WP 13/14

# A discrete latent factor model for smoking, cancer and mortality.

Daniel Howdon and Andrew Jones

June 2013

# A discrete latent factor model for smoking, cancer and mortality.

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

[1]Department for Economics and Related Studies, University of York

[2]Centre for Health Economics, University of York

June 25, 2013

## Summary

This paper investigates the relationships between social circumstances, individual behaviours, and ill-health later in life, with a particular focus on the development of cancer. A discrete latent factor model incorporating individuals' smoking and health outcomes (lifespan and time-to-cancer) is jointly estimated, using the 1984/5 British Health and Lifestyle Survey (HALS) dataset and its July 2009 follow-up, allowing for unobservable factors to affect decisions regarding smoking behaviours as well as health outcomes. Results from this discrete latent factor model are found to be substantially different to those derived from single-equation modelling, suggesting the presence of unobserved heterogeneity. Contrary to previous work on the relationship between circumstances and the development of cancer, a social gradient in time-to-cancer is observed, with individuals in the lowest two social classes developing cancer significantly sooner than individuals in the highest social class. The reduction in estimated median time-to-cancer between individuals in the highest social class, and those in the lowest social class, is found to be between 4 and 4.5 years; approximately twice as many individuals in the lowest social classes as in the highest social class are predicted to develop cancer by an observed age of 75. Those in lower social classes are found to be more likely to smoke, smoke earlier in life, and smoke more cigarettes before quitting.

**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. This research was funded by the Economic and Social Research Council under the Large Grant Scheme (RES-060-25-0045).

*Corresponding author: Tel.: +44 1904 32 1411; email daniel.howdon@york.ac.uk.

# 1 Introduction

This paper develops a joint model of smoking, mortality and cancer, with a particular focus on socioeconomic inequality in 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 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 that include 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). 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, which is line with the epidemiological literature such as Doll et al. (2004).

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. As in Howdon (2012), who finds that using lifetime incidence rather than survival times may underestimate inequalities in cancer, cancer outcomes are estimated as elapsed time to the onset of cancer, rather than lifetime incidence of the disease. 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.

The existence of social gradients in cancer is more controversial than that of social gradients in lifespan and, consequently, this paper focuses on the former. The methods used here to model smoking behaviours and cancer outcomes jointly have not been used in the investigation of social inequality in the literature.

Estimates for smoking behaviours reveal that individuals in lower social classes are more likely to start smoking, start smoking earlier in life when they do smoke, and are exposed to more pack-years of smoking, than those in higher social classes. A statistically significant relationship between social class and accelerated onset of cancer is revealed, with those in the lowest social classes found to develop cancer sooner, conditional on survival to 45, than those in the highest social class. The reduction in estimated median time-to-cancer for counterfactual estimates for the highest social class, and for the lowest social class, is almost four and a half years for both men and women – greater than the reduction when counterfactual estimates for non-smoking and for 20 pack-years of exposure are compared. These estimates further imply that approximately twice as many individuals in the lowest social classes, compared to the highest social class, are predicted to develop cancer by the age of 75. 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.

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

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). These gradients are potentially a matter of concern, depending on the conception of equity used (see, for instance, Rosa Dias and Jones (2007)), and the existence of, and direction of, a causal link between socioeconomic inequality (primarily, in income) and inequality in health (see, for instance, Deaton (2002)). 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 survival time-to-cancer (for instance, Singh et al. (2003); Banks et al. (2006); Dalstra et al. (2005); Howdon (2012)). Establishing links, and the strength of those links, between circumstances, individual behaviours and health outcomes is also complicated by a number of factors. Individuals who smoke may disproportionately come from more at-risk groups, either due to inherent (perhaps, genetic) characteristics that leave them

[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.

more prone to early death, due to their social status (for instance, less than 14% of individuals in the highest social class were current smokers in the dataset used here, compared to 39% in the lowest two social classes), or due to complementary lifestyle characteristics.

Link and Phelan (1995) argue that insufficient attention has been paid to social conditions as themselves being fundamental causes (with respect to which, ‘even if one effectively modifies intervening mechanisms or eradicates some diseases, an association between a fundamental cause and disease will reemerge’) of ill-health, and that individual risk-factors must be contextualised. Link and Phelan (1995) point to a treadmill of risk factors, with new risk factors being suggested to explain persistent social gradients in health as old risk factors were eradicated. Counterfactual estimates, therefore, based on individuals adopting apparently fully healthy lifestyles rest upon *ceteris paribus* assumptions that may not hold. Seeking to isolate the effect of lifestyles adopted disproportionately by one particular stratum of society from the inherent effect of being in that particular social stratum is to ignore the possibility of dynamic effects that would cause at least part of the social gradient associated with that particular lifestyle to persist were the disparity in lifestyles to be eliminated. This is potentially problematic for empirical analysis: while it may be clear what identifies a particular parameter (for instance, the parameter estimated for a variable related to pack-years of smoking is identified by the change in lifespan associated with exposure to an additional pack-year of smoking), it may not be clear how that parameter should be interpreted (see, for instance, Keane (2010)).

## 3 Data

This paper uses 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 the United Kingdom (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 complete 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 B1, Appendix B), 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.

[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 limitation of the HALS1 data is the 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 simply 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 'effort' 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.

A further problem 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. For instance, and in particular in this case, 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. 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.

# 4 Methods

## 4.1 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 the approach proposed by Heckman and Singer (1984) and Mroz (1999) for dealing with the effect of unobserved heterogeneity in systems of equations. 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.

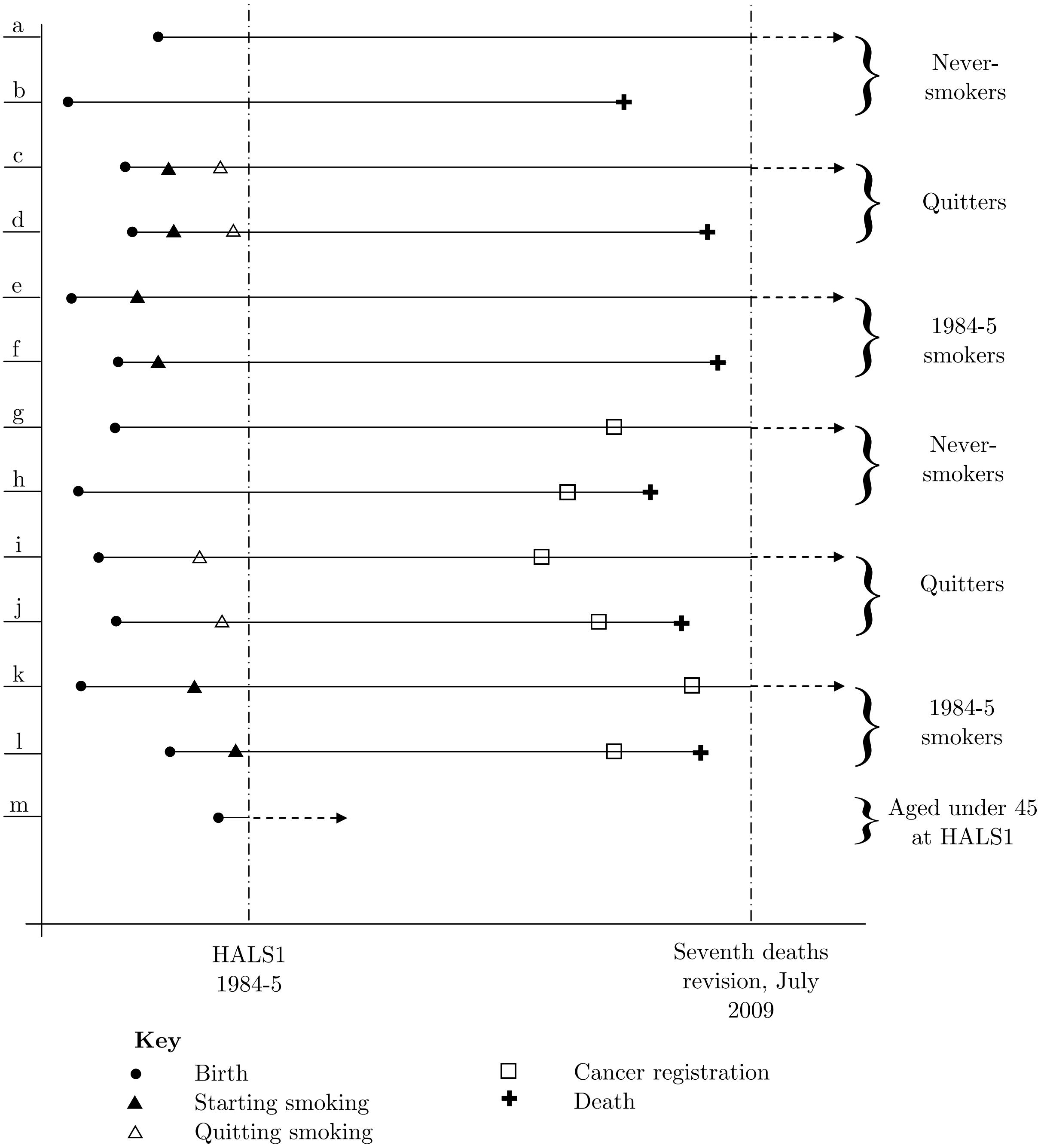


Figure 1: Types of observed outcomes

Figure 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.

### 4.1.1 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 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 $\varphi_1$ is an individual-specific intercept term, reflecting unobserved individual characteristics that influence the probability of ever smoking.

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]:

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

[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).

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

where:

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

and $\varphi_2$ 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 $\gamma_1$ is the loglogistic duration dependence parameter. Individuals who are not observed to start smoking before HALS1 ($a$, $b$, $g$ and $h$ in Figure 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.

### 4.1.2 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 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 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 to quit before HALS1 ($c$, $d$, $i$ and $j$ in Figure 1) have a "complete spell" for this function and individuals who are observed as current smokers ($e$, $f$, $k$ and $l$ in Figure 1) at HALS1 are censored observations. The overall contribution of each individual to the loglikelihood is the logged Gompertz likelihood function,

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

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 $\gamma_2$ is the Gompertz shape parameter.

### 4.1.3 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 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:

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

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

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

and $\alpha$ is the Weibull shape parameter.

### 4.1.4 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

[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*.

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 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:

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

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 $\gamma_4$ 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

[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.

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).

### 4.2 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 (2001) 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 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 (2001) 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

[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).

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 $\tau$ and $\rho$ are the factor loadings.

Mixing probabilities, $\pi_k$, 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 $\theta_1$ and $\theta_2$ 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, $\zeta_p$. The structure of the latent factor model is summarised in Table 1.

| Mass point, $k$ | $u$ | $v$ | $\varphi_j$ |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | $\tau_j$ |
| 3 | 0 | 1 | $\rho_j$ |
| 4 | 1 | 1 | $\nu_j$ |

Table 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. Balia and Jones (2011) fix mass points at 0 and 1 (i.e. where $u = v = 1$ and $\tau + \rho = 1$), 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 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 lognormal 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.

[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.

[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.

## 4.3 Key covariates and interpretation of parameters

Summary statistics for the variables used in the analysis are presented in Table 2:

| 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 |
| alcboth | Both parents heavy drinkers | 0.00 | 0.05 | 0.00 | 1.00 |
| alcmoonly | Only mother heavy drinker | 0.00 | 0.06 | 0.00 | 1.00 |
| alcfaonly | Only father heavy drinker | 0.06 | 0.24 | 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 |
| 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 |

Table 2: Summary statistics (all 3784 observations)

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 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

[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 missing for this variable at HALS2.

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].

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. Individuals who exercised for more than 5 hours in

[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`

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

the previous two weeks are classed as having exercised for the recommended period of time in this model[16]. 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.

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[17].

## 5 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 and 4. Table 3 shows the coefficients associated with the covariates and Table 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 5.

[16] 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`

[17] 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.

| Variable | smoker | starting | pack-years | lifespan | cancer |
|---|---|---|---|---|---|
| mothm | 0.562*** | -0.008 | | | |
| mothf | 0.557*** | -0.047 | | | |
| fathm | 0.472*** | -0.041* | | | |
| fathf | 0.280*** | -0.057** | | | |
| bothm | 0.523*** | -0.048* | | | |
| bothf | 0.682*** | -0.105*** | | | |
| sc23 | 0.305** | -0.087*** | -0.263 | 0.128 | -0.030 |
| sc45 | 0.538*** | -0.126*** | -0.413** | 0.394** | -0.046* |
| lhqdeg | -0.438*** | 0.067** | 0.139 | -0.530** | 0.054* |
| lhqoth | -0.244 | 0.059 | -0.290 | 0.006 | 0.061 |
| lhqA | 0.100 | 0.045* | 0.177 | -0.342 | 0.026 |
| lhqO | -0.169** | 0.046** | -0.019 | 0.011 | -0.000 |
| lhqhnd | -0.264 | 0.117*** | 0.017 | -0.269 | 0.010 |
| male | 0.649*** | -0.201*** | -0.079 | 0.442*** | -0.026*** |
| bc20 | 0.302*** | -0.050*** | -0.019 | -0.008 | -0.023* |
| bc30 | -0.001 | -0.085*** | 0.175 | -0.046 | -0.045*** |
| bc40 | -0.070 | -0.265*** | 0.052 | 0.267 | -0.016 |
| strtpostdoll | | 0.347*** | | | |
| strtpostpubhealth | | 0.921*** | | | |
| starting | | | 0.049*** | | |
| othersmok | | | -0.752*** | 0.109 | -0.015 |
| ltunemp | | | -0.479** | 0.548** | -0.061** |
| sick | | | -0.364** | 0.785*** | -0.025 |
| retd | | | 0.113 | -0.136 | 0.023* |
| single | | | -0.187 | 0.257** | 0.015 |
| sepdiv | | | -0.729*** | -0.027 | -0.005 |
| widowed | | | -0.394*** | 0.078 | 0.017 |
| rural | | | 0.256*** | -0.091 | -0.004 |
| suburb | | | 0.130* | -0.041 | -0.003 |
| pack_years_quit | | | | 0.014*** | -0.001** |
| pack_years_quit2 | | | | -0.557** | 0.027 |
| pack_years_start | | | | 0.037*** | -0.003*** |
| pack_years_start2 | | | | -3.032*** | 0.243*** |
| NPAD | | | | 0.184* | -0.022 |
| recex | | | | -0.402*** | |
| redmeat3 | | | | -0.123** | 0.011 |
| constant | -0.706*** | 3.126*** | -4.786*** | -56.533*** | 4.718*** |
| $\gamma$ | | 0.141*** | 0.008*** | | 0.065*** |
| $\alpha$ | | | | 12.327*** | |
| N. of cases | | | 3784 | | |

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

Table 3: DLFM results – main coefficients

|  | Latent class, k |  |  |  |
|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 |
| $\varphi_1$ | 0 | 0.287** | -0.144 | 0.143 |
| $\varphi_2$ | 0 | -0.075*** | 0.010 | -0.065*** |
| $\varphi_3$ | 0 | -0.463*** | 0.577** | -0.114 |
| $\varphi_4$ | 0 | 2.356*** | 1.341*** | 3.697*** |
| $\varphi_5$ | 0 | -0.276*** | -0.211*** | -0.487*** |
| $\pi_k$ | 0.353*** | 0.443*** | 0.090*** | 0.113*** |

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

Table 4: DLFM results (2) – latent factor coefficients and class membership probabilities

| Variable | Coefficient |
|---|---|
| pack_years_quit | -0.002** |
| pack_years_quit2 | 0.025 |
| pack_years_start | -0.005*** |
| pack_years_start2 | 0.320** |
| othersmok | -0.013 |
| NPAD | -0.028 |
| redmeat3 | 0.015 |
| lhqdeg | 0.073 |
| lhqoth | 0.050 |
| lhqA | 0.050 |
| lhqO | 0.004 |
| lhqhnd | 0.016 |
| ltunemp | -0.096** |
| sick | -0.025 |
| retd | 0.023 |
| male | -0.036** |
| sc23 | -0.054 |
| sc45 | -0.073 |
| single | -0.003 |
| sepdiv | 0.004 |
| widowed | 0.027 |
| rural | -0.019 |
| suburb | -0.001 |
| bc20 | 0.031 |
| bc30 | 0.043 |
| bc40 | 0.187 |
| Constant | 4.559*** |
| $\gamma$ | 0.164*** |
| N. of cases | 3784 |

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

Table 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). However, there appears to be no shift in the intercept caused by smoking (either for ex-smokers or current smokers at HALS1). 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[18].

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. Conversely to the results for lifespan, the shifts caused by dummy variables for starting and quitting smoking appear to explain part of the elevation in risk caused by smoking.

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[19]. 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.

### 5.1 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 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 cancer, despite this class being made up of individuals with approximately similar smoking characteristics and

[18] $12\beta_{pack\_years\_start} - 12^2 \left(\beta_{pack\_years\_start2}/10000\right) \approx \beta_{sc45}$.

[19] 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 6: Descriptive statistics, by most probable latent class based on posterior probabilities.

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 – possibly genetic – 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.

# 6 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 or social class to the same value, 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:

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

where $X_{cf}$ refers to the counterfactual values for variables. These probabilities are multiplied by the associated prior probability of class membership. These product 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, with men and women considered separately, are presented in Table 7, with estimated median survival curves presented in Figures 2 to 5.

| | **Male** | | **Female** | |
|---|---|---|---|---|
| | Estimated survival time | *Difference from full sample* | Estimated survival time | *Difference from full sample* |
| Full sample | 85.0 | – | 88.7 | – |
| *Counterfactuals* | | | | |
| Social class 1 | 87.8 | *+2.8* | 91.9 | *+3.2* |
| Social class 2/3 | 85.2 | *+0.20* | 89.2 | *+0.5* |
| Social class 4/5 | 83.9 | *-1.1* | 87.8 | *-0.9* |
| 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* |
| Social class 1, non-smoker | 90.6 | *+5.6* | 93.7 | *+5.0* |
| Social class 4/5, 30 pack-years | 81.0 | *-4.0* | 83.7 | *5.0* |

Table 7: Counterfactual estimates – median survival time to onset of cancer (years)

The reduction in estimated median survival time between counterfactual estimates for the highest social class, and for the lowest social class, is approximately four and a half years for both men and women – greater than the reduction when individuals who do not smoke, and individuals with 20 pack-years of exposure, are compared.

While the $\varphi_k$ 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 $\varphi_k$ 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. 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 4, 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 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, as shown in Figure 3, 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.

While, at an age of 75, over 78% of males in social class 1 (Figure 2) are predicted to be cancer free, this probability falls to 75% in social classes 2 or 3 and, in social class 4 or 5, to just over 72%. While this variation is not as immediately dramatic as the differences between the unobserved factors generating latent classes, it does mean that over 25% more men in social

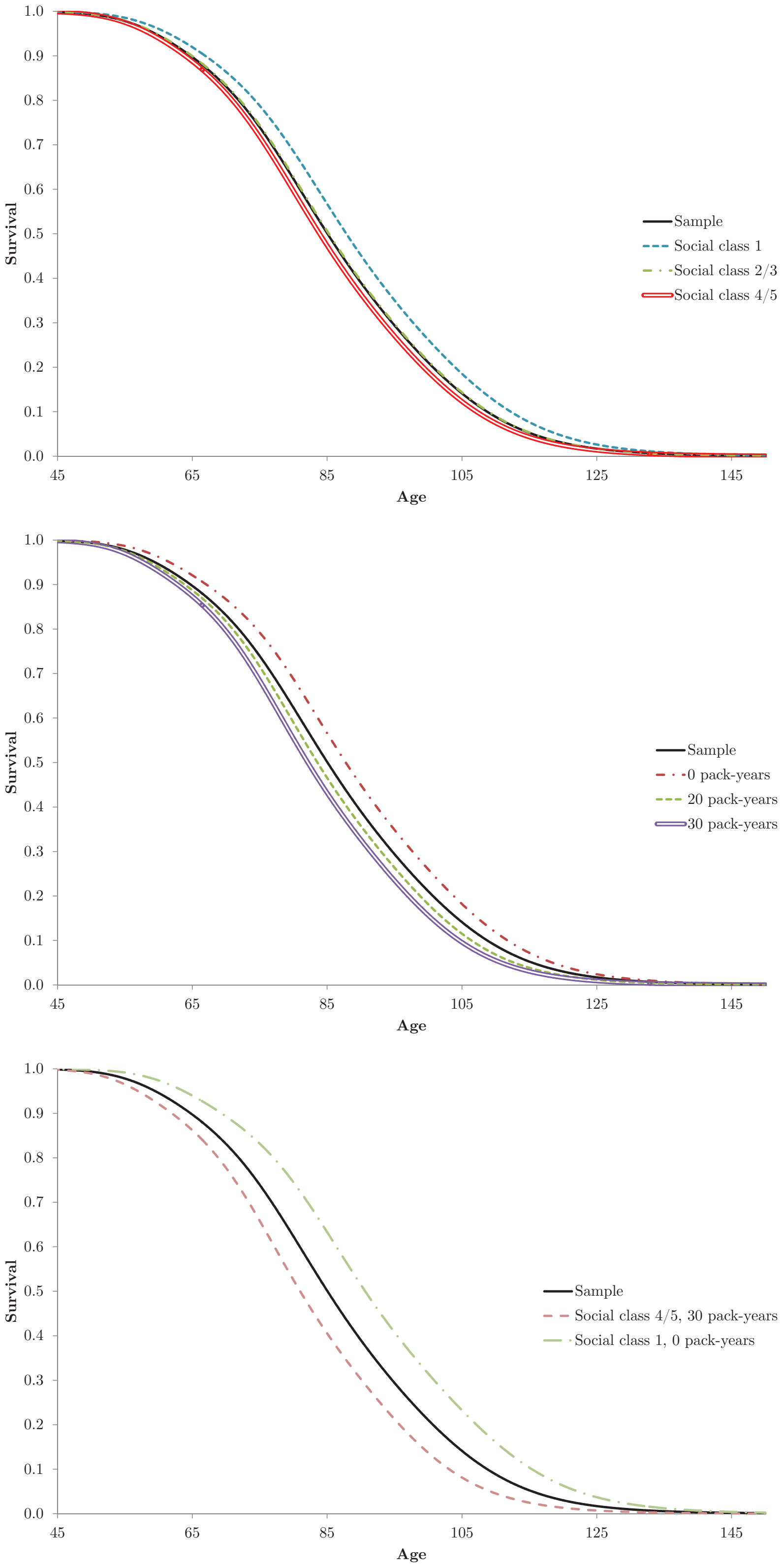


Figure 2: Estimated survival curves (males)

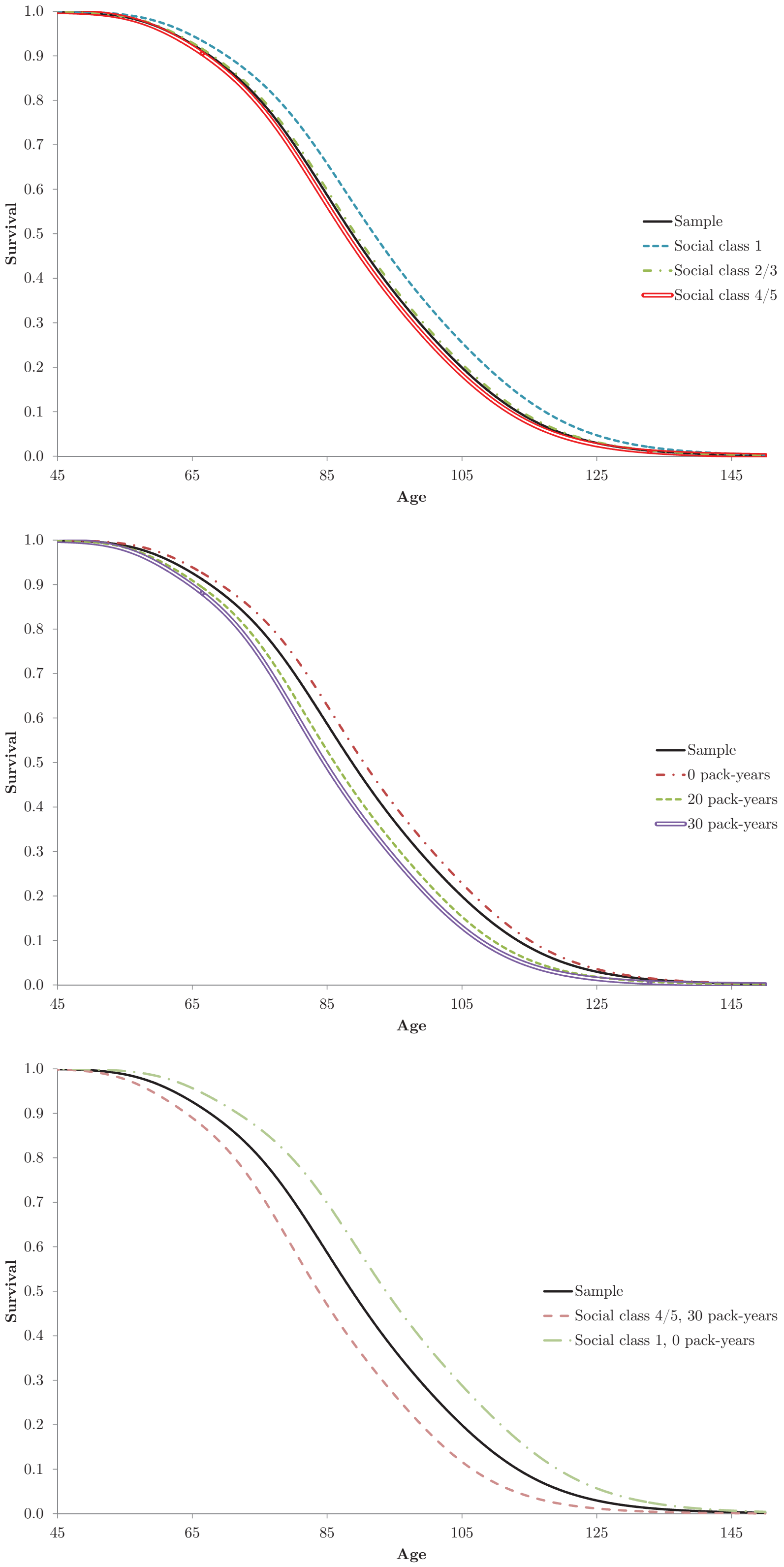


Figure 3: Estimated survival curves (females)

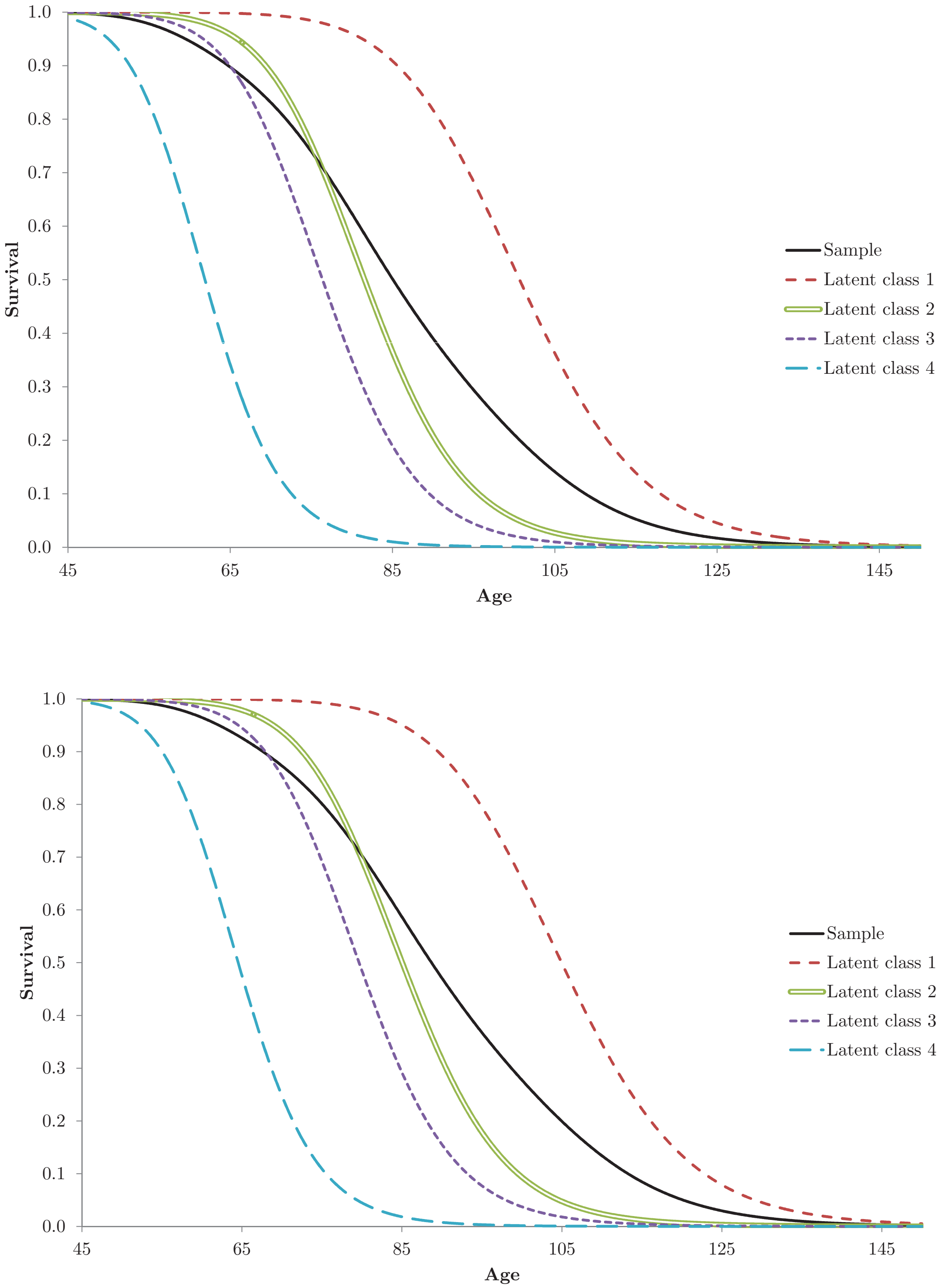
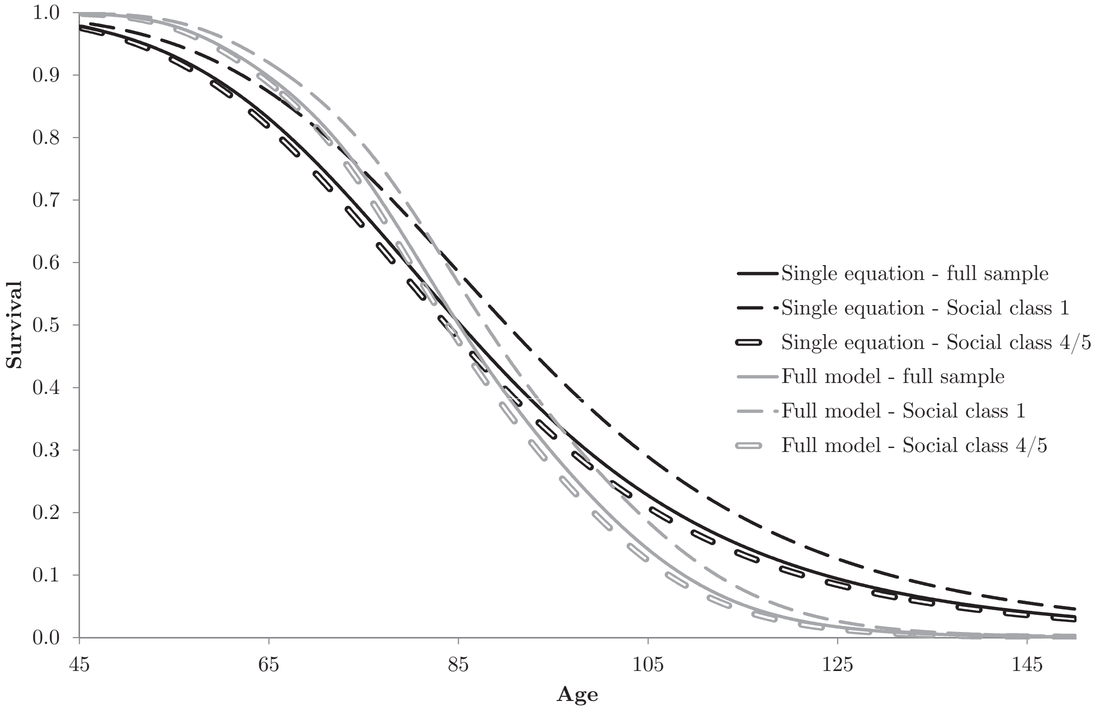


Figure 4: Estimated survival curves by latent class (top – male, bottom – female)



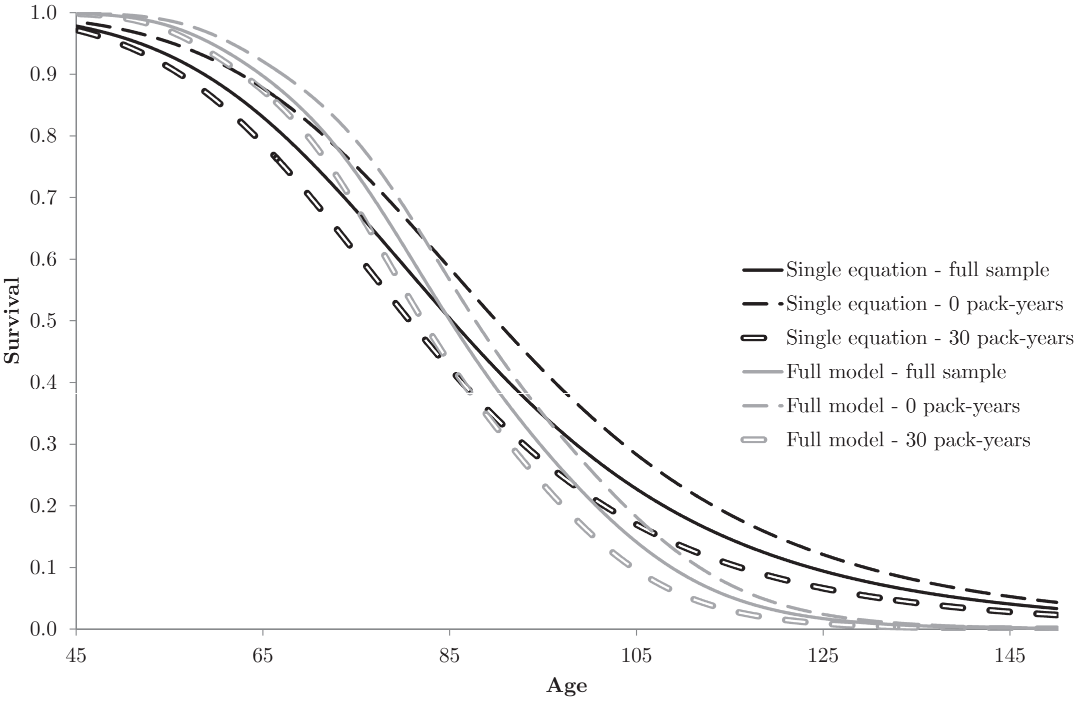


Figure 5: Estimated survival curves (males) – comparison of single-equation and full model results

class 4 or 5 are predicted to develop cancer by age 75 as those in the highest social class. For women (Figure 3), these respective probabilities are 84%, 81% and 79%. At the age of 95, these probabilities are 35% for men (43% for women) in social class 1, 30% for men (38% for women) in social class 2 or 3, and 27% for men (35% for women) in social class 4 or 5. This social inequality in cancer outcomes, using assumptions that are apt to underestimate any such inequality, is clear and striking.

The difference between results obtained using single equation estimates and those from the full DLFM (Figure 5) for men is also notable. The different duration dependence ($\gamma$) 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 both being in different social classes or 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 counterfactual predictions for the highest and lowest social class is 6.4 years in the single equation model, and 3.9 in the joint model. Similarly, 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. These results suggest a role for unobserved heterogeneity in explaining differences in survival times. Failure to account for this unobserved heterogeneity would cause differences in survival times between both individuals in different social classes, and individuals with different smoking exposures, to be overestimated.

## 7 Discussion

Existing literature on the relationship between social class, smoking behaviours and cancer is very limited: we are aware of no existing research employing duration techniques to examine such relationships. Research using data from the British Doctors Study, while employing a large sample over a long time period, looks at only a small stratum of society – male doctors in the UK – and smoking data in this dataset is much less rich than that contained in HALS. Notably, contrary to the claims of Deaton (2002), a social gradient in cancer outcomes is observed even after controls are made for smoking behaviours.

The findings for smoking and lifespan equations are broadly in line with those in Balia and Jones (2011). In addition to introducing cancer outcomes, we, here, build on their 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 is more than twice as great as that in the joint model, leading to a much flatter estimated survival function, and more early and late failures. Further, the differences in estimated survival times associated with social class and 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 then-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 associated reduction in healthy time without cancer is estimated to be 4 years for men

(4.2 years for women) for those in the lower social classes compared to those in the highest. Around twice as many men (and more than twice as many women in corresponding groups) in the lowest social classes as in the highest social class are predicted to develop cancer by age 75. Despite this social gradient, unobservable factors seem to swamp this. 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.

# A AIC and BIC scores

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 A1: Comparison of baseline hazards

# B Cases where cancer present on death certificate, without cancer registry record.

| 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 B1: 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. (2001), Smoking and endogenous mortality: Does heterogeneity in life expectancy explain differences in smoking behavior?, Economics Series Working Paper 77, University of Oxford, Department of Economics.

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. volume IV. Descriptive epidemiology.’, *IARC Scientific Publications* **128**.

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), 2335–2350.

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.

Howdon, D. (2012), Time and chance happen to them all? Duration modelling versus lifetime incidence of cancer, Health, Econometrics and Data Group (HEDG) Working Paper 12/06, HEDG, Department of Economics, University of York.

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.

Keane, M. P. (2010), 'Structural vs. atheoretic approaches to econometrics', *Journal of Econometrics* **156**(1), 3–20.

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.

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

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), 178–182.

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', http://www.ctsu.ox.ac.uk/ tobacco/contents.htm.

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.

Rosa Dias, P. and Jones, A. M. (2007), 'Giving equality of opportunity a fair innings', *Health Economics* **16**(2), 109–112.

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', http://seer.cancer.gov/publications/ses/.

Thomas, B., Dorling, D. and Smith, G. D. (2010), 'Inequalities in premature mortality in Britain: observational study from 1921 to 2007', *BMJ* **341**(c3639), 291.

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', http://progressreport.cancer.gov/doc_detail.asp?pid=1&did=2007&chid=73&coid=721&mid.

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), 1189–1198.

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.