow-up (to April 2005) with data held in this latest follow-up shows, however, that no cancer registrations were late – i.e. were included in the July 2009 follow-up with a date of April 2005 or earlier – but that 7 death registrations were late by this measure. Furthermore, the age at the time of an individual's first cancer registration is not the same as the age of the individual first developing cancer. Diagnosis of cancer does not immediately take place upon the individual developing the disease, nor does it occur at the same stage of development of the cancer across different individuals, or over time. In particular, the stage at diagnosis has varied over time, with US National Cancer Institute (2006) showing declines in the rates of late-stage diagnoses of cases of cancers of the cervix, colon, prostate and rectum between 1980 and 2006.

There is censoring of smoking variables at the time of the survey, with no follow-up made on smoking habits. Consequently, for instance, an individual who is recorded as having quit at the time of HALS1 may take up smoking again, or an individual recorded as a current smoker at the time of HALS1 may quit soon after. The value for years spent smoking only considers the known years of smoking at the time of HALS1. Further, and similarly, circumstantial variables in the model such as social class and marital status, and lifestyle variables such as alcohol consumption and time spent exercising are effectively assumed to be time-invariant: there is no way to observe how these variables changed over time. The reliability of the HALS1 data further is enhanced by accurate recall and reporting of individuals' smoking habits: evidence on this suggests that, while smoking status is generally recalled accurately, the number of cigarettes smoked per day over time is frequently recalled with some error, with relatively poorer recall for ex-smokers (Krall et al., 1989), potentially introducing bias at the point of data collection.

[2]This data is obtained using the Stata icd9 command to search for individuals whose death certificate shows any cancer (codes in the range 140 to 239.99).

A further challenge posed by the possibility of unobservable hetereogeneity is the potential for the introduction of bias in that individuals can only appear in the HALS dataset if they were alive at the time of HALS1. While observables may suggest a balanced sample, this dataset may reflect the omission of certain groups who differ in important unobservable characteristics. In particular, individuals who would have been of age to be included in HALS1 and who had smoked are more likely to have died before HALS1 took place. While this sample may, for instance, show a representative sample of smokers in the UK at the time of HALS1, if individuals select into smoking based on their life expectancy, HALS1 may exclude frailer or less frail individuals (depending on the joint distribution of underlying frailty and the effect of smoking on the health of such individuals). While the number of smokers may be representative, therefore, the makeup of these smokers in terms of their unobserved frailty, may not. As in Chapter 2, only individuals aged 45 or over at the time of HALS1 are included in this model, to reduce the confounding of mortality and cancer registrations with genetic factors unrelated to the covariates used in the health outcome models, and to ensure that as full a spell of smoking as possible is observed for individuals in the sample. The existence of this unobserved heterogeneity motivates the use of the joint model, detailed below.

3.4 Theoretical model

As in the previous chapter, a second health outcome of cancer is introduced, and the individual is said to develop cancer at some time period m. The individual is assumed to maximise a utility function which is explicitly conceived of in terms of optimal choices around health-related goods, and two types of consumption goods: those not affecting health, and those affecting health. We assume that individuals' choices around consumption goods that affect health are affected by the individual's exogenously-determined, but unobserved to the researcher, health constraints: his initial (genetic, or early-life) stock of health, and the individual's discount rate at any future time period. This means that both health outcomes (cancer and lifespan) and decisions regarding consumption goods that affect health (in our model, decisions regarding smoking) are jointly affected by these unobserved factors. This motivates the use of a joint model which allows the impact of

such unobservable factors to be estimated and recovered.

The individual's utility optimisation problem is characterised as

$\max U = \sum_1^n U_t\left(h_t, Z_t, L_t\right) \quad t = 1...n$

where h_t is the individual's flow of health ('healthy days', in Grossman's original paper) in time period t, Z_t is a vector of consumption that does not affect health, and L_t is a vector of consumption that affects health.

The individual maximises this utility function subject to a number of constraints, as detailed below.

Individual are assumed to invest in health (I_t) – take decisions that, positively or negatively, affect their health – where this investment is a function of medical care (M_t), a vector of consumption goods affecting health (L_t), and a vector of exogenous characteristics, such as the individual's level of education (E_t).

$I_t = I_t\left(M_t, L_t; E_t\right)$

An individual's health in future time periods is a function of this investment, and health from the current time period (H_t), with some depreciation of this due to natural processes, where the depreciation rate is δ_t.

$H_{t+1} = I_t + \left(1 - \delta_t\right) H_t$

In some final time period, ($t = n$), the individual's health stock will fall below some level, $H_{\min}$, a minimum level of health stock required to survive.

$H_n < H_{\min}$

Further, these decisions are made subject to a constraint that discounted lifetime expenditure cannot exceed discounted lifetime income (where income in time period t is Y_t), plus some level of wealth held at time $t = 0$, termed A_0.

$\sum_{t=1}^n \frac{p_t M_t + q_t Z_t + s_t L_t}{(1+r)^t} = \sum_{t=1}^n \frac{Y_t}{(1+r)^t} + A_0$

where p_t is the price of medical treatment, q_t is a vector of prices of consumption goods that do not affect health, s_t is a vector of prices of consumption goods that do affect health, y_t is the individual's income and A_0 is an initial level of wealth held by the individual. For the purposes of this model, a solution in terms of the individual's optimal smoking behaviour alone is of relevance.

We also assume a positive correlation between n and m: that an earlier time of death is also associated with an earlier cancer registration.

If the individual's depreciation rate δ_t is greater, the model predicts a lower optimal level of health stock, H_t in each time period ($t \neq 0$) and also accelerates the point at which $H_n \leq H_{\min}$: the point at which the individual dies. Furthermore, a lower level of inherited health, H_0 will cause a lower level of optimal health in future time periods, and also accelerates the point at which $H_n < H_{\min}$.

Individuals' preferences for consumption affecting health are affected by the inherited stock of health (H_0) and current and future depreciation rates, as well as a vector of factors assumed exogenous, such as social class, education, and public health actions affecting individuals' consumption of goods affecting health (SC_t).

$L_t = L_t\left(H_0, \delta_t(t = 0...n); SC_t\right)$

Similarly, in an innovation employed here, different expected future depreciation rates and different levels of inherited health stock are permitted to affect the individual's optimal demand not only for health but also for L_t, lifestyles affecting health. Owing to potential problems of endogeneity, we directly model individual's lifestyles with respect to smoking decisions, and allow decisions regarding smoking and lifestyle to be correlated with each other in a flexible framework which allows different levels of (for instance) expected depreciation rates and inherited health stock to be positively or negatively associated, through their effect on health outcomes (for the purposes of this paper, expected lifespan and expected time-to-cancer), with decisions regarding smoking.

3.5 Methods

3.5.1 The model

A system of five equations, including a binary outcome of whether an individual ever smoked, as well as duration models for starting smoking, quitting smoking, mortality, and cancer registration, is estimated. This extends the approach of Balia and Jones (2011), who estimate similar models, but without cancer registration, for an earlier HALS follow-up. The model adopts a discrete latent factor approach for dealing with the effect of unobserved heterogeneity in systems of equations Heckman and Singer (1984) and Mroz (1999). Deaths data are included to allow for the competing risk of mortality in the model for cancer, and also to make use of all information regarding future health outcomes that

may be considered by individuals as they make decisions regarding their smoking.

This section outlines each of the components of the overall loglikelihood function for the model, which includes contributions for the probability of ever-smoking and the hazards for age of starting smoking, pack-years exposure to smoking, age of onset of cancer and age at death. These contributions are bound together by the latent factor specification of unobserved heterogeneity in the joint likelihood function.

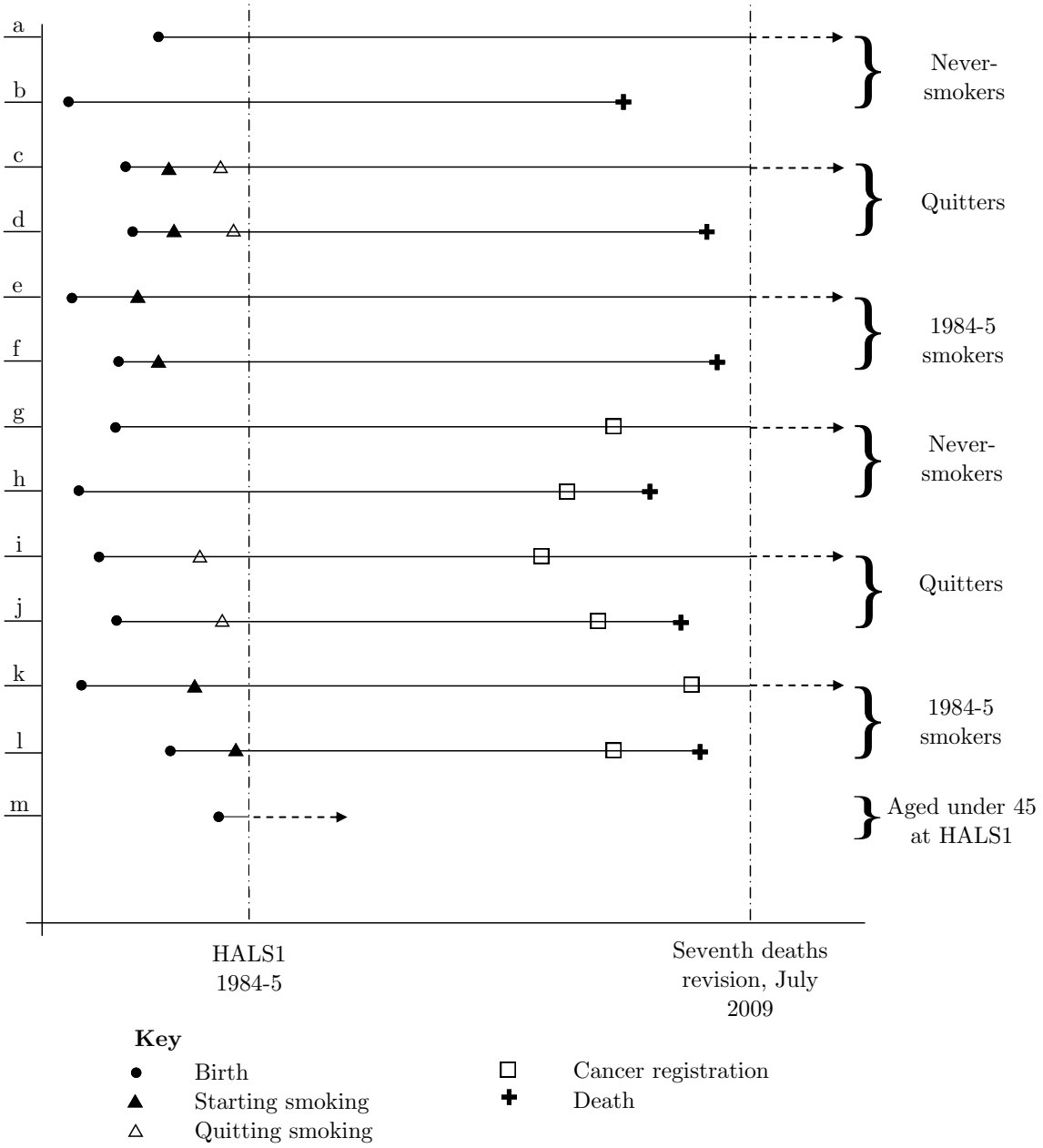

Figure 3.1: Types of observed outcomes

Figure 3.1 illustrates the basic possibilities for observed durations for different types of individual. The horizontal axis represents time, with events to the left occurring before events to the right, and examples of subject types appear on the vertical axis. Date of birth and dates of starting and quitting smoking were collected in the initial HALS1 survey, and date of death in subsequent follow-ups. Using this information, a solid line denotes known years alive (survival time in the lifespan model), with a solid circle denoting birth, a hollow square denoting cancer registration (failure in the cancer registration model), and a cross denoting death (failure in the lifespan model). The dashed line beyond July 2009 represents the fact that these observations are right-censored at this point as such individuals' status as alive or dead (or registered cancer sufferers or not) is not known beyond this. Individuals of type m are not included in the sample due to being aged under 45 at the time of HALS1. Individuals of type n also do not appear in HALS (and are not used in this analysis), due to their having died prior to HALS1.

Starting smoking

Individuals become 'at risk' in this model at the time of their birth, as indicated by the solid circle. Given that, in this sample, individuals are (due to exclusions) aged at least 45, with a mean age of 60, they are likely to have started to smoke if they were ever to smoke. The dependent variable in the duration model is years observed without starting smoking. A solid triangle on the diagram indicates that an individual is recorded to have started to smoke before HALS1 (failure in this model). Such individuals (c to f and i to l in Figure 3.1) score 1 on the *ever_smoker* variable. This is modeled by a probit model with loglikelihood contribution[3]:

$$l_1 = \ln\left(\Phi\left(\omega_1\right)\right)$$

where:

$$\omega_1 = \beta_1' x_1 + \varphi_1$$

and φ_1 is an individual-specific intercept term, reflecting unobserved individual characteristics that influence the probability of ever smoking.

[3]This split population approach to modelling the initiation of smoking follows Douglas and Hariharan (1994); Forster and Jones (2001) and Balia and Jones (2011).

Those who started smoking are also used in the starting duration model (in which all are failures) and all contribute to the loglikelihood with their logged loglogistic density function[4]:

$$l_2 = -\ln\left(1+(\omega_2 t_1)^{1/\gamma_1}\right) + \left(\frac{1}{\gamma_1}-1\right)\ln\omega_2 + \left(\frac{1}{\gamma_1}-1\right)\ln t_1 - \ln\gamma_1 - ln\left(1+\omega_2 t_1^{1/\gamma_1}\right)$$

where:

$$\omega_2 = \exp\left(-\left[\beta_2' x_2 + \varphi_2\right]\right)$$

and φ_2 is again individual-specific intercept term, reflecting unobserved individual characteristics that influence the age at starting to smoke.
t_1 is time to censoring or failure, and γ_1 is the loglogistic duration dependence parameter. Individuals who are not observed to start smoking before HALS1 (a, b, g and h in Figure 3.1) score 0 on the ever_smoker variable, enter the probit model and provide loglikelihood contribution:

$$l_1 = \ln\left(\Phi\left(-\omega_1\right)\right)$$

These individuals are not used in the duration model for starting smoking.

Exposure to smoking

Only those who scored 1 on the *ever_smoker* variable (those who had ever smoked, i.e. types c to f and i to l in Figure 3.1) contribute to the likelihood function for this part of the model. The dependent variable here is not time spent smoking (*smoke_years*), but total exposure to smoking before quitting (for individuals with a complete spell) or before HALS1 (for individuals whose observations are censored). In Figure 3.1, *smoke_years* is denoted by the length of the solid line between the solid triangle, denoting starting smoking, and either the hollow triangle, denoting quitting, or the point at which HALS1 was conducted. The dependent variable, *pack_years*, is *smoke_years* multiplied by individuals' self-reported average number of packs of (20) cigarettes smoked per day (*n_cigs*/20), giving a more complete picture of total exposure to smoking. Individuals who are observed

[4]Hazard functions for each duration model are selected according to statistical criteria to find the best-fitting parametric distribution. See the Appendix.

to quit before HALS1 (c, d, i and j in Figure 3.1) have a "complete spell" for this function and individuals who are observed as current smokers (e, f, k and l in Figure 3.1) at HALS1 are censored observations. The overall contribution of each individual to the loglikelihood is the logged Gompertz likelihood function,

$$l_3 = q \cdot (\ln(\omega_3) + \gamma_2 t_2) - \frac{\omega_3}{\gamma_2}(\exp(\gamma_2 t_2) - 1)$$

where q denotes an individual has quit smoking, t_2 is time to failure or censoring,

$$\omega_3 = \exp\left(-\left[\beta_3' x_3 + \varphi_3\right]\right)$$

and γ_2 is the Gompertz shape parameter.

Age of death

All individuals are included in this model, and are entered into the model conditional on survival at the time of HALS1[5]: individuals are only 'at risk' from this time onwards as they cannot be observed to have died before the point at which the survey is completed. The dependent variable here is time observed alive (*lifespan*). In Figure 3.1, lifespan is denoted by the distance between the solid circle, denoting birth, and either a cross, denoting death, or the point at which the July 2009 follow-up was conducted. Individuals whose death has been reported at the time of the HALS follow-up in July 2009 (b, d, f, h, j and l) have a complete spell for this outcome and individuals whose death has not been reported (a, c, e, g, i and k) are censored at this time. The overall contribution to the loglikelihood is the logged left-truncated Weibull likelihood function:

$$l_4 = d \cdot (\ln(\omega_4) + \ln(\alpha) + (\alpha - 1)\ln(t_3)) - \omega_4(t_3^\alpha + t_0^\alpha)$$

where t_0 is the age of the individual at HALS1, d denotes whether an individual has died:

$$\omega_4 = \exp\left(\beta_4' x_4 + \varphi_4\right)$$

[5]Additional data that are not included in the original HALS1 dataset provided by the Economic and Social Data Service, regarding the date of the initial interview was provided by Brian Cox and merged into the HALS1 dataset, matching by serial number. This allows greater accuracy in the measurement of *smoke_years*.

and α is the Weibull shape parameter.

Cancer registration

All individuals are included in this model, and are entered into the model conditional on survival at the time of HALS1. While the intuition behind this is not as straightforward as that in the mortality model (individuals can be, and indeed are, observed to have developed cancer before the survey began), individuals who had developed cancer before HALS1 are much more likely to have died before the survey took place. Those 147 individuals with pre-existing cancer registrations are dropped from the sample: the inclusion of such individuals would lead to some negative survival times in the left-truncated survival model. Individuals who are registered as dead at the time of the most recent follow-up are checked for any mention of a cancer on their death certificate. Such individuals are treated as failures in this model, with a failure time of their age at death. The dependent variable here is healthy time observed (*cancer_age*): i.e. time before an individual is observed to have developed cancer. Individuals who have been registered as developing cancer at the time of the July 2009 HALS follow-up (g to l in Figure 3.1), or who have a cancer included on their death certificate, have a complete spell observed for this model (the distance from birth to cancer registration, denoted by a hollow square) while individuals who have never been registered as developing cancer at this time (a to f) are censored. The overall contribution to the loglikelihood is the logged left-truncated loglogistic likelihood function:

$$\begin{aligned} l_5 = &\ln\left(1+(\omega_5 t_0)^{1/\gamma_4}\right) - \\ &\ln\left(1+(\omega_5 t_4)^{1/\gamma_4} + c\left[\frac{1}{\gamma_4}\ln\omega_5 + \left(\frac{1}{\gamma_4}-1\right)\ln t_4 - \ln\gamma_4 - \ln\left(1+\omega_5 t_4^{\frac{1}{\gamma_4}}\right)\right]\right) \end{aligned} \tag{3.1}$$

where

$$\omega_5 = \exp\left(-\left[\beta_5' x_5 + \varphi_5\right]\right)$$

t_0 is again the age of the individual at HALS1, t_5 is time to censoring or failure, and γ_4 is the loglogistic duration dependence parameter.

The cancer registration model is clearly more problematic than the mortality model in terms of interpretation. While cancer registration, if it occurs, must clearly precede death,

death cannot precede cancer registration[6]. Consequently, individuals can be censored in this model for two reasons: that they are not registered as having developed cancer at the time of the follow-up (a, c and e), or that they have died without developing cancer (b, d and f). These two types of censorings clearly differ. While survival (i.e., being alive and not registered as a cancer sufferer) at HALS1 is plausibly non-informative, death (particularly from certain causes) is not: for instance, cardiovascular disease and some cancers (such as lung cancer) share risk factors. Death from such diseases is therefore likely to be correlated with cancer registration; those dying from, for instance, CVD are likely to, absent such a death, have developed cancer. The example of CVD is particularly pertinent given that smoking causes CVD with a relatively short lag and lung cancer with a much longer lag (Cutler et al., 2006). As such, deaths are not accurately characterised as non-informative censorings but, where the cause of death is etiologically similar to cancers or the individual has innate susceptibilities to both the cause of death and cancers (Estève et al., 1994), death is likely to be correlated with the potential for cancer registration absent death. Although the model employed does allow for four latent classes of individuals to exist, each of which could potentially have the same or opposing directional effects on lifespan and time-to-cancer, a formal specification of the joint distributions of survival times for cancers and deaths is required to entirely eliminate any biases. Such information is, however, inherently unavailable (Estève et al., 1994; Honoré and Lleras-Muney, 2006).

3.5.2 Joint likelihood

While some of the potential effect of unobservable heterogeneity is muted by including only those aged over 45 at the time of HALS1 (the most frail individuals being those likely to die earliest (Gutierrez, 2002)), as discussed in Contoyannis and Jones (2004), Balia and Jones (2008, 2011) and Adda and Lechene (2013) unobservable heterogeneity poses potential problems for any analysis. If unobservable heterogeneity exists and is ignored, estimated coefficients may be biased. With particular regard to the effect of smoking, this includes factors which affect life expectancy – such as underlying congenital and hereditary conditions leaving individuals prone to early death – and also affect, for

[6] Although, as discussed, individuals can have a cancer registration age equal to their age at death, where cancer appears on the death certificate without the disease ever being previously diagnosed.

instance, the decision to smoke.

Individuals with lower prior life expectancies may select disproportionately into smoking due to the relatively low opportunity cost of smoking in terms of life years foregone, an effect which is potentially greater if the individual also considers morbidity as a future health outcome (Contoyannis and Jones, 2004; Balia and Jones, 2011)[7]. Alternatively, frailer individuals may disproportionately fail to select into smoking as the marginal value of additional good health is greater for such people. Adda and Lechene (2013) present evidence suggesting that the former, even when factors such as social class are controlled for, more accurately characterises smoking behaviour: individuals with lower life expectancies disproportionately take up smoking, smoke more cigarettes and are less likely to quit than those with longer life expectancies. Contoyannis and Jones (2004), however, present evidence suggesting that frailer individuals select out of smoking and are more likely to quit sooner. In either case, the consequence is that smoking behaviours are potentially endogenous in health outcomes. Further, the probability of starting smoking may be endogenous in both the time at which an individual starts and the total pack-years exposure of the individual, and the age at starting smoking may be endogenous in the total exposure to smoking.

The joint model is estimated by using a latent factor specification for the joint distribution of the random intercepts in each equation, $\varphi_1 \dots \varphi_5$, where $\varphi_j = \tau_j u + \rho_j v$ $(j = 1, \dots, 5)$, u and v are discrete factors, and τ and ρ are the factor loadings.

Mixing probabilities, π_k, representing the proportions of the sample composing each latent class, are recovered via estimation of the joint probabilities of observing combinations of the Bernoulli random variables u and v, taking a value 1 with probability θ_1 and θ_2 respectively. These probabilities are given a logistic form:

$$\theta_p = \frac{e^{\zeta_p}}{1 + e^{\zeta_p}} \quad (p = 1, 2)$$

and are recovered by estimation of the parameters, ζ_p. The structure of the latent factor

[7]While this model does allow individuals to make decisions based on any information regarding their future probability of developing cancer, individuals are likely to have less private information regarding this than regarding future mortality. Hereditary or congenital factors affecting an individual's chance of developing cancer are less common: only a small proportion (5-10%) of cancers are attributable to genetic defects, with the remainder attributable to environment and lifestyle (Anand et al., 2008).

model is summarised in Table 3.1.

Mass point, k	u	v	φ_j
1	0	0	0
2	1	0	τ_j
3	0	1	ρ_j
4	1	1	ν_j

Table 3.1: Mass points: 4 points of support

When combined, the final total likelihood function is:

$$L = \sum_{k=1}^{4} \pi_k \left(\exp l_{1,k}\right) \left(\exp l_{2,k}\right) \left(\exp l_{3,k}\right) \left(\exp l_{4,k}\right) \left(\exp l_{5,k}\right)$$

Further assumptions are required to identify the distribution of latent factors. Mass points at 0 and 1 (i.e. where $u = v = 1$ and $\tau + \rho = \nu$) are fixed by Balia and Jones (2011), and the same approach is employed here. While, as argued by Balia and Jones (2011), the model should in principle be identified by the non-linear form of each equation with no need for exclusion restrictions, in order to aid identification, the full model is estimated using three procedures. Each equation in the model is first singly estimated, using the preferred baseline hazard function according to AIC and BIC scores[8]. The derived parameter estimates from this stage are used as starting values (along with postulated approximate latent class parameters) in a second model, which estimates the full model with various parameter restrictions[9]. All of these estimates, including the estimated latent factor parameters, are used as starting values to estimate the final model, without parameter restrictions. Various different parameter restrictions in the initial stages are employed, and the final results are found to be robust to changes to these.

Where possible the generalised gamma, Gompertz, Weibull, lognormal and loglogistic distributions are compared for each duration equation. Gompertz and Weibull distributions are commonly used in duration analysis of human mortality (see, for example, Wilson (1994) who finds, using 1988 US Census data, that Weibull, Gompertz and loglogistic

[8]See the Appendix.

[9]The effect of each latent class parameter is, for example, initially postulated to be the in the same direction for cancer and lifespan. Where $\beta_{\text{variable},j}$ denotes the coefficient estimate for the given variable in equation j, the restrictions invoked are: $\beta_{\text{sc12},5} = \beta_{\text{sc12},4}$; $\rho_4 = -4\rho_5$; $\tau_4 = -4\tau_5$; $\tau_1 = -1.1\rho_1$. Different combinations of these restrictions are invoked, with no effect on the final parameters derived.

distributions provided good fits in simple models of human mortality). The generalised gamma distribution is compared, where possible, with the other forms of the baseline hazard, but given its heavy computational demands, particularly within the context of a jointly-modelled system of five equations such as this, estimation is not always possible[10]. In addition to these commonly-used distributions, the expopower distribution (Saha and Hilton, 1997), a flexible parametric distribution, nesting the exponential, Weibull and log-normal distributions is also compared. While a bathtub-shaped hazard is less plausible given the exclusion of all individuals aged under 45 at the time of HALS1, some cancers (such as testicular cancer) are more likely to occur earlier in life and, as such, it is useful to include such a distribution which allows for this while also remaining less computationally-intensive than, for example, the generalised gamma distribution. A comparison of BIC and AIC scores for all of these distributions is presented in the Appendix.

3.5.3 Key covariates and interpretation of parameters

Summary statistics for the variables used in the analysis are presented in Table 3.2:

Table 3.2: Summary statistics (all 3784 observations)

label	description	mean	std dev	min	max
mothm	Mother smoked, male child	0.02	0.14	0.00	1.00
mothf	Mother smoked, female child	0.02	0.15	0.00	1.00
fathm	Father smoked, male child	0.28	0.45	0.00	1.00
fathf	Father smoked, female child	0.32	0.47	0.00	1.00
bothm	Both parents smoked, male child	0.09	0.28	0.00	1.00
bothf	Both parents smoked, female child	0.11	0.31	0.00	1.00
othersmok	Other smokers in house	0.33	0.47	0.00	1.00
rural	Lives in the countryside	0.21	0.41	0.00	1.00
suburb	Lives in a surburban area	0.46	0.50	0.00	1.00
strtpostdoll	Started smoking after 1954 (first Doll et al BMJ article) but before 1971	0.04	0.20	0.00	1.00
strtpostpubhealth	Started smoking after 1971 (first smoking public health campaign)	0.00	0.05	0.00	1.00
starting	Number of years non-smoking	34.08	22.48	4.00	96.00
smoke_years	Years of smoking exposure	21.77	19.98	0.00	72.00
n_cigs	Average number of cigarettes smoked per day	10.41	12.46	0.00	97.00

Continued on next page

[10] In fact, the generalized gamma is not preferred by AIC or BIC scores for any of the single-equation models for which it provides parameter estimates. While it nests many of the other distributions, the expopower distribution (which also nests the Weibull and log distribution) often outperforms it even on its loglikelihood score.

cancer_dc	Registered as cancer sufferer or cancer on death certificate	0.27	0.44	0.00	1.00
cancer_age	Age of cancer registration or age of censoring (July 2009)	77.22	9.01	47.20	115.23
death	Dead	0.58	0.49	0.00	1.00
lifespan	Observed lifespan: censoring at July 2009	78.33	8.61	48.50	115.23
smoker	Ever-smoker	0.63	0.48	0.00	1.00
start	Smoker	0.31	0.46	0.00	1.00
quit	Ex-smoker	0.33	0.47	0.00	1.00
pack_years	Pack-years of exposure	18.50	24.11	0.00	236.00
pack_years_quit	Pack-years (HALS1 quitter)	8.68	20.50	0.00	236.00
pack_years_quit2	Pack-years squared / 10000 (HALS1 quitter)	0.05	0.23	0.00	5.57
pack_yearss	Pack-years (HALS1 current smoker)	9.82	18.21	0.00	138.00
pack_years_start2	Pack-years squared / 10000 (HALS1 current smoker)	0.04	0.12	0.00	1.90
NPAD	Heavy alcohol drinker	0.09	0.29	0.00	1.00
redmeat3	Eats red meat 3+ times per week	0.52	0.50	0.00	1.00
recex	At least 5 hours of exercise in last two weeks	0.09	0.28	0.00	1.00
lhqdeg	Highest qualification is degree	0.03	0.17	0.00	1.00
lhqoth	Other highest qualification	0.01	0.08	0.00	1.00
lhqA	Highest qualification is A-Level	0.03	0.17	0.00	1.00
lhqO	Highest qualification is O-level/CSE	0.07	0.26	0.00	1.00
lhqhnd	Highest qualification is HND/HNC	0.02	0.13	0.00	1.00
ltunemp	Long term unemployed	0.02	0.14	0.00	1.00
sick	Not working due to permanent sickness/disability	0.04	0.19	0.00	1.00
retd	Retired	0.43	0.49	0.00	1.00
male	Male	0.45	0.50	0.00	1.00
sc23	Social class 2 or 3	0.66	0.48	0.00	1.00
sc45	Social class 4 or 5	0.32	0.46	0.00	1.00
single	Single	0.07	0.25	0.00	1.00
sepdiv	Separated/Divorced	0.05	0.22	0.00	1.00
widowed	Widowed	0.16	0.37	0.00	1.00

In the health outcomes equations, *pack_years* (and its squared term) is interacted with being a current smoker, and separately with being an ex-smoker. These variables are separated to mark those individuals for whom *smoke_years* is complete rather than right-censored at the time of HALS1: smoking status is unknown beyond the point at which such data was collected[11]. The separation of current smokers and quitters is useful due to

[11]Examination of the HALS2 dataset, a follow-up on the original sample seven years later in which similar data was again collected, reveals that – of those in the sample here whose smoking status could be ascertained – 27% of those who were current regular smokers at HALS1 had quit smoking by the time of this survey in 1991-1992. It must be noted that, however, over 45% of regular smokers at HALS1 were

the fact that risk of death for certain cancers, such as lung cancer, has been found to be elevated for ever-smokers over never-smokers for a period of up to 20 years, but declines with time after quitting smoking (Reid et al., 2006).

While the identification of the parameter estimates of coefficients of the various pack-years variables seems clear, interpretation of these coefficients is not as straightforward. Due to the censoring of the smoking duration variables at the time of HALS1, this does not represent the elevated hazard (or acceleration of time to failure) of exposure to one additional pack-year of smoking. This coefficient represents the association of an increase of one pack-year of observed smoking on the increased hazard of failure, conditional on smoking status in 1985. While this model could be estimated using smoking status at HALS1 (i.e. whether an individual is a current smoker, quitter, or has never smoked) as the only smoking-related regressors, this would seem to discard useful information: that some individuals smoke for longer and with greater intensity than others.

Balia and Jones (2011) model the influence of parental smoking but do not allow for different relationships for male and female offspring. Here, parental smoking is interacted with gender to investigate any differential result of effects of different parents smoking on different genders of children. Brown and van der Pol (2010) suggest that, at least for mothers and daughters, the intergenerational transfer of risk and time preference explains a significant part of the correlation between smoking outcomes.

In addition to variables regarding smoking status[12], another key lifestyle variable, a dummy variable for heavy consumption of alcohol, is included in the model. This is defined as those drinking over 20 units per week[13] – the NHS describe alcohol consumption over this level as 'high' [14]. While moderate consumption of alcohol may be protective against some diseases (Doll et al., 1994, 2005), evidence suggests up to 40% higher all-cause mortality for heavy consumers (Doll et al., 1994)[15].

missing for this variable at HALS2.

[12]With smoking take-up defined as ever having smoked on average at least one cigarette per day, for a period of at least six months (Cox et al., 1987).

[13]The mean consumption of alcohol by those in the sample recorded as drinking over 20 units per week is 38 units.

[14]See, for instance, `http://www.nhs.uk/Conditions/Alcohol-misuse/Pages/Treatment.aspx` Results are robust to a definition of heavy alcohol consumption that defines this as 14 units for women and 21 units for men.

[15]Doll et al. (1994) group the heaviest consumers of alcohol as those drinking 43 or more units per week.

As well as alcohol consumption, a variable for individuals' exercising habits is included in the lifespan model. This exercise dummy is derived from a composite measure of hours of exercise spent in the last two weeks, *tothrsex*, created from HALS data for total time spent involved in: keep fit exercises, cycling, golf, jogging, swimming, table tennis, basketball, football, rugby, badminton, tennis, squash, fives, rackets, cricket, windsurfing, sailing, self-defence, boxing, wrestling, backpacking, hiking and dancing[16]. Individuals who exercised for more than 5 hours in the previous two weeks are classed as having exercised for the recommended period of time in this model[17]. Further, consumption of red meat (*redmeat3*, defined as consuming red meat at least three times per week), linked to colorectal cancer, the second most common form of the disease (Cutler, 2008), is included in the cancer registration model.

Variables for socioeconomic class, based on the then-prevalent Registrar-General's classification by occupation, are also included in the model. The existence of socio-economic gradients in health is well-established (Marmot, 2007; Thomas et al., 2010; Wilkinson, 1996; Wilkinson and Pickett, 2010), with the socio-economic gradient in smoking explaining part of this (Schaap and Kunst, 2009). Such inequality in health outcomes is potentially of greatest concern where equality of opportunity in society is considered to be the appropriate goal. One useful model of this allows for some variation in health to be due to effort and some to be due to circumstances (Roemer, 1998; Rosa Dias, 2009). While strong evidence exists regarding a social gradient in lifespan overall and illnesses such as cardiovascular disease, the existence of a social gradient in cancer is more controversial. Deaton (2002) argues that the Whitehall Studies (Marmot et al., 1978, 1991) show no social gradient in any cancer apart from lung cancer, the gradient in which is entirely explained by differential smoking behaviours between the occupational grades. Despite finding social gradients in health overall and in many diseases, Wilkinson and Pickett (2010) find no social gradient in breast cancer, and 'only small class differences' in prostate cancer. Further, much attention has focused on incidence of cancer rather than time-to-cancer (for instance, Singh et al. (2003); Banks et al. (2006); Dalstra et al. (2005)).

[16] While this is likely to only partially capture a measure of exercise conducted by individuals, it is serves as a limited, but useful, proxy.

[17] The NHS recommend that adults exercise for 30 minutes, five times a week. More details are available at `http://www.nhs.uk/Livewell/fitness/Pages/Howmuchactivity.aspx`

A variable for long-term unemployment (those unemployed for a period of one year or more) is included in the model to exclude individuals who may have been suffering from only a short spell of worklessness. While correlation between long-term unemployment and ill-health is well-established, evidence differs regarding the direction of causality. Gordo (2006) claims that, accounting for endogeneity, long-term unemployment has a significant and negative effect on the health of individuals (using German data), while Böckerman and Ilmakunnas (2009) (using Finnish data) conversely suggest that individuals with poor health prospects are sorted into unemployment[18].

3.6 Results

Five equations are jointly estimated: a probit model for smoking initiation, and duration models for time before smoking initiation (for smokers only), pack-years of exposure to smoking (for smokers only), time until death (conditional on being alive and cancer free at HALS1) and time until developing cancer (conditional on being alive and cancer free at HALS1).

The Appendix presents AIC and BIC scores for the single equations estimates of the full range of survival distributions that could be estimated for each outcome: age of starting, exposure before quitting, age of cancer registration, and age of death. Those models with the best AIC and BIC scores are italicised. Accordingly, a loglogistic baseline hazard function is chosen for starting smoking, Gompertz for smoking exposure, Weibull for mortality, and loglogistic for cancer registration.

Full results for the parameter estimates from the five equation DLFM are provided in Tables 3.3 and 3.4. Table 3.3 shows the coefficients associated with the covariates and Table 3.4 shows the factor loading and probabilities of class membership for the latent factor model. Single-equation estimates for the cancer registration model are provided, for comparison, in Table 3.5.

The interpretation of a coefficient – and the sign (positive or negative) of a coefficient – depends upon the type of model being interpreted. A key is provided to this effect, in a row labelled as 'Model type'. A positive (negative) coefficient for the probit model

[18]See Mathers and Schofield (1998) and Böckerman and Ilmakunnas (2009) for a review of the evidence on the relationship and possible direction of causation between unemployment and health.

for starting smoking has the usual interpretation: an increased probability of a positive (negative) result, i.e. of starting smoking. The remaining four models are duration models, which fall into two categories: accelerated failure time (AFT) or proportional hazard (PH). A positive (negative) coefficient in an accelerated failure time type of model implies an increase (decrease) in time before 'failure', the point at which the event of interest occurs. A positive (negative) coefficient in a proportional hazard type of model implies an elevation (reduction) in the hazard of 'failure', the occurrence of the event of interest. More specifically, the exponent of a coefficient in a PH model gives the multiplicative effect on the hazard of failure (the 'hazard ratio'), while the exponent of a coefficient in an AFT model gives the multiplicative effect on the expected lifespan (the 'time ratio').

This means that the intuitive interpretations of positive and negative coefficients are reversed, depending upon the type of model being employed. For instance, a positive coefficient in the lifespan model implies a higher instantaneous probability of death, conditional upon survival to the current time period (consistent with a *decrease* in median survival time, relative to the reference case), and a positive coefficient in the cancer model implies an increased length of expected cancer-free time (consistent with an *increase* in median survival time, relative to the reference case).

Variable	smoker	starting	pack-years	lifespan	cancer
Model type	Probit	AFT	PH	PH	AFT
Mother smoked, male child	0.562***	-0.008			
Mother smoked, female child	0.557***	-0.047			
Father smoked, male child	0.472***	-0.041*			
Father smoked, female child	0.280***	-0.057**			
Both parents smoked, male child	0.523***	-0.048*			
Both parents smoked, female child	0.682***	-0.105***			
Social class 2 or 3	0.305**	-0.087***	-0.263	0.128	-0.030
Social class 4 or 5	0.538***	-0.126***	-0.413**	0.394**	-0.046*
Highest qualification is degree	-0.438***	0.067**	0.139	-0.530**	0.054*
Other highest qualification	-0.244	0.059	-0.290	0.006	0.061
Highest qualification is A-Level	0.100	0.045*	0.177	-0.342	0.026
Highest qualification is O-level/CSE	-0.169**	0.046**	-0.019	0.011	-0.000
Highest qualification is HND/HNC	-0.264	0.117***	0.017	-0.269	0.010
Male	0.649***	-0.201***	-0.079	0.442***	-0.026***
Born in 1920s	0.302***	-0.050***	-0.019	-0.008	-0.023*
Born in 1930s	-0.001	-0.085***	0.175	-0.046	-0.045***
Born in 1940s	-0.070	-0.265***	0.052	0.267	-0.016
Started smoking after 1954 but before 1971		0.347***			
Started smoking after 1971		0.921***			
Number of years non-smoking			0.049***		
Other smokers in house			-0.752***	0.109	-0.015
Long term unemployed			-0.479**	0.548**	-0.061**
Not working due to permanent sickness/disability			-0.364**	0.785***	-0.025
Retired			0.113	-0.136	0.023*
Single			-0.187	0.257**	0.015
Separated/Divorced			-0.729***	-0.027	-0.005
Widowed			-0.394***	0.078	0.017
Lives in the countryside			0.256***	-0.091	-0.004
Lives in a surburban area			0.130*	-0.041	-0.003
Pack-years (HALS1 quitter)				0.014***	-0.001**
Pack-years squared /10000 (HALS1 quitter)				-0.557**	0.027
Pack-years (HALS1 current smoker)				0.037***	-0.003***
Pack-years squared / 10000 (HALS1 current smoker)				-3.032***	0.243***
Heavy alcohol drinker				0.184*	-0.022
At least 5 hours of exercise in last two weeks				-0.402***	
Eats red meat 3+ times per week				-0.123**	0.011
Constant	-0.706***	3.126***	-4.786***	-56.533***	4.718***
γ		0.141***	0.008***		0.065***
α				12.327***	
N. of cases			3784		

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.3: DLFM results – main coefficients

	Latent class, k			
	1	2	3	4
φ_1	0	0.287**	-0.144	0.143
φ_2	0	-0.075***	0.010	-0.065***
φ_3	0	-0.463***	0.577**	-0.114
φ_4	0	2.356***	1.341***	3.697***
φ_5	0	-0.276***	-0.211***	-0.487***
π_k	0.353***	0.443***	0.090***	0.113***

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.4: DLFM results (2) – latent factor coefficients and class membership probabilities

Variable	Coefficient
Pack-years (HALS1 quitter)	-0.002**
Pack-years squared / 10000 (HALS1 quitter)	0.025
Pack-years (HALS1 current smoker)	-0.005***
Pack-years squared / 10000 (HALS1 current smoker)	0.320**
Other smokers in house	-0.013
Heavy alcohol drinker	-0.028
Eats red meat 3+ times per week	0.015
Highest qualification is degree	0.073
Other highest qualification	0.050
Highest qualification is A-Level	0.050
Highest qualification is O-level/CSE	0.004
Highest qualification is HND/HNC	0.016
Long term unemployed	-0.096**
Not working due to permanent sickness/disability	-0.025
Retired	0.023
Male	-0.036**
Social class 2 or 3	-0.054
Social class 4 or 5	-0.073
Single	-0.003
Separated/Divorced	0.004
Widowed	0.027
Lives in the countryside	-0.019
Lives in a surburban area	-0.001
Born in 1920s	0.031
Born in 1930s	0.043
Born in 1940s	0.187
Constant	4.559***
γ	0.164***
N. of cases	3784

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.5: Single equation - cancer registration

Relative to the benchmark of latent class 1 (35% of the sample), latent classes 2 (44% of the sample) and 4 (11% of the sample) consist of individuals who are more likely to start smoking, start earlier in life, smoke more cigarettes after starting, die sooner, and get cancer earlier in life. Latent class 3 (9% of the sample) consists of individuals who are less likely to start smoking, start later in life, smoke fewer cigarettes if they do start, but die sooner and get cancer earlier in life.

Different relationships between parental smoking and individuals' smoking behaviours are observed according to the gender of the parent and the gender of the offspring. The relationship with the probability of starting smoking of one of either a mother or father smoking on the offspring is found to be greater on men than women. The correlation with the probability of smoking of the offspring is found to be greater for a mother who smokes than for a father. The relationship with time to starting is greater for women than men. While these results are broadly in line with those of Balia and Jones (2011), a major difference lies in the large divergence observed between the relationships according to the genders of parents and children. Further, while Balia and Jones (2011) find a cohort effect for those born subsequent to the publication of the first evidence showing a link between smoking and ill-health in 1954, a much larger deceleration in time to starting smoking is observed (over the cohort born between 1954 and the first public health campaign) for the cohort born after the first anti-smoking public health campaign in 1972.

Parental smoking has little direct relationship with total exposure to smoking (the dependent variable in equation three) conditional on starting smoking. Those in social class 4 or 5, and those with other smokers in the house, are observed to have a significantly lower hazard of quitting.

As would be expected, additional exposure to smoking increases the hazard of death, with a stronger relationship observed for current smokers than for quitters, and a declining relationship with total exposure on the increase in hazard (as shown by the opposing coefficient on the squared terms). The interpretation of these coefficients is complicated by the censoring of durations of current smokers at HALS1 (as well as the lack of data regarding whether quitters ever started smoking again, and, if so, for how long). Social class continues to be correlated, independent of lifestyle choices, with an elevation in the hazard of death for those in social class 4 or 5 roughly equivalent to that of an exposure

of approximately 12 observed pack-years (for HALS1's current smokers) at the time of HALS1, compared to those in social class 1[19].

Results on cancer registration differ somewhat. Being male, and being long-term unemployed at HALS1 are significantly related with reducing time to failure in this model. Evidence of a social gradient in cancer is found – with those in social class 4 or 5 having a significantly shorter (by approximately 5%) predicted healthy time before developing cancer than those in the highest social class – even after accounting for the effect of disproportionate smoking among those in a lower social class, and before accounting for the effect of reduced lifespans in preventing the observation of cancer registrations among those who would, had they not died, have been more prone to suffer from such a disease. This is equivalent to an exposure to smoking of approximately 19 pack-years[20]. One crucial problem with the HALS follow-up dataset, which could lead to the underestimation of the social gradient in cancer, is the number of individuals (107) who die with cancer present (according to death certificate data) but without ever being registered as suffering from the disease, suggesting a disproportionate failure to diagnose (and, presumably, therefore, to treat) those in lower social classes.

3.6.1 Posterior probabilities

Individuals are here sorted into the most likely latent class to which they belong, based on their observed outcomes. This means, for each class k and individual i:

$$P_{ki} = \frac{\pi_k \cdot L_{ki}}{\sum_{l=1}^{4} \pi_l \cdot L_{li}}$$

.

Sorting individuals into their most likely class based on these posterior probabilities – that is, assigning each individual i to class k for which P_{ki} is highest – results in Figure 3.6 are obtained:

Those individuals most likely to be part of class 1 are highly unlikely to ever develop cancer: only 2% of individuals most likely to be in class 1 are observed to have developed

[19] $12\beta_{pack\_years\_start} - 12^2\,(\beta_{pack\_years\_start2}/10000) \approx \beta_{sc45}$.

[20] This is calculated using the same method as in footnote 16. However, caution should be attached to this, given that smoking and social class are likely to affect both time-to-cancer and lifespan.

Class	1	2	3	4
n	1247	1968	101	468
HALS1 age	60.22	61.52	59.77	58.38
Social class 1	0.02	0.02	0.02	0.02
Social class 2/3	0.65	0.66	0.65	0.67
Social class 4/5	0.32	0.31	0.32	0.30
Ever-smoker	0.69	0.55	1.00	0.71
Smoker at HALS1	0.40	0.20	0.85	0.37
Quitter at HALS1	0.29	0.36	0.15	0.34
Pack-years of exposure (ever-smokers only)	31.74	24.31	59.60	30.26
Developed cancer	0.02	0.31	0.59	0.71
Age of cancer (developed cancer)	87.14	76.94	70.95	65.30
Lifespan (dead only)	88.19	79.76	71.82	66.83

Table 3.6: Descriptive statistics, by most probable latent class based on posterior probabilities.

cancer, despite this class being made up of individuals with approximately similar smoking characteristics and social class, and of similar ages, to those most likely to be members of class 4, of which 71% of individuals are observed to have developed cancer by July 2009. Furthermore, differences in observed lifespan are striking, with a difference of over 20 years between individuals in class 1 and class 4. This points to unobservable factors which explain large elevations in an individual's hazard of suffering cancer and early death, even when such individuals are in the same social class and adopt similar lifestyles.

3.7 Counterfactual simulations

This section presents counterfactual predictions of survival times – healthy years without cancer. This is done by amending the observed values for all individuals' smoking behaviours and holding other individual characteristics (and the estimated coefficients associated with these characteristics) constant, in a post-estimation analysis.

Survival probabilities are estimated for each of the $k(k = 1, \ldots, 4)$ latent classes, using the loglogistic survival function:

$$S_k(t) = \left(1 + \left[t \cdot \exp\left(-\beta\prime X_{cf} + \varphi_k\right)\right]^{(1/\gamma)}\right)^{-1}$$

where X_{cf} refers to the counterfactual values for variables. These probabilities are multiplied by the associated prior probability of class membership. These products are summed

to calculate a survival function for the full distribution:

$$S(t) = \sum_{k=1}^{4} \pi_k \cdot S_k(t)$$

Results for median survival times to onset of cancer, with men and women considered separately, are presented in Table 3.7, with estimated median survival curves presented in Figures 3.2 to 3.4.

	Male		**Female**	
	Estimated survival time	*Difference from full sample*	Estimated survival time	*Difference from full sample*
Full sample	85.0	–	88.7	–
Counterfactuals				
Non-smoker	87.8	*+2.8*	90.4	*+1.7*
20 pack-years	83.5	*-1.5*	86.1	*-2.6*
30 pack-years	82.1	*-2.9*	84.6	*-4.1*

Table 3.7: Counterfactual estimates – median survival time to onset of cancer (years)

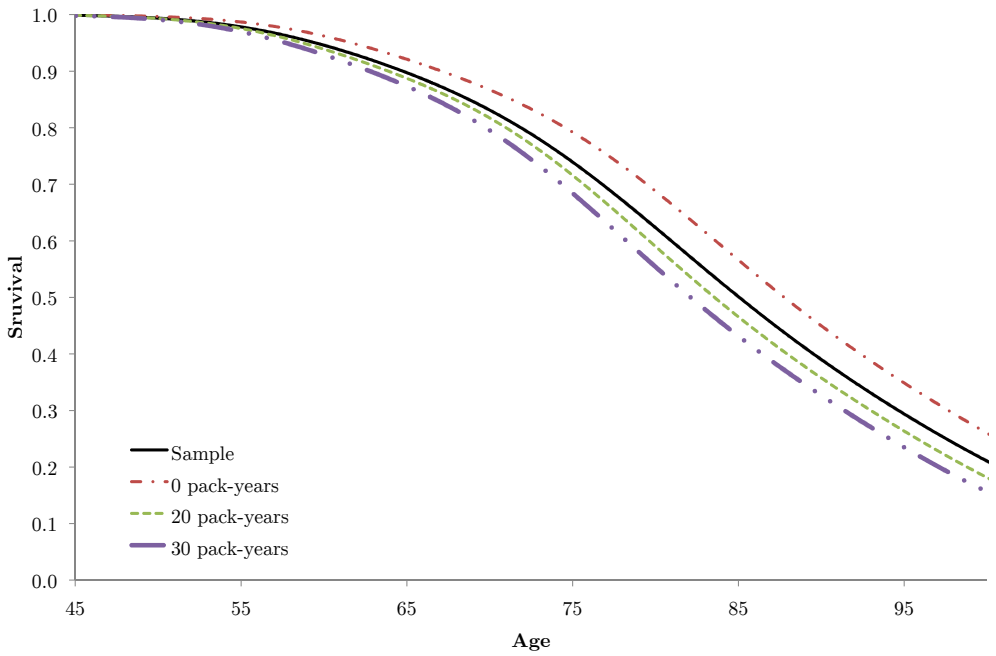

Figure 3.2: Estimated survival curves for cancer onset by smoking behaviour (males)

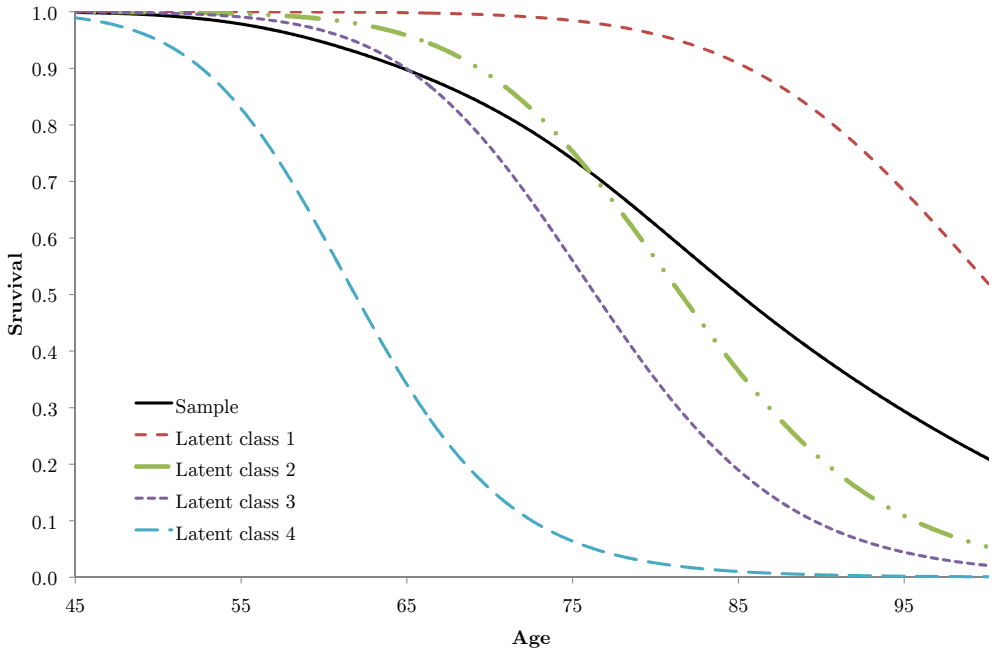

Figure 3.3: Estimated survival curves for cancer onset by latent class (males)

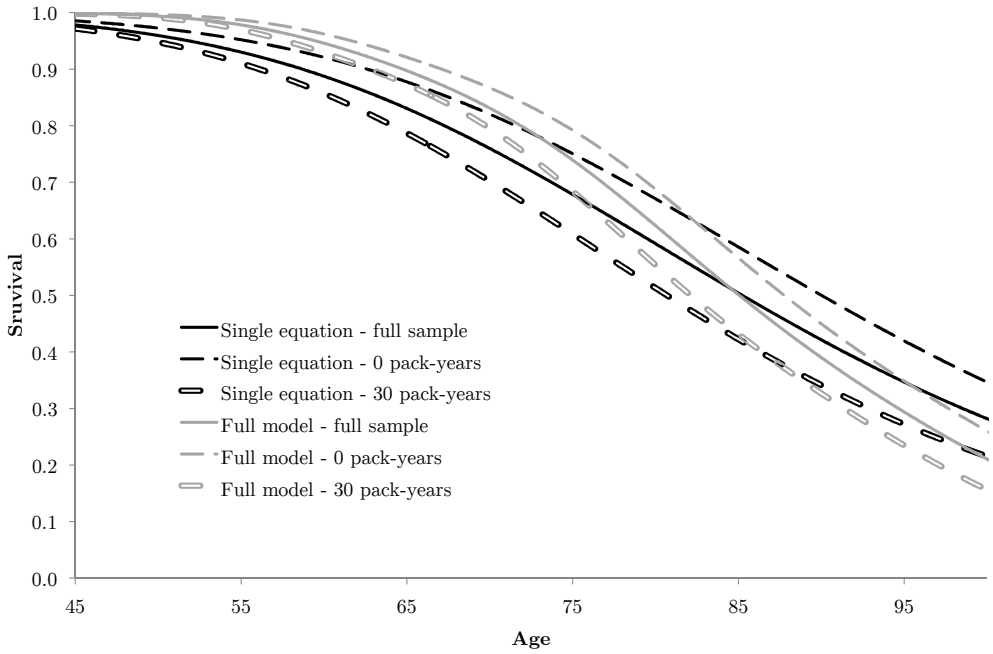

Figure 3.4: Estimated survival curves for cancer onset (males) – comparison of single-equation and full model results

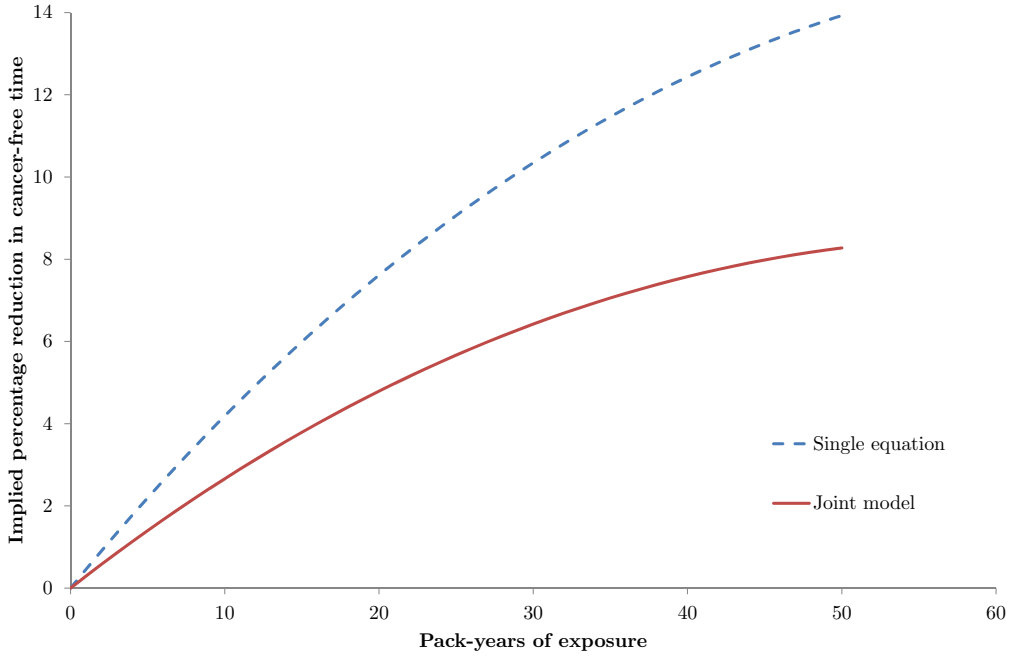

Figure 3.5: Estimated percentage reduction in time-to-cancer by smoking exposure – comparison of single-equation and full model results for current-smokers

While the φ_k parameter is, for each latent class, estimated as a constant, these estimated survival curves do not represent parallel shifts of each other, due to the non-linear relationship between φ_k and $S(t)$. Individuals in latent class 1, in particular, exhibit large increases in survival probabilities at all ages over others in the sample.

Any use of terms such as "time-to-cancer" or "age", with regard to this model, requires some clarification. What is being modelled in the cancer model is time to cancer in the absence of death. Individuals who die before developing cancer are treated as non-informative censored observations within the model, and contribute to the modelled likelihood as such. This means that, for instance, a predicted probability of survival at age 75 is calculated under the assumption that people could be observed to be at risk of cancer forever, and would not die and thus be censored in this way (that is, that individuals are infinitely-lived or that $m \leq n$ for all individuals, in the terminology used in the theoretical model). Any use of the term "age" must be seen in this light.

The difference between survival probabilities at older ages is particularly striking. As illustrated in Figure 3.3, at the age of 75, 98% of males in latent class 1 are predicted to have survived; in latent class 4, the corresponding probability is just 6%. For women, survival at 75 is predicted to be over 99% in latent class 1, and 11% in latent class 4. At the age of 95, these probabilities are 68% for men (79% for women) in latent class 1 and below 0.2% (below 0.4%) in latent class 4.

As illustrated in Figure 3.2, at an age of 75, 68% of males who are observed to have an exposure of 30 pack-years at the time of HALS1 are predicted to remain cancer-free, compared to 79% of those who had not smoked. For women, these respective probabilities are 74% and 83%. At the age of 95, these probabilities are 23% for men (28% for women) with an exposure of 30 pack-years and 35% (40%) for non-smokers.

The difference between results obtained using single equation estimates and those from the full DLFM (Figures 3.4 and 3.5) for men is also notable. The different duration dependence (γ) parameters estimated by the two models cause the implied survival functions from the two models to have a completely different shape: the single equation model implying more early failures but also more very late failures. Furthermore, the reduction in survival time, at the median, from having different observed smoking exposures is predicted to be smaller when using the joint model rather than single equation estimates.

The reduction in estimated median survival time (for males) between the counterfactual estimates for non-smokers and those with 30 pack-years of exposure is 9.3 years in the single equation model, and 5.7 in the joint model. Figure 3.5 further illustrates this using non-counterfactual methods, displaying the implied reduction in cancer-free time for different levels of smoking exposure. These results suggest a role for unobserved heterogeneity in explaining differences in survival times. Failure to account for this unobserved heterogeneity leads to differences in survival times between both individuals in different social classes, and individuals with different smoking exposures, to be overestimated.

3.8 Discussion

In addition to introducing cancer outcomes to existing research employing HALS (Balia and Jones, 2011), we here build on this existing work by modelling smoking exposure by pack-years rather than simply duration, and allowing health outcomes to vary with different exposures to smoking, rather than by whether the individual was a current smoker, former smoker, or never-smoker at the time of HALS1. Further, different relationships are found between parental smoking and the probability of a child smoking and the time to the child starting, depending on the gender both of parents and of their offspring.

The use of a joint model for smoking behaviours and health outcomes changes results substantially. The duration dependence parameter in the single equation model for cancer is more than twice as great as that in the joint model, leading to a much flatter estimated cancer survival function, and more early and late estimated failures. Further, the differences in estimated survival times associated with smoking exposure are higher when using single equation estimation rather than a joint model. Single-equation estimation yields estimates (for men) of this difference that are 2.4 years greater for the gap between the highest and lowest social classes, and 2.6 years greater for those with 20 observed pack-years of exposure than those with no observed years of exposure.

Assuming that individuals are rank-identical in the elevation of their respective hazards for cancer and death, the coefficients obtained in the main cancer model should be seen as lower bounds on the actual effect on healthy survival time without cancer, given that some individuals – who were likely to be registered as a cancer sufferer sooner than others

who remained at-risk – died before such a registration was possible. Interpretation of coefficients in the cancer registration model is complicated by the way in which those who do not develop cancer are censored: (at least some) deaths are informative censorings, and are symptomatic of the tendency of the individual to develop cancer, in the absence of death.

The reduction in time to cancer is estimated to be 5.7 years for male current smokers (5.8 years for women) at the time of HALS1 with 30 observed pack-years of exposure, compared to those who had never smoked at this time. At an age of 75, 93% of men with no observed smoking exposure are predicted to be cancer free, compared to only 82% of those with an observed exposure of 30 pack-years.

The latent class model appears to separate out some groups of individuals who are highly likely to develop some form of cancer due to unobserved factors, and others of those highly unlikely to do so. For instance, latent class 1 is composed of individuals of whom, under counterfactual simulations, almost 99% of men (over 99% of women) do not develop cancer by age 75, while the corresponding probability for individuals in latent class 4 is below 5% for men (below 10% for women). When posterior probabilities of class membership are estimated, and individuals sorted into their most likely class based on these probabilities, these differences are made even more stark: despite very similar lifestyle and circumstances for such individuals, only 2% of individuals most likely to be members of latent class 1 are observed to have developed cancer in the most recent follow-up, compared to 71% of those in latent class 4. The difference in lifespan for those individuals in each group who are observed to be deceased is approximately 20 years. These results point strongly to unobservable factors explaining a large part of the differences in health outcomes. This strongly highlights the importance of taking account of individual unobserved heterogeneity by modelling smoking behaviours jointly with health outcomes – not only in terms of lifespan (as in Balia and Jones (2011)) – but also in disease-specific (in this paper, cancer) outcomes.

Our results suggest a fruitful avenue of future research that would arise from collecting richer, long-panel data regarding smoking behaviours, and health outcomes. While these results point to the importance of modelling pre-diagnosis cancer-free lifespan rather than simply lifetime incidence, limitations of the dataset – a single cross-section from almost 25

years prior to the most recent cancer follow-up – employed again mean that care should be attached to directly interpreting the magnitude of the estimated coefficients, particularly those attached specifically to smoking behaviours. Further, larger datasets would allow more information to be collected on specific types of cancer, rather than merely grouping these into a single category. Such a dataset would allow the methods employed in this paper to be used to examine cancer-specific outcomes.

3.A Appendix

AIC and BIC scores for single-equation models are presented below:

Model	Observations	Loglikelihood	d.f.	AIC	BIC
Starting					
Expopower	2388	-7306.624	21	14655.25	14776.59
Exponential	2388	-9282.463	20	18604.93	18720.49
Loglogistic	2388	-6964.628	21	13971.26	14092.6
Weibull	2388	-7300.492	21	14642.98	14764.33
Gompertz	2388	-7967.207	21	15976.41	16097.76
Smoking exposure					
Generalised gamma	2388	-6063.621	24	12175.24	12313.92
Expopower	2388	-6058.478	24	12164.96	12303.63
Exponential	2388	-6069.637	22	12183.27	12310.39
Loglogistic	2388	-6119.277	23	12284.55	12417.45
Weibull	2388	-6069.346	23	12184.69	12317.59
Gompertz	2388	-6059.03	23	12164.06	12296.96
Cancer registration					
Generalised gamma	3784	-4469.158	29	8996.316	9177.233
Expopower	3784	-4472.943	29	9003.887	9184.804
Exponential	3784	-4544.547	27	9143.093	9311.534
Loglogistic	3784	-5045.162	28	10146.32	10321
Weibull	3784	-4471.475	28	8998.949	9173.628
Gompertz	3784	-4477.419	28	9010.838	9185.517
Mortality					
Generalised gamma	3784	-8598.828	30	17257.66	17444.81
Exponential	3784	-9021.817	28	18099.63	18274.31
Loglogistic	3784	-8943.939	28	17943.88	18118.56
Weibull	3784	-8599.991	29	17257.98	17438.9
Gompertz	3784	-8603.764	29	17265.53	17446.45

Table A3.1: Comparison of baseline hazards

Year of death	No. of deaths	Percentage
1984	0	0.00
1985	5	3.45
1986	17	11.72
1987	14	9.66
1988	18	12.41
1989	27	18.62
1990	22	15.17
1991	5	3.45
1992	1	0.69
1993	4	2.76
1994	4	2.76
1995	1	0.69
1996	1	0.69
1997	2	1.38
1998	0	0.00
2000	2	1.38
2001	1	0.69
2002	2	1.38
2006	2	1.38
2007	1	0.69
2008	10	6.90
2009	4	2.76
Total	145	

Table A3.2: Deaths where cancer is listed on an individual's death certificate, with the individual never registered as developing cancer

References

Adda, J. and Lechene, V. (2013), 'Health selection and the effect of smoking on mortality', *The Scandinavian Journal of Economics* **115**(3), 902–931.

Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B. and Aggarwal, B. B. (2008), 'Cancer is a preventable disease that requires major lifestyle changes', *Pharmaceutical Research* **25**(9), 2097–2116.

Balia, S. and Jones, A. M. (2008), 'Mortality, lifestyle and socio-economic status', *Journal of Health Economics* **27**(1), 1–26.

Balia, S. and Jones, A. M. (2011), 'Catching the habit: a study of inequality of opportunity in smoking-related mortality', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**(1), 175–194.

Banks, J., Marmot, M., Oldfield, Z. and Smith, J. P. (2006), 'Disease and disadvantage in the United States and in England', *JAMA: The Journal of the American Medical Association* **295**(17), 2037–2045.

Böckerman, P. and Ilmakunnas, P. (2009), 'Unemployment and self-assessed health: evidence from panel data', *Health Economics* **18**(2), 161–179.

Brown, H. and van der Pol, M. (2010), Intergenerational transfer of health: the role of time and risk preferences, *in* 'Health Economists Study Group 2011 (Winter)', York.

Contoyannis, P. and Jones, A. M. (2004), 'Socio-economic status, health and lifestyle', *Journal of Health Economics* **23**(5), 965–995.

Cox, B., Blaxter, M., Buckle, A., Fenner, N. P., Golding, J., Gore, M., Huppert, F., Nickson, J., Roth, M., Stark, J. et al. (1987), *The Health and Lifestyle Survey. Preliminary report of a nationwide survey of the physical and mental health, attitudes and lifestyle of a random sample of 9,003 British adults.*, Health Promotion Research Trust, London.

Cox, B., Huppert, F. and Whichelow, M., eds (1993), *The Health and Lifestyle Survey: Seven Years on: A Longitudinal Study of a Nationwide Sample, Measuring Changes in Physical and Mental Health, Attitudes and Lifestyle*, Dartmouth Pub. Co, Aldershot, Hants., England.

Cutler, D. (2008), 'Are we finally winning the war on cancer?', *Journal of Economic Perspectives* **22**(4), 3–26.

Cutler, D., Deaton, A. and Lleras-Muney, A. (2006), 'The determinants of mortality', *The Journal of Economic Perspectives* **20**(3), 97–120.

Dalstra, J., Kunst, A., Borrell, C., Breeze, E., Cambois, E., Costa, G., Geurts, J., Lahelma, E., Van Oyen, H., Rasmussen, N., Regidor, E., Spadea, T. and Mackenbach, J. (2005), 'Socioeconomic differences in the prevalence of common chronic diseases: an overview of eight European countries', *International Journal of Epidemiology* **34**(2), 316 –326.

Deaton, A. (2002), 'Policy implications of the gradient of health and wealth', *Health Affairs* **21**(2), 13–30.

Department of Health and Human Services (1989), *Reducing the health consequences of smoking. 25 years of progress*, US Government Printing Office, Washington, DC.

Department of Health and Human Services (2004), 'Smoking and Tobacco Use; 2004 Surgeon General's Report', http://www.cdc.gov/tobacco/data_statistics/sgr/2004/index.htm.

Department of Health and Human Services (2010), 'A Report of the Surgeon General: How Tobacco Smoke Causes Disease', http://www.surgeongeneral.gov/library/tobaccosmoke/report/index.html.

Doll, R. (1998), 'Uncovering the effects of smoking: historical perspective', *Statistical Methods in Medical Research* **7**(2), 87–117.

Doll, R. and Hill, A. B. (1954), 'The mortality of doctors in relation to their smoking habits: a preliminary report', *BMJ* **328**(7455), 1529–1533.

Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2004), 'Mortality in relation to smoking: 50 years' observations on male British doctors', *BMJ* **328**(7455), 1519.

Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2005), 'Mortality in relation to alcohol consumption: a prospective study among male British doctors', *Int. J. Epidemiol.* **34**(1), 199–204.

Doll, R., Peto, R., Hall, E., Wheatley, K. and Gray, R. (1994), 'Mortality in relation to consumption of alcohol: 13 years' observations on male British doctors', *BMJ* **309**(6959), 911–918.

Douglas, S. and Hariharan, G. (1994), 'The hazard of starting smoking: Estimates from a split population duration model', *Journal of Health Economics* **13**(2), 213–230.

Estève, J., Benhamou, E. and Raymond, L. (1994), 'Statistical methods in cancer research: descriptive epidemiology.', *IARC scientific publications* **IV**(128), 1.

Forster, M. and Jones, A. M. (2001), 'The role of tobacco taxes in starting and quitting smoking: Duration analysis of british data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **164**(3), 517–547.

Gordo, L. R. (2006), 'Effects of short- and long-term unemployment on health satisfaction: evidence from German data', *Applied Economics* **38**(20), 2335.

Gutierrez, R. G. (2002), 'Parametric frailty and shared frailty survival models', *Stata Journal* **2**(1), 22–44.

Heckman, J. and Singer, B. (1984), 'A method for minimizing the impact of distributional assumptions in econometric models for duration data', *Econometrica* **52**(2), 271–320.

Honoré, B. E. and Lleras-Muney, A. (2006), 'Bounds in competing risks models and the war on cancer', *Econometrica* **74**(6), 1675–1698.

Jones, A. M., Rice, N., Bago d'Uva, T. and Balia, S. (2007), *Applied Health Economics*, Routledge advanced texts in economics and finance, Routledge, London.

Kaprio, J. and Koskenvuo, M. (1989), 'Twins, smoking and mortality: A 12-year prospective study of smoking-discordant twin pairs', *Social Science & Medicine* **29**(9), 1083–1089.

Krall, E. A., Valadian, I., Dwyer, J. T. and Gardner, J. (1989), 'Accuracy of recalled smoking data', *American Journal of Public Health* **79**(2), 200–202.

Marmot, M. (2007), 'Achieving health equity: from root causes to fair outcomes', *The Lancet* **370**(9593), 1153–1163.

Marmot, M., Rose, G., Shipley, M. and Hamilton, P. J. (1978), 'Employment grade and coronary heart disease in British civil servants', *Journal of Epidemiology and Community Health* **32**(4), 244–249.

Marmot, M., Smith, G. D., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E. and Feeney, A. (1991), 'Health inequalities among British civil servants: the Whitehall II study', *The Lancet* **337**(8754), 1387–1393.

Mathers, C. D. and Schofield, D. J. (1998), 'The health consequences of unemployment: the evidence.', *The Medical Journal of Australia* **168**(4), 178.

Mroz, T. A. (1999), 'Discrete factor approximations in simultaneous equation models: Estimating the impact of a dummy endogenous variable on a continuous outcome', *Journal of Econometrics* **92**(2), 233–274.

Peto, R., Lopez, A., Boreham, J. and Thun, M. (2006), 'Mortality from smoking in developed countries, 1950-2000'.

Reid, A., de Klerk, N. H., Ambrosini, G. L., Berry, G. and Musk, A. W. (2006), 'The risk of lung cancer with increasing time since ceasing exposure to asbestos and quitting smoking', *Occupational and Environmental Medicine* **63**(8), 509 –512.

Roemer, J. E. (1998), *Equality of Opportunity*, Harvard University Press, Cambridge, Mass.

Rosa Dias, P. (2009), 'Inequality of opportunity in health: evidence from a UK cohort study', *Health Economics* **18**(9), 1057–1074.

Saha, A. and Hilton, L. (1997), 'Expo-power: A flexible hazard function for duration data models', *Economics Letters* **54**(3), 227–233.

Schaap, M. and Kunst, A. (2009), 'Monitoring of socio-economic inequalities in smoking: Learning from the experiences of recent scientific studies', *Public Health* **123**(2), 103–109.

Singh, G. K., Miller, B. A., Hankey, B. F., Edwards, B. K., Singh, G. K., Miller, B. A., Hankey, B. F. and Edwards, B. K. (2003), 'Area socioeconomic variations in US cancer incidence, mortality, stage, treatment, and survival, 1975-1999'.

Thomas, B., Dorling, D. and Smith, G. D. (2010), 'Inequalities in premature mortality in Britain: observational study from 1921 to 2007', *BMJ* **341**(c3639), c3639.

University of Cambridge Clinical School (2009), 'HALS deaths and cancer working manual, june 2009'.

US National Cancer Institute (2006), 'Cancer trends progress report - stage at diagnosis'.

Vallejo-Torres, L. and Morris, S. (2010), 'The contribution of smoking and obesity to income-related inequalities in health in England', *Social Science & Medicine* **71**(6).

Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y. T., Gupta, P. C., Hackshaw, A., Matos, E., Samet, J., Sitas, F., Smith, J., Stayner, L., Straif, K., Thun, M. J., Wichmann, H. E., Wu, A. H., Zaridze, D., Peto, R. and Doll, R. (2004), 'Tobacco and cancer: Recent epidemiological evidence', *Journal of the National Cancer Institute* **96**(2), 99 –106.

Wilkinson, R. G. (1996), *Unhealthy Societies: The Afflictions of Inequality*, Routledge, London.

Wilkinson, R. and Pickett, K. (2010), *The Spirit Level: Why Equality Is Better for Everyone*, Penguin, London.

Wilson, D. L. (1994), 'The analysis of survival (mortality) data: Fitting Gompertz, Weibull, and logistic functions', *Mechanisms of Ageing and Development* **74**(1-2), 15–33.

Chapter 4

Age, proximity to death and ill-health, in the presence of compression of morbidity.

Daniel Howdon* [1,2] & Nigel Rice [1,2]

1. Department of Economics and Related Studies, University of York, YO10 5DD, UK
2. Centre for Health Economics, University of York, YO10 5DD, UK

Tel.: +44 1904 32 1411.
**E-mail address:* daniel.howdon@york.ac.uk

Abstract

This paper uses Hospital Episode Statistics, English administrative data from the Department of Health, to further investigate the red herring thesis, as advanced by Zweifel et al. (1999). We use a sample of over 100,000 individuals who used healthcare in the financial year 2005/06 and had died by the end of the financial year 2012/13. We use a panel structure to follow individuals over seven years of this administrative data, containing estimates of inpatient healthcare expenditures (HCE), information regarding individuals' age, time-to-death (TTD), and morbidities at the time of their admission. We find that, while TTD might better explain HCE than does age, TTD itself merely proxies for individuals' morbidities, and no longer explains differences in HCE once we condition on morbidity characteristics. Our results point to an important role for including estimates of morbidity (and changes in their respective prevalences) when estimating future HCE growth associated with an ageing population, rather than relying on a ceteris paribus role for TTD.

JEL codes: H51; J11; I19.

Keywords: healthcare expenditures, ageing, time-to-death.

Acknowledgements: This is an independent report commissioned and funded by the Policy Research Programme in the Department of Health from the Economics of Social and Health Care Research Unit (ESHCRU). ESHCRU is a joint collaboration between the University of York, London School of Economics and University of Kent. The views expressed are those of the authors and may not reflect those of the funders. Daniel Howdon acknowledges PhD funding from the Economic and Social Research Council under the Large Grant Scheme (RES-060-25-0045).

4.1 Introduction

There is concern that the demographic pressures of population ageing will lead to an unprecedented rise in public expenditures to levels unsustainable under current financing arrangements. In the UK in 2013 approximately 17% of the population (11 million individuals) were aged 65 years or over. This represents a rise of 17.3% in this age group on a decade earlier. Projections suggest that by 2050 this group will have increased disproportionately to younger age groups accounting for approximately 25% of the population (Cracknell, 2010)). The growth in the proportion of older individuals is partly due to increased longevity and partly due to the age structure of the population, particularly ageing of the generation of baby boomers of the post war period to the early 1970s. Health care expenditures have also risen substantially over time both in real terms and proportional to economic growth. Close to the inception of the National Health Service (NHS) net expenditure (net of patient charges and receipts) on the UK NHS in 2050/51 was £11.7b (in 2010/11 prices); representing 3.5% of Gross Domestic product (GDP). This rose to £121.3b in 2010/11; approximately 8.2% of GDP. Over the twenty-five year period from 1999/00 to 2014/15 (forecast) expenditure in England has almost doubled to £103.7b (2010/11 prices) with an average expenditure per head of population of £1,900 (Harker, 2012). Abstracting from issues such as technological innovation, the concern is that as the share of the population at older ages rises, the economic consequences will become increasingly unsupportable.

Interest in the link between ageing populations and health care expenditures can be traced back 25 years when the International Monetary Fund (IMF) asserted that 'demographic pressures [in the UK] of an aging population will be associated with increased demand for medical services', and presented descriptive statistics from various countries, showing that older patients, on average, had greater health care costs than younger patients (Heller et al., 1986). A report by the Organisation for Economic Co-operation and Development (OECD) predicted that across Europe population ageing will create a rise in age-related social expenditures from around 19% of GDP in 2000 to around 26% by 2050. Old-age pension payments and expenditure on health and long term care was deemed responsible for approximately half this increase (Dang et al., 2001). Approaches to pre-

dicting expenditure growth vary, but in a simplistic form consists of computing observed expenditures per head for different age-sex groups and multiplying by projections of the number of people expected to fall into each group. This approach, however, fails to consider the underlying drivers of heath care expenditures and the relative role of age, or, as has been suggested, proximity to death, or underlying levels of disability and ill-health, in determining expenditures and its likely growth (see Gray (2005)).

Additional to projections of population ageing is the potential change in the health profile of the population over time. An 'expansion of morbidity' hypothesis has proposed that the 'net contribution of our successes has actually been to worsen the people's health', as improvements in health care tend to lengthen the lives of those living with illness disproportionately to its effect on those living without (Gruenberg, 2005). Should population ageing occur alongside a deterioration of health at older ages, then this will exacerbate impacts on public expenditures. While subsequent academic research into these claims – notably, research in the 'compression of morbidity' and 'red herring' strands of literature – have given reason to suggest that such concerns may have been misplaced or exaggerated, concern over the impact of an ageing population on HCE has persisted. Indeed, even in 2012, the UK's then-Secretary of State for Health claimed that the fact that 'the number of people aged over 85 in this country will double in the next 20 years' was one of two factors in 'costs... rising at an unaffordable rate' (Lansley, 2012). He further argued that 'age is *the* principal determinant of health need'[1], and that local NHS budgets should be recalibrated to be based on this, as a result (Williams, 2012).

This paper uses UK administrative data from Hospital Episode Statistics (HES) to consider research relating to the 'red herring' thesis advanced by Zweifel et al. (1999). This seeks to explore the determinants of health care expenditures, with particular attention to the role played by age, time-to-death (TTD), and morbidity. We do this in a unique way by following a sample of the population of users of inpatient hospital services in England in 2005/06 over seven years and constructing a panel on individual health care expenditures and morbidity over this period. We merge information on date of death, where relevant, obtained from ONS statistics from which TTD for decedents is computed. We show that TTD dominates age as a key driver of health care expenditures and mor-

[1] Emphasis ours.

bidity characteristics dominate TTD. This finding extends the 'red herring' literature by showing that TTD is itself a 'red-herring' and acts as a proxy for morbidity.

4.1.1 Compression of morbidity

The 'compression of morbidity' strand of literature beginning with Fries (1980) suggests that individuals are likely to see increased morbidity in a fewer number of years at the end of their life and, as a consequence of this, the implications for HCE of an ageing population become less clear. While individuals born into later cohorts may enjoy a longer lifespan, they will not have an increased number of years living with chronic conditions, due to being able to postpone their onset compared to those in earlier cohorts, and possibly even avoid them entirely. Fries (2005) identifies three separate 'eras' of illness and well-being experienced during the 20th Century and beyond: an era of infectious disease, followed by an era of chronic disease, followed by an era described by the author as 'directly related to the process of senescence, where the aging process itself, independent of specific disease, will constitute a major burden of disease'. Senescence – the process of ageing – is characterised by the 'decline of maximal function of [all] vital organs', beginning before any chronic disease takes hold: deaths where this function declines below a level necessary to sustain life, in the absence of any disease occasioning this, may be termed 'natural deaths' (Fries, 2005). Freedman et al. (2002), in a systematic review covering research that had been conducted between 1990 and 2002 found that many measures of disability and limitations in old age had seen declines in recent years: in particular, a change of -1.55% to -0.92% per year in those reporting any disability during the late 1980s and 1990s. Romeu Gordo (2011) observe a cohort-on-cohort fall in the number of individuals with high levels of disability-related functional problems in their everyday life for those born between 1924 and 1947 in the US. Cutler et al. (2013), using Medicare records from the US, present evidence of an increase in disability-free life between 1991 and 2009. Cross-country international evidence on the changing patterns of disability rates across nine OECD countries is provided by Jacobzone et al. (2000). Consistent with the above literature, they report evidence of significant falls in severe disability rates. The importance of this issue for forecasting HCE depends upon how changes in mortality and changes in morbidity occur and interact with each other. If the onset of chronic conditions

– those imposing large costs on health systems – can be postponed out of an individual's lifetime, then health care costs may fall as later cohorts enjoy a longer lifespan, with a reduced level of necessary treatment for chronic conditions.

The morbidity profile of individuals, according to this research, at any given age has improved over time, leading to health problems being experienced later in life and more closely to death. In the illustrated case (Figures 4.1 and 4.2[2]), individuals live up to a longer observed maximum age (indicated by the shift out of the survival curve from S_1 to S_2 in Figure 4.1), and have a higher observed level of health at all ages (indicated by the shift out of the health status curve from H_1 to H_2 in Figure 4.2). Both survival curves and health status curves have become increasingly rectangular. The effect on health care expenditure (HCE) is ambiguous, given that generally more healthy ageing – a decrease in morbidity at any given age – puts downward pressure on HCE, while an increase in life expectancy, *ceteris paribus*, puts upward pressure on HCE. The actual relationship between health care costs and changes in morbidity and mortality profiles at every given age depends upon the changing shape of these two curves. The use of age *per se* in predicting future health care costs should be approached with caution, as a result.

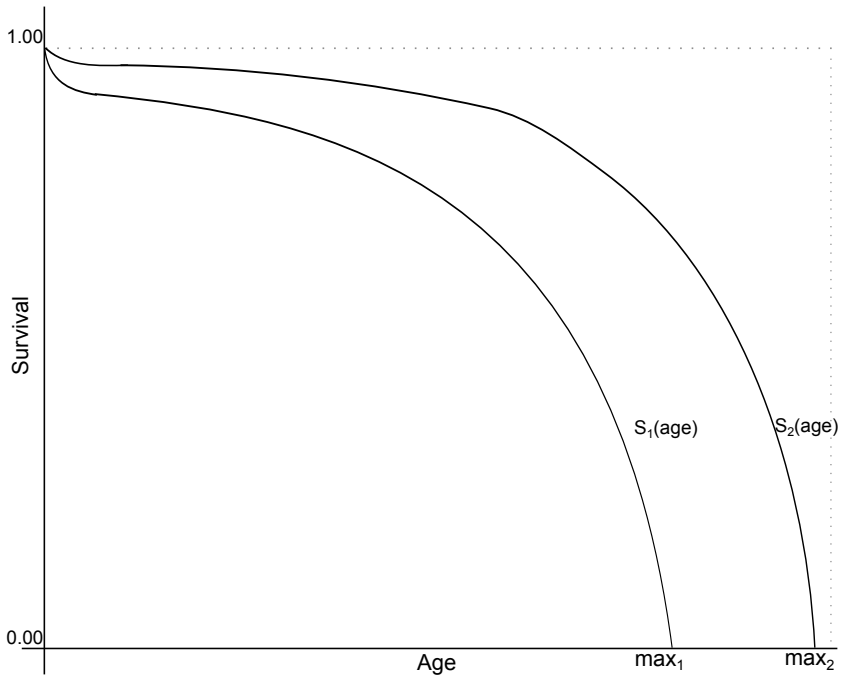

Figure 4.1: Stylised change in survival curves

[2]Adapted from Fries (1980) and `http://www.aei.org/files/2008/06/27/20080626_WashingtonAEI.pdf`.

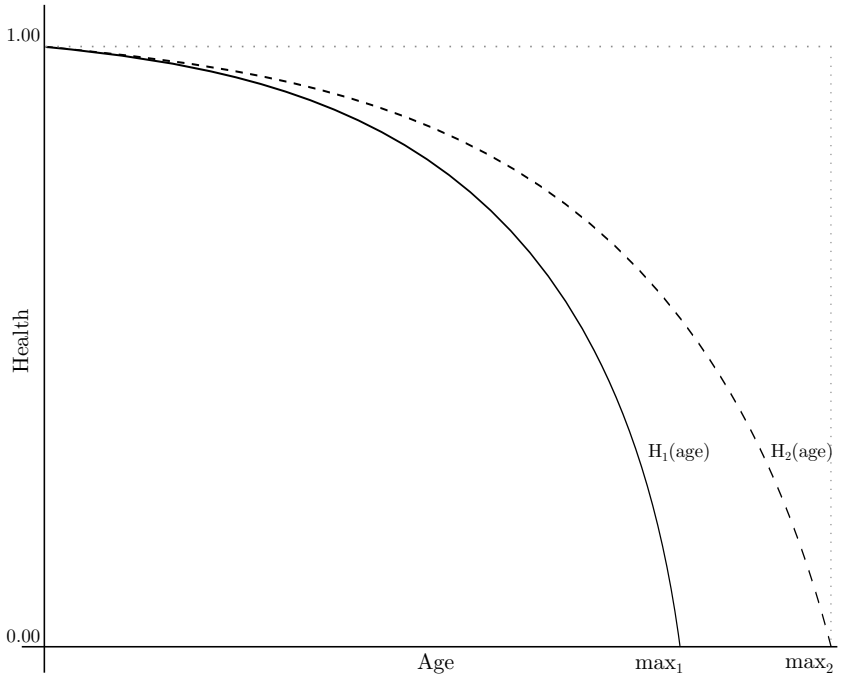

Figure 4.2: Stylised change in health profiles

4.1.2 Age and time-to-death

The 'red herring' strand of literature further gives empirical reason to suggest that claims of steeply-rising future HCE due to population ageing.[3] may have been exaggerated, potentially owing to morbidity being concentrated in later years of life. Zweifel et al. (1999), using Swiss sickness fund data, finds that no effect of age on health care expenditures existed after controlling for 'time-to-death' (TTD), i.e. the difference in time between when treatment occurs and the individual's death. Owing to the number of individuals with zero HCE, a two-step model (with a probit first stage and OLS second stage) was employed, with only deceased patients included in the model. Such work was criticised on the grounds of potential endogeneity, with time-to-death affected by both present, previous (and, due to the nature of how TTD must be measured, future) HCE. In a subsequent paper, Zweifel et al. (2004) seek to test for such problems, finding that while TTD is endogenous, their results were 'fairly robust' to the error this induces. Werblow et al. (2007) find that age is a small (but statistically significant) determinant of HCE after controlling for TTD for patients using long-term care (LTC), such as those in care homes, and is not associated

[3]HCE may rise due to technological change brought about by new expensive innovations in health care treatments, or due to shifting patterns of morbidity.

with HCE for non-LTC patients. More complicated methods, such as those employing generalised linear models, have since been used, for example by Werblow et al. (2007), in order to deal with the non-normal properties (such as positive skewness) exhibited in the distribution of HCE. These papers have corroborated results obtained using probit and OLS two-step models. Felder et al. (2010), in a recent paper in this series, first predict individuals' survival based on observed HCE and socioeconomic characteristics (in early waves), before using predicted values based on this as an instrument for TTD in explaining HCE in later waves. The authors find that, while TTD cannot be deemed exogenous, any effect of age on HCE becomes insignificant when TTD (or instrumented TTD) is included in the model.

While use has been made of morbidity markers in models of long-term care expenditures (LTCE) (see de Meijer et al. (2011)), such use has not been made in models explicitly investigating the link between HCE and population ageing. One possibility is that TTD is itself a red herring, in that it is simply a proxy for morbidity, unobserved in existing HCE models. This seems intuitively plausible: in the years before death, it is likely that morbidity will increase, leading to more treatment, and that comorbidities complicating the treatment of the disease bringing about the hospital episode will also increase. Shwartz et al. (1996), in work predating the original red herring hypothesis, note that the inclusion of variables for comorbidities increase substantially the explanatory power of models. It seems likely that variables incorporating 'time-to-death' in more recent models of HCE are picking up, in large part, these comorbidities, which are unobserved in existing HCE models: indeed, de Meijer et al. (2011) conclude that time-to-death 'largely approximates disability' in models of LTCE. Dixon et al. (2011), in proposing individual-level formulae for resource allocation in the UK's National Health Service (often termed 'Person-Based Resource Allocation', or PBRA) include individual level morbidity markers, finding that these have a 'powerful effect... in predicting individual level expenditure'.

The process generating HCE is clearly not a simple function of those explanatory variables used in existing 'red herring' research: the actual data-generating process behind these health care expenditures is unlikely to be characterised accurately by a simple use of age, historical time and time-to-death. In addition to the aforementioned problems surrounding TTD and age as a proxy for morbidity, as Breyer et al. (2014) note, many ex-

isting models are likely to be characterised with substantial endogeneity problems, which lead to potential bias in the estimation of the change in HCE as an individual ages or approaches death. The authors control for potential endogeneity introduced by differential treatment based on a physician's view of the patient's expected health benefits from treatment, proxied by actuarial tables of life expectancy conditional on age. If physicians expect individuals to respond differently to treatment, this may cause those who are more likely to respond to treatment to be treated more intensely than those who are not, thus increasing expected HCE for individuals who are younger, further-from-death or with fewer comorbidities because of physician selection. Conversely, HCE for older individuals – or, more likely, individuals in the final years of life – may rise as intensity of treatment becomes stronger with heroic efforts to save an individual's life, possibly motivated by ethical 'rule of rescue' concerns when faced with an identifiable, gravely sick individual (Jonsen, 1986). Breyer et al. (2014) jointly estimate this possible physician selection based on life expectancy alongside a model for health care expenditures, incorporating both age and time-to-death as explanatory variables. They find that increasing survival rates for the elderly in Germany have positive impacts on HCE, arguing that this is explained by physician selection: treating patients more intensively if they expect positive results from treatment over a longer time span.

Datasets used within the 'red herring' literature are, in general, sickness fund datasets, with only Seshamani and Gray (2004) using population-level (for users of NHS treatment) data, the Oxford Record Linkage Study, a longitudinal dataset of all individuals within an area of Oxfordshire. We believe our paper to be the first in this strand of literature to use a sample of individuals from a comprehensive national-level dataset of health care users.

The extent to which 'red herring' and related issues are of interest depends upon the intended use of such research. Much existing literature focuses on projections of future health care costs given an ageing population, with the headline results of some papers (such as Stearns and Norton (2004) and Seshamani and Gray (2004)) being the overestimation of expected costs for a given future year when TTD is an omitted variable. This is due to the collinearity between TTD and age for a given individual: an individual who gets one year closer to death also gets one year older, and so the impact of TTD is picked up by age in such models. The inclusion of morbidity markers in addition to, or

replacing, TTD would allow greater precision of future estimates where reliable estimates of morbidity prevalence, and the cost of treatments, conditional on age and TTD were known. Certainly, if the compression of morbidity hypothesis holds, and individuals are able to postpone the onset of chronic diseases – with associated higher HCE – to a time period closer to their death, or even indefinitely, explicitly considering morbidity rather than proxying this by age and/or TTD becomes ever more important.

We build upon the compression of morbidity and red herring strands of existing literature, seeking to further examine the relationship between ageing, time-to-death and health care expenditures. The original red herring hypothesis is that, once time-to-death is included in models of HCE, age *per se* does not explain changes in HCE. While models intended for resource allocation (Dixon et al., 2011) have already included morbidity as an explanatory variable in HCE for the general population, other applications of models of HCE have not – in particular, those focusing explicitly ageing populations, or costs in the years approaching death. This paper seeks to bridge the gap between the red herring strand of literature and models of resource allocation, treating morbidity measures as omitted variables in models of current health care expenditure, and examining what the relationship between age, TTD and HCE is once morbidity is included in these models. This has important implications for the prediction of future health care expenditures, especially in the presence of a compression of morbidity over time. While, due to the shortness of our panel, we cannot directly investigate the existence of a compression of morbidity, successive papers have suggested its existence (Fries, 1980; Freedman et al., 2002; Romeu Gordo, 2011) and, consequently, it is necessary to examine the relationship between morbidity and HCE to inform predictions of future HCE. Assuming that future changes in age-related morbidity-specific ill-health can be anticipated, the inclusion in these models of age, morbidity, as well as any residual association with TTD, will allow for better prediction of future HCE.

4.2 Data

4.2.1 Data sources

Information on patient-level hospital use and associated Reference Costs for treatment are derived from the Hospital Episodes Statistics (HES) dataset, published by the Health and Social Care Information Centre (HSCIC). This is complemented with small-area data on years of potential life lost (YPLL) published by the ONS, and individual level mortality information, jointly published by the HSCIC and the ONS.

We use successive years (financial years 2005/06 to 2011/12) of the HES dataset, along with the associated years of the NHS's Reference Costs (RC) data. HES has been published for each financial year since 1989/90 and is available for admitted patient care, outpatient, accident and emergency and maternity cases. The admitted patient care ('inpatient') HES dataset that we use provides information on individual-level patient characteristics and diagnoses and procedures undergone for all patients admitted to hospitals in England.

HES contains diagnostic data, categorised (since 1995/96) according to the tenth revision of the World Health Organization's International Classification of Diseases (ICD-10). Details of procedures and interventions are recorded according to the fourth revision of the Office of Population, Censuses and Surveys' Classification of Intervention and Procedures (OPCS-4) (Health & Social Care Information Centre, 2013).

HES is broken down by completed "episode" – each record consists of a continuous period of care at a single provider of treatment under the same consultant. A new record is generated when a patient is either transferred to the care of either a new consultant, transferred to a new provider, or is discharged from hospital. Although individuals are not identifiable, the `hesid` variable allows individuals to be tracked across episodes, to create spells – multiple episodes unseparated by a temporal break outside of hospital. Dawson et al. (2005) note that this data can be used to generate different units of analysis:

- The aforementioned episodes (or consultant episodes, CEs) themselves, including episodes that are incomplete within the financial year in question.
- Finished consultant episodes (FCEs) – those which are finished within the financial year in question, but may have begun earlier.

- Provider spells (PSs) – all episodes in the same spell and within the same hospital, but potentially under the care of different consultants.

- Continuous inpatient spells (CIPSs) – all episodes in the same spell, but potentially under the care of different consultants and/or at a different provider.

Although PSs or CIPSs are often preferred when looking at the event of a patient's hospitalisation, the costing of a patient's time in hospital and the recording of their diagnoses and procedures undergone are made at the episode level.

The `hesid` variable further enables the tracking of patients over different years of the HES dataset, and consequently the creation of a panel structure for the data. Information within the HES dataset – most commonly, information regarding diagnosis, treatment and age of the patient – is used to apply the most appropriate HRG categorisation to the dataset. We use the Health and Social Care Information Centre's Consultation 'Grouper' software to carry out this first step.

We use the most recent version of this Grouper – for the 2011/12 financial year – for all seven of the years we use, to categorise patients into Healthcare Resource Groups (HRGs). HRGs are used to categorise patient spells not only by broad procedure or diagnosis, but by the type and complexity of the patient's spell, into one of over 1,400 groupings. This allows us to apply the current best-practice methods for grouping patients into HRGs based on the information available.

In a final stage, we apply available estimates of hospital costs for each inpatient spell, using reference costs data for the relevant tax year. These costs are based on each NHS provider's estimates of their own costs for each patient spell, categorised by HRG. These reference costs are derived from accounting costs for each HRG, submitted by each organisation providing secondary care in England (Department of Health, 2012). The NHS Costing Manual provides guidance to all providers to support the calculation of reference costs and to enforce more uniform standards for costing methodologies. We use the estimate provided by the hospital providing treatment as our estimated cost for the patient's episode.

The DH's Reference Cost data is submitted on a full absorption basis – that is, taking account of all direct and indirect costs relating to the activities in question, as well as

a proportion of an estimate of all overhead costs relating to the overall running of the provider. Further, to account for the fact that costs will vary even within HRGs, hospitals are required to provide *per diem* costs for longer admissions that exceed a given 'trim point', which differs by each HRG. This trim point is defined as the upper quartile of length of stay, plus 1.5 times the inter-quartile range for length of stay for that HRG (Department of Health, 2012). Even within the same HRG, costs are not identical but differ according to the patient's length of stay. An estimate of costs for each inpatient spell is obtained by matching data on costs for that provider in the Reference Costs database to HRG for each episode in the relevant year's HES data.

We add information regarding an individual's death from linked HES-ONS mortality data. Again, the merging process is carried out using the individual's `hesid`. The latest version of this data provides information on deaths to the end of the 2012 calendar year, and therefore provides information on some individuals whose deaths are known to have occurred after the end of the final wave in our dataset. Where individuals are known to have died, they are included up to and including the final quarter of their life, and not included in the panel in following years. TTD can only be measured – for decedents – retrospectively, using information available at the time of the individual's death. Within a panel data structure, TTD in any given wave is unknown at the time at which data within that wave is collected. Existing literature on the relationship between HCE and TTD uses observed TTD (or an instrument for TTD) as an explanatory covariate in models along with age and other characteristics.

We make use of the Office for National Statistics' Indices of Multiple Deprivation (IMD), by Lower Super Output Area (LSOA) in order to construct an instrument for TTD. LSOAs are defined at the time of the UK's decennial Census and are made up of similarly-sized small areas of the country. HES data, for the years used in our dataset, provides information on the individual's LSOA of residence at the time of the 2001 Census. At this time, LSOAs in England consisted of 32,482 areas of populations between 1,000 and 3,000, with between 400 and 1,200 households (Office for National Statistics, 2011).

Indices of Multiple Deprivation, at this LSOA level, are measures of the levels of deprivation in these small areas. Although made up of seven domains (income, employment, health and disability, education, housing, living environment and crime (Department for

Communities and Local Government, 2011)), we primarily make use of one of the indicators that forms part of the health and disability IMD score: years of potential life lost (YPLL). This consists of a standardised measure of premature mortality calculated using information for all individuals to have died before the age of 75, as described in Blane and Drever (1998)[45]. Although the LSOAs themselves are defined every ten years at the time of the UK's census, statistics for each domain are collected and published for these areas more regularly: we make use of those published in 2007 (produced using data from 2001-2005 inclusive), and 2010 (produced using data from 2004-2008 inclusive) (Department for Communities and Local Government, 2008, 2011). For each of these years, we use LSOAs as defined in the 2001 UK Census. While these figures are comparable within years, the data collector (the UK's Department for Communities and Local Government) caution against using this data for trend analysis. We therefore include each wave of this measure separately as instruments.

Our random sample consists of 111,136 individuals (55,039 men and 56,097 women) aged 50 years and older, and is taken from all inpatient episodes in England in the financial year 2005-2006. The size of sample taken and age cut-off are somewhat arbitrary. Sample size was selected to enable computations not to become burdensome, and the age cut-off was selected to ensure sufficient deaths were observed in the data to make meaningful influence. We follow all sampled individuals until their death. We collapse all inpatient episodes for a given quarter into a single observation in our data. This observation contains a sum of all hospital costs incurred in all episodes finishing in that quarter, as well as diagnostic information contained in the ICD-10 codes for those episodes. In principle, the ICD-10 classification allows for up to 14,400 different diagnoses. To make these more manageable for analysis, however, we further collapse this information using the US Agency for Healthcare Research and Quality's Clinical Classifications Software (CCS) method to convert ICD-10 codes to CCS codes (US Agency for Healthcare Research and Quality, 2009). This reduces the number of different groupings to a more manageable 260 mutually-exclusive, and clinically mean-

[4]The Office for National Statistics, however, use 75 rather than 65 years, in their implementation of this method, as the age at which mortality is considered to be premature (Department for Communities and Local Government, 2011).

[5]Details of the method employed by the ONS were obtained in personal communication with the study's author, Chris Dibben.

ingful, categories[6]. Where individuals do not have any episodes in a quarter, they are recorded as having zero hospital costs, and as having zero observed morbidities arising from diagnostic information. The recording of zero morbidities might be unrealistic for patients observed to have hospitalisations in recent periods and for whom there is likely to exist an underlying, albeit less grave, health problem. In the absence of additional information on the gravity of any residual health problem, we assume that such health issues are insignificant relative to those leading to a hospitalisation. While we include a sum of all hospital costs for episodes ending in the quarter in question, we include only a maximum of three diagnoses for each individual, for a maximum of five episodes ending in that quarter. Using the merged mortality data, we are able to add a variable for the individual's time-to-death, measured in number of quarters to death.

Descriptive statistics for the sample of inpatient healthcare users in the first wave are presented in Tables 4.1 & 4.2.

Table 4.1: Summary statistics (Quarter 1, men. n=55039.

Variable	Mean	Std. Dev.	Min	Max
HCE	613.91	1801.81	0	92886.46
log(HCE)	2.03	3.28	0	11.44
TTD (quarters)	9.71	7.84	0	27
log(TTD)	2.04	0.88	0	3.33
Age	75.09	10.13	50	105.66
YPLL (IMD 2007)	65.83	15.70	33.30	191.50

Table 4.2: Summary statistics (Quarter 1, women. n=56097.)

Variable	Mean	Std. Dev.	Min	Max
HCE	666.85	1879.67	0	71405.11
log(HCE)	2.05	3.32	0	11.18
TTD (quarters)	9.89	7.91	0	27
log(TTD)	2.05	0.89	0	3.33
Age	78.28	10.78	50	111.16
YPLL (IMD 2007)	66.06	15.60	33.30	191.50

As is usual, the distribution of HCE is positively skewed, with this skewness reduced

[6]A full list of these CCS groupings is provided in Appendix 4.A

somewhat when we take a logarithmic transformation. As would be expected due to their longer lifespan, on average, the average age of women in the sample is somewhat higher than that for men (approximately 78 years, compared to 75). Similarly, women are observed for, on average, slightly more waves (mean TTD for women is 10.89 compared to 10.71 for men).

Diagrams, presented in Figures 4.3 and 4.4, based on descriptive statistics, provide some illustration of the existing red herring thesis. When we take a 5% sample of all users of inpatient services (118,263 men and 125,931 women), including both decedents and survivors, HCE appear to increase with age (top-left panel). This is the usual age-expenditure curve that is used to infer rising costs with population ageing: the assumption being that as the population ages, ignoring the drop in expenditures at very high ages as this is likely due to low sample sizes, the curve continues to rise as an extrapolation of the observed trend. The observation that expenditures rise with age, however, is an artifact of a compositional effect. The naïve age-expenditure curve is composed of individuals who are known to have died during the period of observation (the sample used in estimation, of 55,039 men and 56,097 women) – who have, on average, high expenditures for this period (top-right panel) – and individuals who are known to have survived (63,224 men and 69,834 women) who have, on average, lower expenditures for this period (bottom-left panel). The average expenditures for individuals observed to have died during the sample preiod are far greater than for individuals who survive. This suggests an important role for time-to-death in explaining HCE. As the proportion of the full population who are decedents increases with age, the näive observed relationship between age and expenditure displays an increasing trend. Note, however, that average expenditures for both decedents and survivors display a flatter profile than that depicted for the full population suggesting a less important role for age.

When we focus on decendents, and consider average HCE by proximity to death, we observe a large increase in costs in terminal quarters – particularly in the year immediately before death. Figure 4.5 shows a similar relationship between expenditures and TTD for men at selected ages. In general, expenditure in quarters preceding the final three average around £500 (although there is variation). In the final three quarters, and particularly the final quarter, we observed a large increase in expenditure. With the exception of 50 year

olds, there is a clear gradient of health expenditures rising most dramatically in the final quarter of life with average increases over the penultimate quarter ranging from £329 for 55 year olds to £1,249 for 90 year olds.

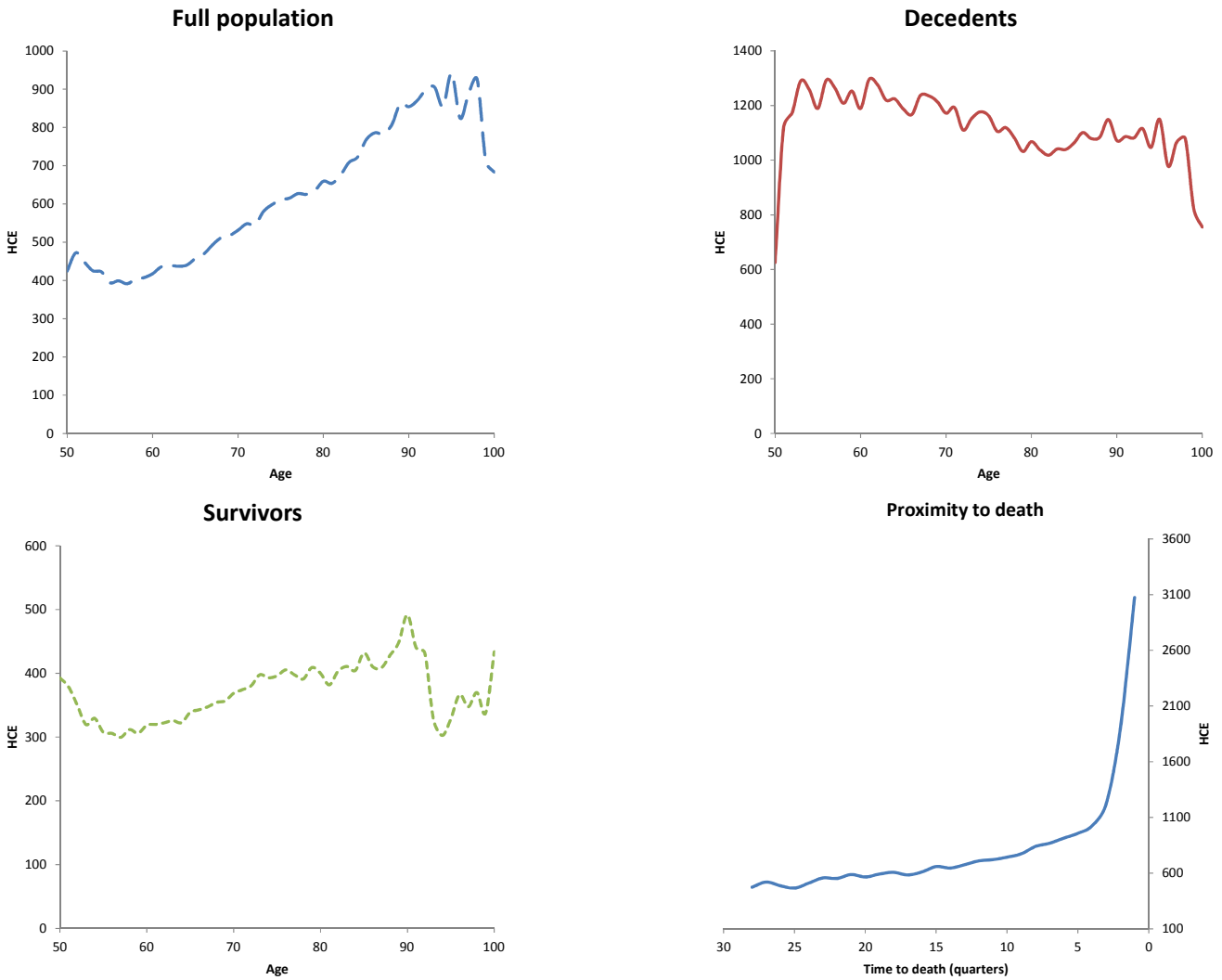

Figure 4.3: Healthcare expenditures by age and proximity to death, males

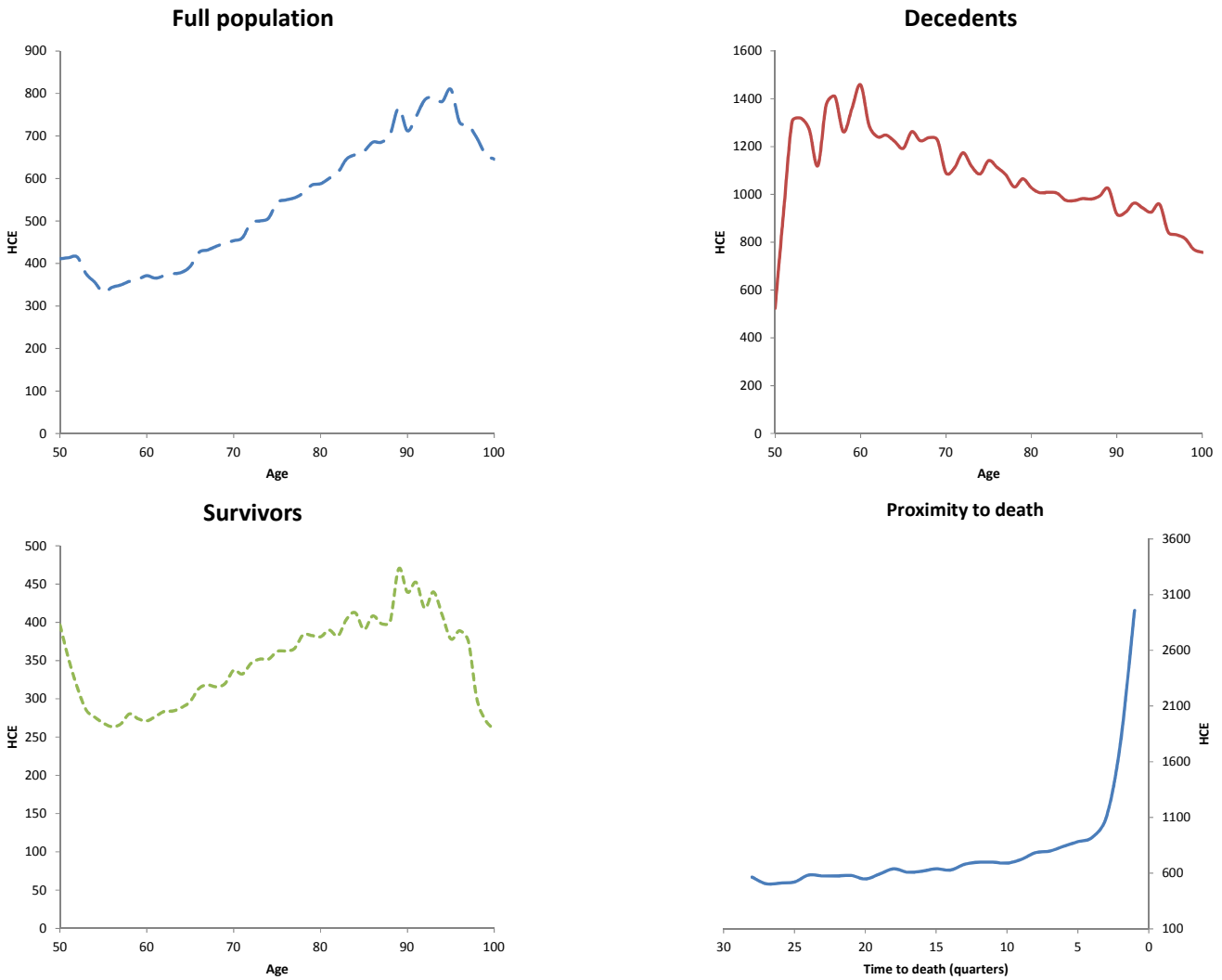

Figure 4.4: Healthcare expenditures by age and proximity to death, females

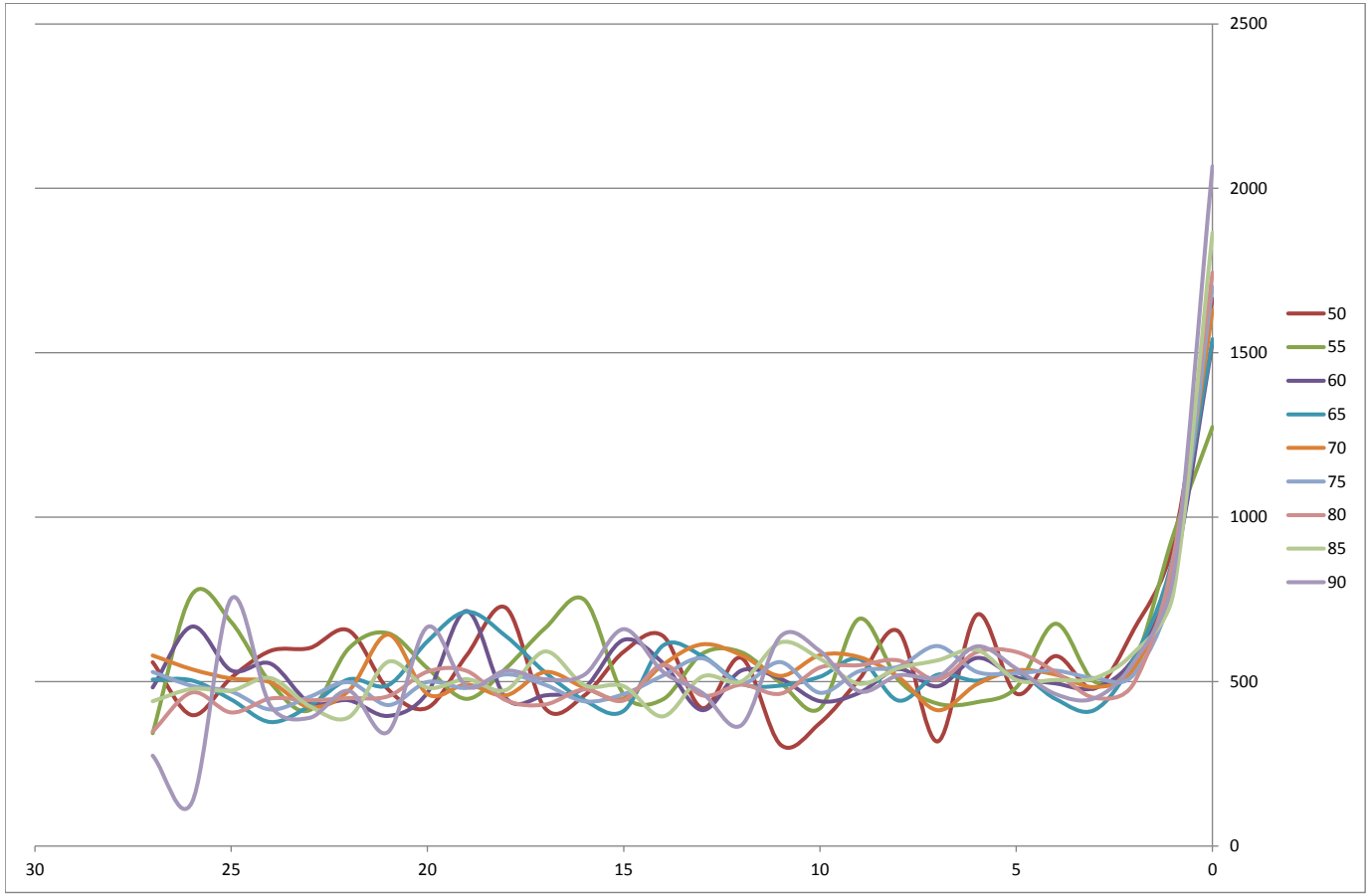

Figure 4.5: Healthcare expenditures (y-axis, GBP) by time to death (x-axis, quarters), males by age

The relationship between HCE and TTD in levels is nonlinear. Figure 4.6 shows that the relationship is approximately linear on the logarithmic scale and in the modelling that follows logarithms of both HCE and TTD are used throughout.

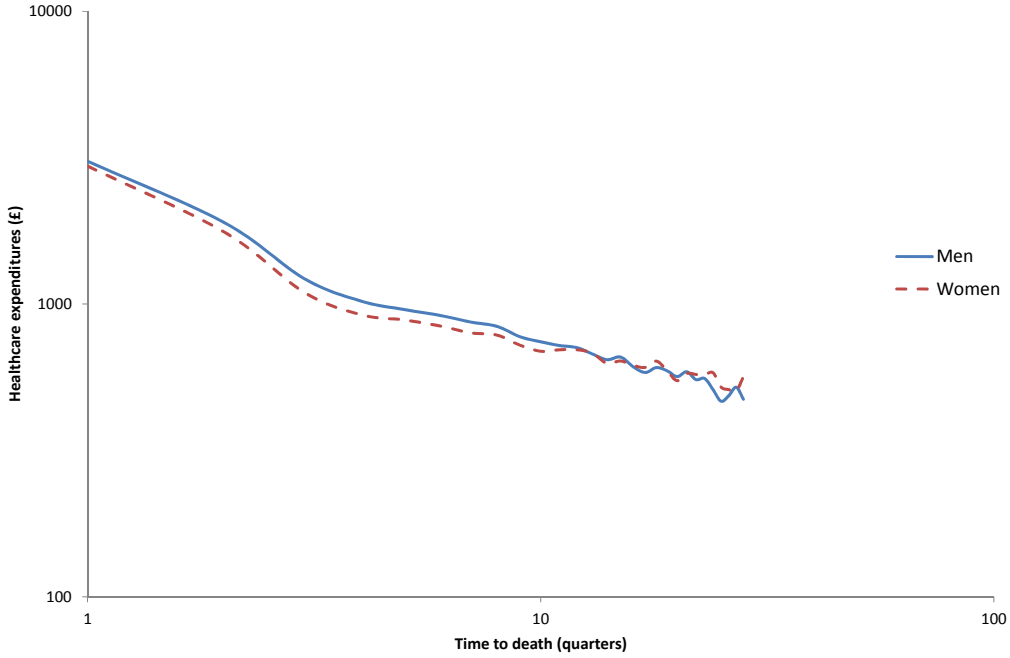

Figure 4.6: Average healthcare expenditures according to quarters to death (log scale for x- and y- axes)

4.3 Theoretical model

This chapter uses a reduced form of the Grossman model in order to more directly consider the meaning of TTD. Consider the Grossman model's conceptualisation of health stock in any given time period (for our purposes, any given quarter), and its conceptualisation of the individual's death. Individuals are assumed to invest in health (I_{it}) – take decisions that, positively or negatively, affect their health – where this investment is a function of medical care (M_{it}) and a vector of exogenous characteristics, such as the individual's level of education (E_{it}).

$$I_{it} = I_{it}\left(M_{it}; E_{it}\right)$$

An individual's health in future time periods is a function of this investment, and health from the current time period (H_{it}), with some depreciation of this due to natural processes, where the depreciation rate is δ_{it}. This depreciation rate is assumed to rise with

age.

$H_{it+1} = I_{it} + (1 - \delta_{it}) H_{it}$

In some final time period, $(t = n)$, the individual's health stock will fall below some level, $H_{\min}$, a minimum level of health stock required to survive.

$H_n < H_{\min}$

We further add the condition that

$M_{it} = f(H_{it}...)$, $\frac{\delta M_{it}}{\delta H_{it}} < 0$

that is, that (in our model, NHS-provided) medical care (and thus HCE) in any given quarter is decreasing in health status.

Death occurs at time period n, and this is affected by I_{it}, $\forall t$, through its effect on H_{it+1}, $\forall t$: at time period n, health stock, H_{it}, falls below $H_{\min}$, the minimum health required to survive.

We observe some (partial) snapshot of a proxy for H_{it} in any given quarter through the morbidity markers recorded in HES, where the presence of morbidity markers proxies for lower levels of health. M_{it} can be seen as our measure of HCE in the current time period. HCE, therefore, affects H_{it+1}: that is, health stock in the next quarter only, and H_{it}, current health stock (as partially proxied by morbidity markers), is therefore not endogenous according to this model.

However, TTD_{it} itself remains endogenous. TTD_{it} (in the original Grossman framework, and previous chapters, n) is increasing in H_{it}, $\forall t$. While M_{it} does not affect H_{it}, it does affect H_{it+1} through I_{it}, and thus TTD. The endogeneity of TTD thus arises from the fact that current HCE influences future health stock, and thus reduces TTD through its positive effect on future health stock. This motivates the use of an instrumental variable approach that instruments TTD with the LSOA measure of YPLL.

4.4 Econometric model

TTD_{it} can only be measured – for decedents – using information available at some time period d, $d > t$, where d is the time period in which individual i is first recorded as dead. In line with existing research into the relationship between HCE and TTD, we use a value for TTD each wave, using information on mortality status known at the end of our final

wave, the financial year 2011-2012. This means, for instance, that an individual who dies in wave 14 would score 14 on TTD in wave 1, 13 in wave 2, and so on, until scoring 1 in wave 14. Owing to the observed relationship between TTD and HCE outlined in the previous chapter, our model for TTD uses a logarithmic transformation of this value.

Existing research in the red herring strand of literature posits a hypothetical model of HCE where

$$HCE_{it} = \alpha + \beta_{age} age_{it} + \mu_i + \varepsilon_{it}, \qquad i = 1, ..., N, t = 1, ..., T_i \quad (1)$$

where μ_i is an individual-specific unobserved effect and ε_{it} is an idiosyncratic error term. Although this model is not estimated in existing papers in the red herring strand of literature, it is claimed that such a model would not adequately explain HCE. TTD is claimed to be an omitted variable in these models, giving rise to models such as:

$$HCE_{it} = \alpha + \beta_{age} age_{it} + \beta_{TTD} TTD_{it} + \mu_i + \varepsilon_{it}, \qquad i = 1, ..., N, t = 1, ..., T_i \quad (2)$$

We argue that individual morbidity is an omitted variable in this type of model. Accordingly, we augment the model as follows:

$$HCE_{it} = \alpha + \beta_{age} age_{it} + \beta_{TTD} TTD_{it} + \sum_{j=1}^{260} \beta_{CCS_j} CCS_{jit} + \mu_i + \varepsilon_{it}, i = 1, ..., N, t = 1, ..., T_i \quad (3)$$

where CCS_n represents a recorded morbidity of CCS type n ($n = 1...260$). We exploit the available data in HES to include detailed information about a patient's morbidities at the time of their hospital stay. We estimate each of these models with random effects.

Modelling HCE as a function of TTD suffers from potential problems of endogeneity. Existing literature suggests that conditional on other covariates, being further from death – i.e. having a high TTD – in time period t are likely to lead to lower levels of HCE in t. This assumes that TTD proxies morbidity which is not observed in the model in question: that those who are healthier are likely to require lower HCE. Higher levels of HCE_{it}, however, are likely to lead to high levels of TTD_{it}: if the hospital activity that

generates healthcare expenditures is effective in improving health then the individual is likely to enjoy a longer remaining lifespan as a result. We therefore posit that actual TTD at time period t has been determined in part by HCE in that time period as well as other time periods.

Other models in the red herring strand of literature model HCE, using TTD and age as explanatory variables, but highlighting this endogeneity problem. Various attempts are made to purge TTD of its endogeneity in HCE (Zweifel et al., 2004; Werblow et al., 2007; Felder et al., 2010). We propose the use of a component of the Health and Disability Index of Multiple Deprivation by Lower Super Output Area – years of potential life lost (YPLL) – as an instrument for TTD. This is highly correlated with TTD and, by virtue of being calculated at an aggregate level, exogenous in a model of HCE. That is, while the level of YPLL at an LSOA level is a strong predictor of an individual's TTD, this YPLL level is not influenced by the HCE for a given individual.

When we estimate an IV model with age, TTD and morbidities, the data suggests a mildly positive relationship between TTD and HCE, though the estimated coefficients are insignificant at even the 5% level.[7]

[7]First-stage regressions in this model for men suggest a positive (but highly non-significant) relationship between TTD and 2007 YPLL figures. We consequently drop 2007 figures from our model and, as a result, are unable to test overidentification restrictions for this model alone.

Table 4.3: Results

Model	RE_AGE	RE_AGE_TTD	RE_AGE_TTD_MORBS	IVRE_AGE_TTD	IVRE_AGE_TTD_MORBS
Men					
First stage regression & diagnostic test results					
2007 IMD YPLL				-.0004**	n/a
2010 IMD YPLL				-.0029***	-.0046***
F-test of instrument relevance				311.67	716.11
Hansen J-statistic (p-value)				0.3354	n/a
Age	0.005	0.0723***	-0.0197***	0.257***	-0.047*
Age^2	0.000	-0.001***	0.000***	-0.002***	0.000
log(TTD)		-0.952***	-0.0427***	-1.826***	0.154
Morbidities			Included		Included
Constant	2.525***	2.145***	1.509***	-1.930***	2.102***
					*N*T=589378, N=55048*
Women					
First stage regression & diagnostic test results					
2007 IMD YPLL				-.0013***	-.0012***
2010 IMD YPLL				-.0007***	-.0008***
F-test of instrument relevance				73.67	63.15
Hansen J-statistic (p-value)				0.6317	0.1256
Age	0.009	0.0774***	-0.0108***	0.549***	-0.010
Age^2	-0.000***	-0.000698***	0.0000307	-0.000***	-0.000
log(TTD)		-0.907***	-0.0252***	-1.767***	0.096
Morbidities			Included		Included
Constant	2.827***	2.225***	1.141***	-11.839***	1.050
					*N*T=610685, N=56104*
					** p<0.05 ** p<0.01 *** p<0.001*

Table 4.4: Diagnoses with highest estimated associated increase in costs, men

CCS grouping	Coefficient estimate, log HCE	
Model	RE_AGE_TTD_MORBS	IVRE_AGE_TTD_MORBS
Leukemia	3.551	3.601
Cancer of brain and nervous system	3.468	3.601
Non-Hodgkin‘s lymphoma	3.399	3.445
Cataract	3.370	3.372
Other non-epithelial cancer of skin	3.339	3.372
Melanomas of skin	3.103	3.123
Osteoarthritis	3.041	3.067
Acute cerebrovascular disease	2.998	3.083
Cancer of pancreas	2.997	3.146
Cancer of bladder	2.929	2.950

Table 4.5: Diagnoses with highest estimated associated costs, women

CCS grouping	Coefficient estimate, log HCE	
Model	RE_AGE_TTD_MORBS	IVRE_AGE_TTD_MORBS
Other non-epithelial cancer of skin	3.707	3.747
Cataract	3.626	3.626
Prolapse of female genital organs	3.592	3.608
Non-Hodgkin‘s lymphoma	3.572	3.598
Leukemia	3.556	3.406
Cancer of brain and nervous system	3.542	3.541
Melanomas of skin	3.408	3.463
Melanomas of skin	3.338	3.445
Cancer of breast	3.163	3.212
Acute cerebrovascular disease	3.141	3.189

4.5 Results

These results (Table 4.3) represent, as far as we are aware, the first reported results in the red herring strand of literature of whether hospital costs increase with age in the aggregate, even before control is made for other factors such as TTD and morbidities (RE_AGE). Existing research broadly states that this is the case, but refer merely to descriptive statistics rather than any kind of econometric analysis. We find a weak and non-significant relationship between age and inpatient costs in a random effects model. Our results are broadly in line with those in the red herring strand of existing research. In a random effects model (RE_AGE_TTD) including TTD and age, we observe a highly significant relationship with TTD. As the individual gets 1% closer to death, HCE increases

by approximately 0.95% for men (0.91% for women)[8]. Confirming the red herring thesis, we find that the relationship between TTD and HCE is much stronger than that between age and HCE: the entire range in which predicted log(HCE) varies with age between the ages of 50 and 100 is equivalent to approximately half the increase in predicted log(HCE) in the final year of life alone.

Conditioning on morbidity markers, we find a reduced role for TTD in explaining HCE. Our estimate on the increase in the TTD elasticity of HCE falls from -0.952 for men (-0.907 for women) (RE_AGE_TTD), to -0.043 (-0.025 for women) when we condition on the individual's observed morbidity in the current time period (RE_AGE_TTD_MORBS). Tables 4.5 and 4.4 list the ten conditions associated with the highest elevation in estimated log HCE, and the estimated associated coefficients for dummy variables for these conditions. In all models – random effects, and random effects with instrumental variables employed – in excess of 90% of the estimated coefficients for the morbidity indicators are significant at the 1% level, and a Wald test of joint significance of coefficients in models for both men and women yields a chi-square value of in excess of 1,000,000, and a p-value reported by Stata 12 as zero. We interpret this as indicating that TTD does indeed serve as a proxy for unobserved morbidity. Further, the estimated coefficients for age see similar falls. This is illustrated in Figure 4.7: the combined relationship of time-to-death and age is severely muted when we condition on current morbidity markers.

[8]Because we aggregate costs by quarter and consequently use discrete values of TTD for each individual in each wave, this elasticity can only be considered as an approximation.

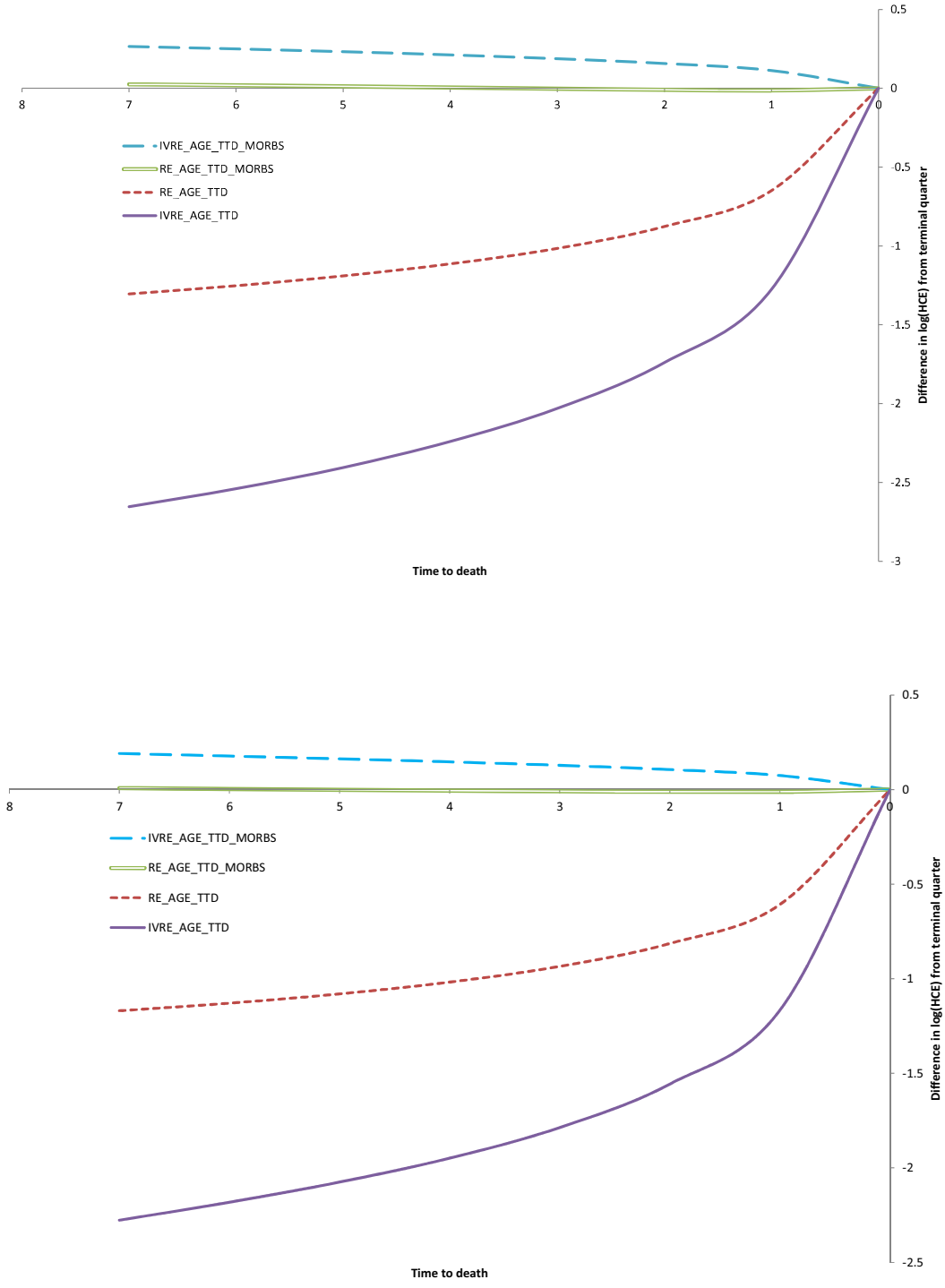

Figure 4.7: Change in HCE according to time-to-death and age, hypothetical individual dying at 75 (top – men, bottom – women)

We anticipate hospital costs to rise as individuals approach death, and as such expect a negative relationship between TTD and HCE. We instrument for TTD in order to deal with the potential endogeneity of TTD in HCE, which would mean that a naïve estimate of the 'effect' of TTD on HCE was likely to be biased towards zero (i.e. that naïve estimates would be expected to be less negative). In a further pair of models, we instrument TTD with YPLL, our small-area measure of premature morbidity.

Previous work (Zweifel et al., 2004; Werblow et al., 2007; Felder et al., 2010) in the red herring strand of literature has highlighted the potential endogeneity of TTD in HCE. While TTD may be negatively associated with HCE through its proxying for morbidity or other events associated with hospital costs in proximity to death, the treatments associated with HCE – if effective – will also increase TTD by increasing the health of the individual receiving treatment. Consequently, if endogeneity does pose problems in this analysis, the coefficient estimate on TTD is likely to be below the true 'effect' of TTD.

When we instrument using both waves of LSOA-level YPLL – IVRE_AGE_TTD – the estimated coefficient of log(TTD) rises (in absolute terms) from -0.952 for men (-0.907 for women) to -1.826 (-1.767 for women). While we confirm the findings of Zweifel et al. (2004) that 'the proximity of death rather than age [being] a main determinant of HCE is fairly robust to endogeneity error,' our results also suggest that failing to account for the endogeneity of TTD in these models leads to a large underestimate of the true 'effect' of TTD in models that do not include morbidity markers. This is also illustrated in Figure 4.7, which shows the large divergence in estimated costs for these two models for an individual who dies at the age of 75. First-stage regressions show, as expected, a negative and significant relationship between YPLL and TTD and an F-test of these instruments in IVRE_AGE_TTD strongly suggests their relevance as a predictor of TTD (an F-statistic of 311.67 for men, and 73.67 for women). Further, a Hansen J-test of overidentification restrictions in this two-instrument model yields a p-value of 0.3354 for men (0.6317 for women), strongly suggesting that these instruments are valid in our model of HCE.

4.6 Conclusions

Ageing populations pose a substantial problem for public service provision, particularly for health and social care. Estimates of how an ageing population will impact HCEs vary considerably. Developing credible predictions is a core component of health systems planning as is allocating resources efficiently and equitably to meet the health care needs of the population. Whilst it is undeniable that health care costs will rise as the baby-boomers age, the impact might not be quite as large as models based on a simple extrapolation of a crude age-expenditure curve suggests. As individuals live longer, all other things equal, they may generate larger cumulative life-time costs. The extent to which this becomes a burden on the health care sector will depend on how morbidity profiles of cohorts change over time. Should a compression of morbidity thesis hold, Fries (1980), Freedman et al. (2002) & Romeu Gordo (2011), on average individuals can expect to live longer and delay the onset of morbidity into later years. This will have the effect of moving the age-expenditure curve to the right as populations age. An expansion of morbidity would have more severe consequences for HCEs with individuals living longer, but also experiencing a greater number of years in ill-health.

Our findings support the red-herring strand of literature that it is not age *per se*, but time-to-death (TTD), particularly the final year of life, that is a strong driver of HCEs. We extend this literature by showing that TTD in large part proxies for morbidity. Our results – showing a weak relationship between HCE and age when TTD is included – fall in line with existing research into the determinants of HCE for ageing populations. However, while TTD clearly plays an important role in explaining HCEs, it is unhelpful in forecasting future expenditure needs. At an individual level TTD is unknown an hence to forecast future expenditure growth assumptions about the proportions of decedents and survivors together with projections of populations within age groups is required. By extending the modelling of HCE to include morbidity characteristics, we show that the impact of TTD is diminished indicating that it acts as a proxy for underlying health status. This is important to allow the planning of future resource requirements and in developing appropriate models for budgets to be allocated equitably across providers of care in response to population health care need. Our results are robust to problems of

endogeneity that exist between HCE and TTD.

Our results strengthen the need to include measures of morbidity in models of HCE. Merely including TTD is insufficient in predicting future HCE. To accurately forecast future expenditure needs, information on changes to profiles of morbidity are required. The existence of a compression of morbidity, along with a tendency for increased life expectancy, suggests competing and opposing pressures on HCE. While increases in life expectancy suggests that a greater number of individuals will be alive at any given age, with associated upward pressure on HCE, a compression of morbidity will tend to, on average, provide downward pressure on HCE for any given individual at any given age.

This work has focused on determinants of the demand for inpatient health care services at an individual level via age, time-to-death and morbidity characteristics. Clearly there is also a substantial role for supply-side impacts on expenditure growth notably through technological advances in health care interventions and the way in which health care services are organized and delivered. We do not address these issues here, but are areas that warrant further investigation at an aggregate level. Inpatient hospital care is one of a number of services provided by the National Health Service in England and other expenditure should also be taken into account when assessing the overall impact of an ageing population, as should costs placed on the Government by long-term care services predominantly accessed by older age groups. The increasing ability to link administrative sources of data provides a potentially valuable resource for future research in this area.

4.A Appendix

Table A4.1: Clinical Classifications Software (CCS) groupings

CCS code	Description
1	Tuberculosis
2	Septicemia (except in labor)
3	Bacterial infection; unspecified site
4	Mycoses
5	HIV infection
6	Hepatitis
7	Viral infection
8	Other infections; including parasitic
9	Sexually transmitted infections (not HIV or hepatitis)
10	Immunizations and screening for infectious disease
11	Cancer of head and neck
12	Cancer of esophagus
13	Cancer of stomach
14	Cancer of colon
15	Cancer of rectum and anus
16	Cancer of liver and intrahepatic bile duct
17	Cancer of pancreas
18	Cancer of other GI organs; peritoneum
19	Cancer of bronchus; lung
20	Cancer; other respiratory and intrathoracic
21	Cancer of bone and connective tissue
22	Melanomas of skin
23	Other non-epithelial cancer of skin
24	Cancer of breast
25	Cancer of uterus
26	Cancer of cervix
27	Cancer of ovary
28	Cancer of other female genital organs
29	Cancer of prostate
30	Cancer of testis
31	Cancer of other male genital organs
32	Cancer of bladder
33	Cancer of kidney and renal pelvis
34	Cancer of other urinary organs
35	Cancer of brain and nervous system
36	Cancer of thyroid
37	Hodgkin‘s disease
38	Non-Hodgkin‘s lymphoma
39	Leukemias
40	Multiple myeloma
41	Cancer; other and unspecified primary
42	Secondary malignancies
43	Malignant neoplasm without specification of site
44	Neoplasms of unspecified nature or uncertain behavior
45	Maintenance chemotherapy; radiotherapy
46	Benign neoplasm of uterus
47	Other and unspecified benign neoplasm
48	Thyroid disorders

Continued on next page

49	Diabetes mellitus without complication
50	Diabetes mellitus with complications
51	Other endocrine disorders
52	Nutritional deficiencies
53	Disorders of lipid metabolism
54	Gout and other crystal arthropathies
55	Fluid and electrolyte disorders
56	Cystic fibrosis
57	Immunity disorders
58	Other nutritional; endocrine; and metabolic disorders
59	Deficiency and other anemia
60	Acute posthemorrhagic anemia
61	Sickle cell anemia
62	Coagulation and hemorrhagic disorders
63	Diseases of white blood cells
64	Other hematologic conditions
65	Mental retardation
66	Alcohol-related mental disorders
67	Substance-related mental disorders
68	Senility and organic mental disorders
69	Affective disorders
70	Schizophrenia and related disorders
71	Other psychoses
72	Anxiety; somatoform; dissociative; and personality disorders
73	Preadult disorders
74	Other mental conditions
75	Personal history of mental disorder; mental and behavioral problems; observation and screening for mental condition
76	Meningitis (except that caused by tuberculosis or sexually transmitted disease)
77	Encephalitis (except that caused by tuberculosis or sexually transmitted disease)
78	Other CNS infection and poliomyelitis
79	Parkinson‘s disease
80	Multiple sclerosis
81	Other hereditary and degenerative nervous system conditions
82	Paralysis
83	Epilepsy; convulsions
84	Headache; including migraine
85	Coma; stupor; and brain damage
86	Cataract
87	Retinal detachments; defects; vascular occlusion; and retinopathy
88	Glaucoma
89	Blindness and vision defects
90	Inflammation; infection of eye (except that caused by tuberculosis or sexually transmitteddisease)
91	Other eye disorders
92	Otitis media and related conditions
93	Conditions associated with dizziness or vertigo
94	Other ear and sense organ disorders
95	Other nervous system disorders
96	Heart valve disorders
97	Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or sexually transmitted disease)

Continued on next page

98	Essential hypertension
99	Hypertension with complications and secondary hypertension
100	Acute myocardial infarction
101	Coronary atherosclerosis and other heart disease
102	Nonspecific chest pain
103	Pulmonary heart disease
104	Other and ill-defined heart disease
105	Conduction disorders
106	Cardiac dysrhythmias
107	Cardiac arrest and ventricular fibrillation
108	Congestive heart failure; nonhypertensive
109	Acute cerebrovascular disease
110	Occlusion or stenosis of precerebral arteries
111	Other and ill-defined cerebrovascular disease
112	Transient cerebral ischemia
113	Late effects of cerebrovascular disease
114	Peripheral and visceral atherosclerosis
115	Aortic; peripheral; and visceral artery aneurysms
116	Aortic and peripheral arterial embolism or thrombosis
117	Other circulatory disease
118	Phlebitis; thrombophlebitis and thromboembolism
119	Varicose veins of lower extremity
120	Hemorrhoids
121	ther diseases of veins and lymphatics
122	Pneumonia (except that caused by tuberculosis or sexually transmitted disease)
123	Influenza
124	Acute and chronic tonsillitis
125	Acute bronchitis
126	Other upper respiratory infections
127	Chronic obstructive pulmonary disease and bronchiectasis
128	Asthma
129	Aspiration pneumonitis; food/vomitus
130	Pleurisy; pneumothorax; pulmonary collapse
131	Respiratory failure; insufficiency; arrest (adult)
132	Lung disease due to external agents
133	Other lower respiratory disease
134	Other upper respiratory disease
135	Intestinal infection
136	Disorders of teeth and jaw
137	Diseases of mouth; excluding dental
138	Esophageal disorders
139	Gastroduodenal ulcer (except hemorrhage)
140	Gastritis and duodenitis
141	Other disorders of stomach and duodenum
142	Appendicitis and other appendiceal conditions
143	Abdominal hernia
144	Regional enteritis and ulcerative colitis
145	Intestinal obstruction without hernia
146	Diverticulosis and diverticulitis
147	Anal and rectal conditions
148	Peritonitis and intestinal abscess
149	Biliary tract disease
150	Liver disease; alcohol-related

Continued on next page

151	Other liver diseases
152	Pancreatic disorders (not diabetes)
153	Gastrointestinal hemorrhage
154	Noninfectious gastroenteritis
155	Other gastrointestinal disorders
156	Nephritis; nephrosis; renal sclerosis
157	Acute and unspecified renal failure
158	Chronic renal failure
159	Urinary tract infections
160	Calculus of urinary tract
161	Other diseases of kidney and ureters
162	Other diseases of bladder and urethra
163	Genitourinary symptoms and ill-defined conditions
164	Hyperplasia of prostate
165	Inflammatory conditions of male genital organs
166	Other male genital disorders
167	Nonmalignant breast conditions
168	Inflammatory diseases of female pelvic organs
169	Endometriosis
170	Prolapse of female genital organs
171	Menstrual disorders
172	Ovarian cyst
173	Menopausal disorders
174	Female infertility
175	Other female genital disorders
176	Contraceptive and procreative management
177	Spontaneous abortion
178	Induced abortion
179	Postabortion complications
180	Ectopic pregnancy
181	Other complications of pregnancy
182	Hemorrhage during pregnancy; abruptio placenta; placenta previa
183	Hypertension complicating pregnancy; childbirth and the puerperium
184	Early or threatened labor
185	Prolonged pregnancy
186	Diabetes or abnormal glucose tolerance complicating pregnancy; childbirth; or the puerperium
187	Malposition; malpresentation
188	Fetopelvic disproportion; obstruction
189	Previous C-section
190	Fetal distress and abnormal forces of labor
191	Polyhydramnios and other problems of amniotic cavity
192	Umbilical cord complication
193	OB-related trauma to perineum and vulva
194	Forceps delivery
195	Other complications of birth; puerperium affecting management of mother
196	Normal pregnancy and/or delivery
197	Skin and subcutaneous tissue infections
198	Other inflammatory condition of skin
199	Chronic ulcer of skin
200	Other skin disorders

Continued on next page

201	Infective arthritis and osteomyelitis (except that caused by tuberculosis or sexually transmitted disease)
202	Rheumatoid arthritis and related disease
203	Osteoarthritis
204	Other non-traumatic joint disorders
205	Spondylosis; intervertebral disc disorders; other back problems
206	Osteoporosis
207	Pathological fracture
208	Acquired foot deformities
209	Other acquired deformities
210	Systemic lupus erythematosus and connective tissue disorders
211	Other connective tissue disease
212	Other bone disease and musculoskeletal deformities
213	Cardiac and circulatory congenital anomalies
214	Digestive congenital anomalies
215	Genitourinary congenital anomalies
216	Nervous system congenital anomalies
217	Other congenital anomalies
218	Liveborn
219	Short gestation; low birth weight; and fetal growth retardation
220	Intrauterine hypoxia and birth asphyxia
221	Respiratory distress syndrome
222	Hemolytic jaundice and perinatal jaundice
223	Birth trauma
224	Other perinatal conditions
225	Joint disorders and dislocations; trauma-related
226	Fracture of neck of femur (hip)
227	Spinal cord injury
228	Skull and face fractures
229	Fracture of upper limb
230	Fracture of lower limb
231	Other fractures
232	Sprains and strains
233	Intracranial injury
234	Crushing injury or internal injury
235	Open wounds of head; neck; and trunk
236	Open wounds of extremities
237	Complication of device; implant or graft
238	Complications of surgical procedures or medical care
239	Superficial injury; contusion
240	Burns
241	Poisoning by psychotropic agents
242	Poisoning by other medications and drugs
243	Poisoning by nonmedicinal substances
244	Other injuries and conditions due to external causes
245	Syncope
246	Fever of unknown origin
247	Lymphadenitis
248	Gangrene
249	Shock
250	Nausea and vomiting
251	Abdominal pain
252	Malaise and fatigue
253	Allergic reactions

Continued on next page

254	Rehabilitation care; fitting of prostheses; and adjustment of devices
255	Administrative/social admission
256	Medical examination/evaluation
257	Other aftercare
258	Other screening for suspected conditions (not mental disorders or infectious disease)
259	Residual codes; unclassified
260	E Codes: All (external causes of injury and poisoning)

References

Blane, D. and Drever, F. (1998), 'Inequality among men in standardised years of potential life lost, 1970-93', *BMJ* **317**(7153), 255–256.

Breyer, F., Lorenz, N. and Niebel, T. (2014), 'Health care expenditures and longevity: is there a Eubie Blake effect?', *The European Journal of Health Economics* pp. (forthcoming, Early View available).

Cracknell, R. (2010), *The ageing population*, Briefing Papers, House of Commons Library, United Kingdom.

Cutler, D. M., Ghosh, K. and Landrum, M. B. (2013), Evidence for significant compression of morbidity in the elderly U.S. population, Working Paper 19268, National Bureau of Economic Research.

Dang, T., Antolin, P. and Oxley, H. (2001), *Fiscal Implications of Ageing: projections of age-related spending*, Economics Department Working Papers No. 305, Organisation for Economic Co-operation and Development, Paris.

Dawson, D., Gravelle, H., O'Mahony, M., Street, A., Weale, M., Castelli, A., Jacobs, R., Kind, P., Loveridge, P. and Martin, S. (2005), 'Developing new approaches to measuring NHS outputs and productivity', *Summary Report. CHE Research Paper* **6**.

de Meijer, C., Koopmanschap, M., d'Uva, T. B. and van Doorslaer, E. (2011), 'Determinants of long-term care spending: Age, time to death or disability?', *Journal of Health Economics* **30**(2), 425–438.

Department for Communities and Local Government (2008), 'The English Indices of Deprivation 2007', http://communities.gov.uk/documents/communities/pdf/733520.pdf.

Department for Communities and Local Government (2011), 'The English Indices of Deprivation 2010', https://www.gov.uk/government/uploads/system/uploads/attachment_data /file/6320/1870718.pdf.

Department of Health (2012), 'Reference costs 2011-12', https://www.gov.uk/government/uploads/system/uploads/attachment_data /file/213060/2011-12-reference-costs-publication.pdf.

Dixon, J., Smith, P., Gravelle, H., Martin, S., Bardsley, M., Rice, N., Georghiou, T., Dusheiko, M., Billings, J., De Lorenzo, M. and Sanderson, C. (2011), 'A person based formula for allocating commissioning funds to general practices in England: development of a statistical model', *BMJ* **343**.

Felder, S., Werblow, A. and Zweifel, P. (2010), 'Do red herrings swim in circles? controlling for the endogeneity of time to death', *Journal of Health Economics* **29**(2), 205–212.

Freedman, V. A., Martin, L. G. and Schoeni, R. F. (2002), 'Recent trends in disability and functioning among older adults in the United States: a systematic review', *Journal of the American Medical Association* **288**(24), 3137–3146.

Fries, J. F. (1980), 'Aging, natural death, and the compression of morbidity', *The New England Journal of Medicine* **303**(3), 130–135. PMID: 7383070.

Fries, J. F. (2005), 'The compression of morbidity', *The Milbank Quarterly* **83**(4), 801–823. PMID: 16279968 PMCID: PMC2690269.

Gray, A. (2005), 'Population ageing and health care expenditure', *Ageing Horizons* (2), 15–20.

Gruenberg, E. M. (2005), 'The failures of success', *The Milbank Quarterly* **83**(4), 779–800.

Harker, R. (2012), *NHS funding and expenditure*, Standard Note, House of Commons Library, United Kingdom.

Health & Social Care Information Centre (2013), 'Hospital Episode Statistics (HES) Analysis Guide', http://www.hscic.gov.uk/media/1592/HES-analysis-guide/pdf/HES_Analysis_Guide_Jan_2014.pdf.

Heller, P. S., Hemming, R., Kohnert, P. and Feldman, R. A. (1986), *Aging and Social Expenditure in the Major Industrial Countries, 1980-2025*, International Monetary Fund.

Jacobzone, S., Cambois, E. and Robine, J. (2000), 'Is the health of older persons in oecd countries improving fast enough to compensate for population ageing?', *OECD Economic Studies* **30**.

Jonsen, A. R. (1986), 'Bentham in a box: Technology assessment and health care allocation', *The Journal of Law, Medicine & Ethics* **14**(3-4), 172–174.

Lansley, H. A. (2012), 'Writing exclusively for BJN, Lansley dismisses Bill critics', *British Journal of Nursing* **21**(6), 333–333.

Office for National Statistics (2011), 'Super Output Areas (SOAs)', http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas–soas-/index.html. Super output areas are a geography designed to improve the reporting of small area statistics.

Romeu Gordo, L. (2011), 'Compression of morbidity and the labour supply of older people', *Applied Economics* **43**(4), 503–513.

Seshamani, M. and Gray, A. (2004), 'Ageing and health-care expenditure: the red herring argument revisited', *Health Economics* **13**(4), 303–314.

Shwartz, M., Iezzoni, L. I., Moskowitz, M. A., Ash, A. S. and Sawitz, E. (1996), 'The importance of comorbidities in explaining differences in patient costs', *Medical Care* **34**(8), 767–782.

Stearns, S. C. and Norton, E. C. (2004), 'Time to include time to death? The future of health care expenditure predictions', *Health Economics* **13**(4), 315–327.

US Agency for Healthcare Research and Quality (2009), 'Clinical classifications software (CCS) for ICD-10', http://www.hcup-us.ahrq.gov/toolssoftware/icd_10/ccs_icd_10.jsp.

Werblow, A., Felder, S. and Zweifel, P. (2007), 'Population ageing and health care expenditure: a school of 'red herrings'?', *Health Economics* **16**(10), 1109–1126.

Williams, D. (2012), 'Lansley: CCG allocations should be based on age, not poverty', *Health Services Journal* .

Zweifel, P., Felder, S. and Meiers, M. (1999), 'Ageing of population and health care expenditure: a red herring?', *Health Economics* **8**(6), 485–496.

Zweifel, P., Felder, S. and Werblow, A. (2004), 'Population ageing and health care expenditure: New evidence on the "Red Herring"', *Geneva Papers on Risk and Insurance - Issues and Practice* **29**(4), 652–666.

Conclusions

This thesis focuses on the longitudinal analysis of health and healthcare data, applied to three different questions, making methodological innovations and producing results that should inform both policy and future empirical work. The thesis as a whole examines three topics, the analysis of each being underpinned by a Grossman framework: social class inequalities in cancer registration, the effect of unobserved heterogeneity on the relationship between smoking behaviours and cancer outcomes, and the implications for healthcare expenditures from demographic changes in the presence of compression of morbidity, when time-to-death is considered directly as a potential proxy for individual morbidity. When these three questions are considered using the novel methods employed in this thesis, substantially different results and implications arise. While imperfections in the data sources employed necessarily mean that further work will be required to pin down causal effects with fewer required assumptions, all three chapters point to the use of these novel methodological approaches in order to more correctly answer existing questions. The remainder of this section provides the specific conclusions to the research questions considered in each chapter, discusses the necessary limitations of the conclusions drawn, and provides a view towards areas for future research.

While previous research finds weaker evidence of socioeconomic inequality in cancer outcomes after controlling for smoking behaviours, Chapter 2 suggests that this can be attributed to the incorrect modelling of cancer outcomes. Different methods and different assumptions regarding the relationship between social class and cancer outcomes are considered, using cross-sectional data, as well as subsequent information regarding individuals' health status. Finally, even after stripping out the effect of smoking, we find that, while evidence regarding differential rates of lifetime incidence of cancer remains weak, individuals in the lowest social classes develop cancer approximately 15% sooner than those

in the highest social class.

The chapter develops a case for a future research agenda in the field of socioeconomic inequality in diseases more generally, and cancers specifically. While smokers are dropped from the analysis in this chapter, the issue of the joint-modelling of smoking behaviours and health outcomes is revisited in Chapter 3. While the use of duration analysis is common in, for instance, cancer outcomes post-diagnosis, this chapter highlights the need to also consider duration analysis, rather than merely probability models, of cancer outcomes in an observation period that begins prior to diagnosis. Further, the imperfections of the Health and Lifestyle Survey dataset employed in this chapter – a single cross-section taken approximately 25 years prior to the latest follow-up – mean that prospects for future research, and confirmation of these results – could lie in the use of better, ideally panel, data to confirm these results. In addition to qualitative confirmation of results, the use of a better dataset would derive quantitative estimates that relied upon fewer assumptions, and could be estimated using a perhaps more appropriate measure of social inequality (class-based, or income-based) than the registrar general classification chosen here for pragmatic reasons.

Chapter 3 builds on the work of Chapter 2, continuing the focus on cancer outcomes. We simultaneously model smoking behaviours and health outcomes – death and cancer – allowing these to be jointly affected by unobservable factors. We find that the joint modelling of smoking behaviours and health outcomes, compared to the single-equation modelling of cancer outcomes, substantially alters the results obtained, suggesting a large role for unobservable factors. This is confirmed by a post-estimation prediction of the probability of class membership for each individual, in which we find that individuals with similar observable characteristics and smoking behaviours exhibit substantially different health outcomes.

The chapter emphasises the need for the joint modelling of individual behaviours and health outcomes where, as in this case, those behaviours are likely to substantially affect outcomes. The potential for unobservable factors to jointly affect individual behaviours and health outcomes should not be ignored, and adopting a modelling method that does not permit outcomes to vary by these unobservable factors is apt to incorrectly estimate causal relationships between smoking and cancer outcomes. Again, the dataset employed

(a single cross-section of individuals in 1985) complicates the exact interpretation of parameter estimates: our count of pack-years is censored in 1985 and thus does not capture a full spell of the individual's smoking behaviours. Clearer interpretations of coefficient estimates of pack-years could be gained if a full spell of the individual's smoking history, rather than one truncated almost 25 years previous to the latest follow-up, was observed.

Chapter 4 focuses on the relationship between age, proximity to death (often time-to-death, or TTD), and morbidity. While existing research examines the link between age, time-to-death and healthcare expenditures (HCE), finding that TTD rather than age determines HCE, we also condition on individual morbidities. By doing so, we find that TTD itself is a 'red herring' in explaining HCE: that what determines HCE is not TTD, but individuals' morbidities.

This chapter has important implications for future research. If the ultimate data-generating process that links TTD and HCE changes – if individuals approach death in a different health state, as the 'compression of morbidity' hypothesis suggests – then assumptions about future HCE may turn out to be incorrect. The inclusion of TTD in models of HCE, as in previous research in the 'red herring' strand of literature, is found to be insufficient in forming predictions about future HCE. The importance of including predictions about – and uncertainty around – future trends in morbidity in estimating future HCE is emphasised: predictions about future health expenditures must incorporate estimates of future changes in morbidity profiles. This has clear policy relevance with regard to the sustainability of financing healthcare for an ageing population, and methodological implications for the modelling of future HCE. While these results are not specifically intended to inform resource allocation methods and would require further assumptions and/or adjustments to be applied to such a use, our results would tentatively confirm previous findings that linking healthcare funding to the age of the local population, as proposed by a recent Secretary of State for Health, would not adequately capture need.

Future research in this area would benefit from the use of a longer panel, which would allow the compression of morbidity hypothesis to be directly tested: the (maximum of) 28 quarters observed in our dataset is unlikely to be sufficient in order to do so. Furthermore, this chapter concentrates solely on hospital inpatient care, which forms only part of NHS expenditure. Future research could use the framework (of modelling according to age, time-

to-death, and morbidity characteristics) adopted in this chapter to consider determinants of other areas of NHS expenditure, as well as other possible effects on long-term social care.

Bibliography

Bibliography

Adda, J. and Lechene, V. (2013), 'Health selection and the effect of smoking on mortality', *The Scandinavian Journal of Economics* **115**(3), 902–931.

Anand, P., Kunnumakara, A. B., Sundaram, C., Harikumar, K. B., Tharakan, S. T., Lai, O. S., Sung, B. and Aggarwal, B. B. (2008), 'Cancer is a preventable disease that requires major lifestyle changes', *Pharmaceutical Research* **25**(9), 2097–2116.

Balia, S. and Jones, A. M. (2008), 'Mortality, lifestyle and socio-economic status', *Journal of Health Economics* **27**(1), 1–26.

Balia, S. and Jones, A. M. (2011), 'Catching the habit: a study of inequality of opportunity in smoking-related mortality', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**(1), 175–194.

Banks, J., Marmot, M., Oldfield, Z. and Smith, J. P. (2006), 'Disease and disadvantage in the United States and in England', *JAMA: The Journal of the American Medical Association* **295**(17), 2037–2045.

Blane, D. and Drever, F. (1998), 'Inequality among men in standardised years of potential life lost, 1970-93', *BMJ* **317**(7153), 255–256.

Böckerman, P. and Ilmakunnas, P. (2009), 'Unemployment and self-assessed health: evidence from panel data', *Health Economics* **18**(2), 161–179.

Breyer, F., Lorenz, N. and Niebel, T. (2014), 'Health care expenditures and longevity: is there a Eubie Blake effect?', *The European Journal of Health Economics* pp. (forthcoming, Early View available).

Brown, H. and van der Pol, M. (2010), Intergenerational transfer of health: the role of time and risk preferences, *in* 'Health Economists Study Group 2011 (Winter)', York.

Contoyannis, P. and Jones, A. M. (2004), 'Socio-economic status, health and lifestyle', *Journal of Health Economics* **23**(5), 965–995.

Cox, B., Blaxter, M., Buckle, A., Fenner, N. P., Golding, J., Gore, M., Huppert, F., Nickson, J., Roth, M., Stark, J. et al. (1987), *The Health and Lifestyle Survey. Preliminary report of a nationwide survey of the physical and mental health, attitudes and lifestyle of a random sample of 9,003 British adults.*, Health Promotion Research Trust, London.

Cox, B., Huppert, F. and Whichelow, M., eds (1993), *The Health and Lifestyle Survey: Seven Years on: A Longitudinal Study of a Nationwide Sample, Measuring Changes in Physical and Mental Health, Attitudes and Lifestyle*, Dartmouth Pub. Co, Aldershot, Hants., England.

Cox, C., Chu, H., Schneider, M. F. and Muñoz, A. (2007), 'Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution', *Statistics in Medicine* **26**(23), 4352–4374.

Cracknell, R. (2010), *The ageing population*, Briefing Papers, House of Commons Library, United Kingdom.

Cutler, D. (2008), 'Are we finally winning the war on cancer?', *Journal of Economic Perspectives* **22**(4), 3–26.

Cutler, D., Deaton, A. and Lleras-Muney, A. (2006), 'The determinants of mortality', *The Journal of Economic Perspectives* **20**(3), 97–120.

Cutler, D. M., Ghosh, K. and Landrum, M. B. (2013), Evidence for significant compression of morbidity in the elderly U.S. population, Working Paper 19268, National Bureau of Economic Research.

Dalstra, J., Kunst, A., Borrell, C., Breeze, E., Cambois, E., Costa, G., Geurts, J., Lahelma, E., Van Oyen, H., Rasmussen, N., Regidor, E., Spadea, T. and Mackenbach, J. (2005), 'Socioeconomic differences in the prevalence of common chronic diseases: an overview of eight European countries', *International Journal of Epidemiology* **34**(2), 316 –326.

Dang, T., Antolin, P. and Oxley, H. (2001), *Fiscal Implications of Ageing: projections of age-related spending*, Economics Department Working Papers No. 305, Organisation for Economic Co-operation and Development, Paris.

Dawson, D., Gravelle, H., O'Mahony, M., Street, A., Weale, M., Castelli, A., Jacobs, R., Kind, P., Loveridge, P. and Martin, S. (2005), 'Developing new approaches to measuring NHS outputs and productivity', *Summary Report. CHE Research Paper* **6**.

de Meijer, C., Koopmanschap, M., d'Uva, T. B. and van Doorslaer, E. (2011), 'Determinants of long-term care spending: Age, time to death or disability?', *Journal of Health Economics* **30**(2), 425–438.

Deaton, A. (2002), 'Policy implications of the gradient of health and wealth', *Health Affairs* **21**(2), 13–30.

Department for Communities and Local Government (2008), 'The English Indices of Deprivation 2007', http://communities.gov.uk/documents/communities/pdf/733520.pdf.

Department for Communities and Local Government (2011), 'The English Indices of Deprivation 2010', https://www.gov.uk/government/uploads/system/uploads/attachment_data /file/6320/1870718.pdf.

Department of Health (2012), 'Reference costs 2011-12', https://www.gov.uk/government/uploads/system/uploads/attachment_data /file/213060/2011-12-reference-costs-publication.pdf.

Department of Health and Human Services (1989), *Reducing the health consequences of smoking. 25 years of progress*, US Government Printing Office, Washington, DC.

Department of Health and Human Services (2004), 'Smoking and Tobacco Use; 2004 Surgeon General's Report', http://www.cdc.gov/tobacco/data_statistics/sgr/2004/index.htm.

Department of Health and Human Services (2010), 'A Report of the Surgeon General: How Tobacco Smoke Causes Disease', http://www.surgeongeneral.gov/library/tobaccosmoke/report/index.html.

Dixon, J., Smith, P., Gravelle, H., Martin, S., Bardsley, M., Rice, N., Georghiou, T., Dusheiko, M., Billings, J., De Lorenzo, M. and Sanderson, C. (2011), 'A person based formula for allocating commissioning funds to general practices in England: development of a statistical model', *BMJ* **343**.

Doll, R. (1998), 'Uncovering the effects of smoking: historical perspective', *Statistical Methods in Medical Research* **7**(2), 87–117.

Doll, R. and Hill, A. B. (1954), 'The mortality of doctors in relation to their smoking habits: a preliminary report', *BMJ* **328**(7455), 1529–1533.

Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2004), 'Mortality in relation to smoking: 50 years' observations on male British doctors', *BMJ* **328**(7455), 1519.

Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2005), 'Mortality in relation to alcohol consumption: a prospective study among male British doctors', *Int. J. Epidemiol.* **34**(1), 199–204.

Doll, R., Peto, R., Hall, E., Wheatley, K. and Gray, R. (1994), 'Mortality in relation to consumption of alcohol: 13 years' observations on male British doctors', *BMJ* **309**(6959), 911–918.

Douglas, S. and Hariharan, G. (1994), 'The hazard of starting smoking: Estimates from a split population duration model', *Journal of Health Economics* **13**(2), 213–230.

Estève, J., Benhamou, E. and Raymond, L. (1994), 'Statistical methods in cancer research: descriptive epidemiology.', *IARC scientific publications* **IV**(128), 1.

Felder, S., Werblow, A. and Zweifel, P. (2010), 'Do red herrings swim in circles? controlling for the endogeneity of time to death', *Journal of Health Economics* **29**(2), 205–212.

Forster, M. and Jones, A. M. (2001), 'The role of tobacco taxes in starting and quitting smoking: Duration analysis of british data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **164**(3), 517–547.

Freedman, V. A., Martin, L. G. and Schoeni, R. F. (2002), 'Recent trends in disability and functioning among older adults in the United States: a systematic review', *Journal of the American Medical Association* **288**(24), 3137–3146.

Fries, J. F. (1980), 'Aging, natural death, and the compression of morbidity', *The New England Journal of Medicine* **303**(3), 130–135. PMID: 7383070.

Fries, J. F. (2005), 'The compression of morbidity', *The Milbank Quarterly* **83**(4), 801–823. PMID: 16279968 PMCID: PMC2690269.

Gordo, L. R. (2006), 'Effects of short- and long-term unemployment on health satisfaction: evidence from German data', *Applied Economics* **38**(20), 2335.

Gray, A. (2005), 'Population ageing and health care expenditure', *Ageing Horizons* (2), 15–20.

Grossman, M. (1972), 'On the concept of health capital and the demand for health', *The Journal of Political Economy* **80**(2), 223–255.

Gruenberg, E. M. (2005), 'The failures of success', *The Milbank Quarterly* **83**(4), 779–800.

Gutierrez, R. G. (2002), 'Parametric frailty and shared frailty survival models', *Stata Journal* **2**(1), 22–44.

Harker, R. (2012), *NHS funding and expenditure*, Standard Note, House of Commons Library, United Kingdom.

Health & Social Care Information Centre (2013), 'Hospital Episode Statistics (HES) Analysis Guide', http://www.hscic.gov.uk/media/1592/HES-analysis-guide/pdf/HES_Analysis_Guide_Jan_2014.pdf.

Heckman, J. and Singer, B. (1984), 'A method for minimizing the impact of distributional assumptions in econometric models for duration data', *Econometrica* **52**(2), 271–320.

Heller, P. S., Hemming, R., Kohnert, P. and Feldman, R. A. (1986), *Aging and Social Expenditure in the Major Industrial Countries, 1980-2025*, International Monetary Fund.

Hiatt, R. A. and Breen, N. (2008), 'The social determinants of cancer: A challenge for transdisciplinary science', *American Journal of Preventive Medicine* **35**(2, Supplement 1), S141–S150.

Honoré, B. E. and Lleras-Muney, A. (2006), 'Bounds in competing risks models and the war on cancer', *Econometrica* **74**(6), 1675–1698.

Jacobzone, S., Cambois, E. and Robine, J. (2000), 'Is the health of older persons in oecd countries improving fast enough to compensate for population ageing?', *OECD Economic Studies* **30**.

Jones, A. M., Rice, N., Bago d'Uva, T. and Balia, S. (2007), *Applied Health Economics*, Routledge advanced texts in economics and finance, Routledge, London.

Jonsen, A. R. (1986), 'Bentham in a box: Technology assessment and health care allocation', *The Journal of Law, Medicine & Ethics* **14**(3-4), 172–174.

Kaprio, J. and Koskenvuo, M. (1989), 'Twins, smoking and mortality: A 12-year prospective study of smoking-discordant twin pairs', *Social Science & Medicine* **29**(9), 1083–1089.

Krall, E. A., Valadian, I., Dwyer, J. T. and Gardner, J. (1989), 'Accuracy of recalled smoking data', *American Journal of Public Health* **79**(2), 200–202.

Laing, R. D. (1967), *The Politics of Experience and the Bird of Paradise*, Penguin UK.

Lansley, H. A. (2012), 'Writing exclusively for BJN, Lansley dismisses Bill critics', *British Journal of Nursing* **21**(6), 333–333.

Link, B. G. and Phelan, J. (1995), 'Social conditions as fundamental causes of disease', *Journal of Health and Social Behavior* **35**, 80–94.

Lyratzopoulos, G., Abel, G. A., Brown, C. H., Rous, B. A., Vernon, S. A., Roland, M. and Greenberg, D. C. (2012), 'Socio-demographic inequalities in stage of cancer diagnosis: evidence from patients with female breast, lung, colon, rectal, prostate, renal, bladder, melanoma, ovarian and endometrial cancer', *Annals of Oncology* p. mds526.

Mackillop, W. J., Zhang-Salomons, J., Boyd, C. J. and Groome, P. A. (2000), 'Associations between community income and cancer incidence in Canada and the United States', *Cancer* **89**(4), 901–912.

Mao, Y., Hu, J., Ugnat, A. M., Semenciw, R., Fincham, S. and Canadian Cancer Registries Epidemiology Research Group (2001), 'Socioeconomic status and lung cancer risk in Canada', *International Journal of Epidemiology* **30**(4), 809–817.

Marmot, M. (2007), 'Achieving health equity: from root causes to fair outcomes', *The Lancet* **370**(9593), 1153–1163.

Marmot, M., Rose, G., Shipley, M. and Hamilton, P. J. (1978), 'Employment grade and coronary heart disease in British civil servants', *Journal of Epidemiology and Community Health* **32**(4), 244–249.

Marmot, M., Smith, G. D., Stansfeld, S., Patel, C., North, F., Head, J., White, I., Brunner, E. and Feeney, A. (1991), 'Health inequalities among British civil servants: the Whitehall II study', *The Lancet* **337**(8754), 1387–1393.

Mathers, C. D. and Schofield, D. J. (1998), 'The health consequences of unemployment: the evidence.', *The Medical Journal of Australia* **168**(4), 178.

Mroz, T. A. (1999), 'Discrete factor approximations in simultaneous equation models: Estimating the impact of a dummy endogenous variable on a continuous outcome', *Journal of Econometrics* **92**(2), 233–274.

Nkosi, T. M., Parent, M.-l., Siemiatycki, J. and Rousseau, M.-C. (2012), 'Socioeconomic position and lung cancer risk: how important is the modeling of smoking?', *Epidemiology (Cambridge, Mass.)* **23**(3), 377–385.

Office for National Statistics (2011), 'Super Output Areas (SOAs)', http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas–soas-/index.html. Super output areas are a geography designed to improve the reporting of small area statistics.

Peto, R., Lopez, A., Boreham, J. and Thun, M. (2006), 'Mortality from smoking in developed countries, 1950-2000'.

Reid, A., de Klerk, N. H., Ambrosini, G. L., Berry, G. and Musk, A. W. (2006), 'The risk of lung cancer with increasing time since ceasing exposure to asbestos and quitting smoking', *Occupational and Environmental Medicine* **63**(8), 509 –512.

Roemer, J. E. (1998), *Equality of Opportunity*, Harvard University Press, Cambridge, Mass.

Romeu Gordo, L. (2011), 'Compression of morbidity and the labour supply of older people', *Applied Economics* **43**(4), 503–513.

Rosa Dias, P. (2009), 'Inequality of opportunity in health: evidence from a UK cohort study', *Health Economics* **18**(9), 1057–1074.

Saha, A. and Hilton, L. (1997), 'Expo-power: A flexible hazard function for duration data models', *Economics Letters* **54**(3), 227–233.

Schaap, M. and Kunst, A. (2009), 'Monitoring of socio-economic inequalities in smoking: Learning from the experiences of recent scientific studies', *Public Health* **123**(2), 103–109.

Seshamani, M. and Gray, A. (2004), 'Ageing and health-care expenditure: the red herring argument revisited', *Health Economics* **13**(4), 303–314.

Shwartz, M., Iezzoni, L. I., Moskowitz, M. A., Ash, A. S. and Sawitz, E. (1996), 'The importance of comorbidities in explaining differences in patient costs', *Medical Care* **34**(8), 767–782.

Sidorchuk, A., Agardh, E. E., Aremu, O., Hallqvist, J., Allebeck, P. and Moradi, T. (2009), 'Socioeconomic differences in lung cancer incidence: a systematic review and meta-analysis', *Cancer Causes & Control* **20**(4), 459–471.

Singh, G. K., Miller, B. A., Hankey, B. F., Edwards, B. K., Singh, G. K., Miller, B. A., Hankey, B. F. and Edwards, B. K. (2003), 'Area socioeconomic variations in US cancer incidence, mortality, stage, treatment, and survival, 1975-1999'.

Stearns, S. C. and Norton, E. C. (2004), 'Time to include time to death? The future of health care expenditure predictions', *Health Economics* **13**(4), 315–327.

Thomas, B., Dorling, D. and Smith, G. D. (2010), 'Inequalities in premature mortality in Britain: observational study from 1921 to 2007', *BMJ* **341**(c3639), c3639.

University of Cambridge Clinical School (2009), 'HALS deaths and cancer working manual, june 2009'.

US Agency for Healthcare Research and Quality (2009), 'Clinical classifications software (CCS) for ICD-10', http://www.hcup-us.ahrq.gov/toolssoftware/icd_10/ccs_icd_10.jsp.

US National Cancer Institute (2006), 'Cancer trends progress report - stage at diagnosis'.

Vallejo-Torres, L. and Morris, S. (2010), 'The contribution of smoking and obesity to income-related inequalities in health in England', *Social Science & Medicine* **71**(6).

Vineis, P., Alavanja, M., Buffler, P., Fontham, E., Franceschi, S., Gao, Y. T., Gupta, P. C., Hackshaw, A., Matos, E., Samet, J., Sitas, F., Smith, J., Stayner, L., Straif, K., Thun, M. J., Wichmann, H. E., Wu, A. H., Zaridze, D., Peto, R. and Doll, R. (2004), 'Tobacco and cancer: Recent epidemiological evidence', *Journal of the National Cancer Institute* **96**(2), 99 –106.

Werblow, A., Felder, S. and Zweifel, P. (2007), 'Population ageing and health care expenditure: a school of 'red herrings'?', *Health Economics* **16**(10), 1109–1126.

Wilkinson, R. G. (1996), *Unhealthy Societies: The Afflictions of Inequality*, Routledge, London.

Wilkinson, R. and Pickett, K. (2010), *The Spirit Level: Why Equality Is Better for Everyone*, Penguin, London.

Williams, D. (2012), 'Lansley: CCG allocations should be based on age, not poverty', *Health Services Journal* .

Wilson, D. L. (1994), 'The analysis of survival (mortality) data: Fitting Gompertz, Weibull, and logistic functions', *Mechanisms of Ageing and Development* **74**(1-2), 15–33.

Zweifel, P., Felder, S. and Meiers, M. (1999), 'Ageing of population and health care expenditure: a red herring?', *Health Economics* **8**(6), 485–496.

Zweifel, P., Felder, S. and Werblow, A. (2004), 'Population ageing and health care expenditure: New evidence on the "Red Herring"', *Geneva Papers on Risk and Insurance - Issues and Practice* **29**(4), 652–666.