HEDG Health, Econometrics and Data Group
THE UNIVERSITY of York

HEDG Working Paper 09/08

Catching the habit: a study of inequality of opportunity in smoking-related mortality

Silvia Balia

Andrew M Jones

May 2009

ISSN 1751-1976

york.ac.uk/res/herc/hedgwp

Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2011)
174, *Part* 1, *pp.* 175–194

Catching the habit: a study of inequality of opportunity in smoking-related mortality

Silvia Balia

Università di Cagliari, Italy

and Andrew M. Jones

University of York, UK

[Received June 2009. Final revision May 2010]

Summary. The paper presents a latent factor model for initiation of smoking, cessation and mortality using the British Health and Lifestyle Survey. This allows us to investigate inequality of opportunity in risk of mortality, focusing on the intergenerational transmission of smoking. We find that the hazard of mortality is higher for current and former smokers relative to never smokers. Furthermore we find that parental smoking plays an important role in the dynamics of smoking and indirectly affects mortality. Predictions derived from the model show that inequality in mortality decreases if individuals adopt the best level of effort (not smoking) or if circumstances are favourable (if parents are non-smokers).

Keywords: Duration analysis; Inequality of opportunity; Latent factors; Mortality; Smoking

1. Introduction

Tobacco smoking causes about 30% of all cancer deaths in developed countries as well as causing deaths from vascular, respiratory and other tobacco-related diseases (Vineis *et al.*, 2004). In European countries all-cause mortality rates have been attributed largely to smoking (Peto *et al.*, 2006). The medical and epidemiological literature shows that smoking offsets recent medical developments by reducing life expectancy for smokers by about 10 years relative to non-smokers (Doll *et al.*, 2004).

Smoking trends are often associated with socio-economic inequalities. In the UK, for example, the highest prevalence of smoking and the highest risk of dying from smoking-related diseases are among the most disadvantaged socio-economic groups (Kunst *et al.*, 2004). The socio-economic gradient in smoking status has been widely investigated (Shaap and Kunst, 2009) and this has implications for inequality of opportunity in health. Unequal opportunities are widely regarded as unfair and shifting the object of interest from socio-economic inequality *per se* to inequality of opportunity helps to broaden the consensus about the need to reduce those inequalities. This entails understanding more about the influence of 'circumstances' on smoking and the contribution that this makes to the effect of smoking on health.

Recent medical evidence shows that parental smoking increases the risk for tobacco use in adolescents (see, for example, Keyes *et al.* (2008)). It could influence smoking behaviour of current generations through the (intergenerational) transmission of preferences for health and

Address for correspondence: Andrew M. Jones, Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD, UK.
E-mail: amj1@york.ac.uk

 0964–1998/11/174175

smoking. From this perspective, current generations are not fully responsible for the subsequent effects of smoking on their own risk of mortality because their family background, which might encourage smoking, is beyond their direct control. Parental smoking can then be considered as one specific *circumstance* factor influencing health outcomes in the sense of Roemer (1998, 2002). According to Roemer's framework, circumstances, such as family background, represent a source of illegitimate variation in individual outcomes. But they are not the sole determinants of the variation in outcomes. Although circumstances may shape behaviour, to some extent, individuals are held responsible for their actions. Roemer called this individual responsibility *effort* and variation in outcomes due to differences in effort, among individuals sharing the same circumstances, is regarded as acceptable. Of course, the level of effort that someone exerts to protect their health may itself be influenced by their personal circumstances. The challenge is to disentangle the separate effects of circumstances and effort on variations in health outcomes.

In the health economics literature, effort is often identified with variation in health-related behaviours, such as smoking, for those who share the same circumstances. Individuals who share the same circumstances, such as parental smoking, may still differ when deciding whether to start smoking and whether to quit. The opportunity approach to inequality has recently become influential in the health economics literature, although it was already implicit in much of the existing works on health equity (see Rosa Dias and Jones (2007)). Empirical applications can be found in Trannoy *et al.* (2010) and Rosa Dias (2009), where evidence of substantial inequality of opportunity in health is presented. Rosa Dias (2010) integrated Roemer's (1998, 2002) conceptual framework for inequality of opportunity with Grossman's (1972) model of health capital and the demand for health. He presented a behavioural model of health taking into account the presence of unobserved heterogeneity.

The main purpose of this paper is to investigate variation in individual mortality and to measure inequality in smoking-related risk of mortality, focusing on intergenerational transmission of smoking habits. We estimate a model for smoking and mortality by using the British Health and Lifestyle Survey (HALS). Our model investigates the dynamics of smoking behaviours in terms of initiation of smoking and cessation and the effect of smoking on the risk of mortality. This is possible thanks to the nature of the survey data that we use, which gives us the scope to specify a joint model for smoking (initiation and cessation) and mortality within a time-to-event survival framework.

The potential endogeneity of indicators of smoking in models for mortality needs to be addressed. The possible biases that are caused by omitted variables and self-selection into smoking are a threat to the validity of inferences about the causal effects of smoking on health outcomes (Auld, 2006). This motivates our use of a latent factor model that accounts for individual-specific unobservable heterogeneity and relies on parental smoking variables as instruments to identify initiation of smoking and cessation. The empirical analysis is extended to analyse inequality of opportunity: this is based on post-estimation predictions from the model that are intended to isolate the partial effects of changes in parental smoking. The Gini coefficient for inequality in overall mortality risk, Sen's welfare index and generalized Lorenz curves are computed for the baseline model and for hypothetical scenarios which compare individual types that differ in terms of this particular set of circumstances and in levels of effort.

We find that the hazard rate for mortality is higher for current and former smokers relative to never smokers, and lower for former relative to current smokers. The duration of smoking is likely to be shorter if smokers take up the habit later in life. We also show that parental smoking behaviour plays an important role in smoking dynamics and indirectly affects mortality. Inequality in mortality is found to decrease if individuals adopt the best effort (not smoking) or, alternatively, if circumstances are favourable (if parents are non-smokers).

The programs that were used to analyse the data can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Survival data in the Health and Lifestyle Survey

The HALS data contain retrospective information about time to death, time to starting consumption of tobacco and time to quitting, for a representative sample of adults in Great Britain (see Cox *et al.* (1987, 1993)). This information is exploited to construct duration variables that indicate the time elapsed before each event occurs.

The initial wave of the HALS was carried out between autumn 1984 and summer 1985 in two home visits (the second by a nurse). The population surveyed comprises individuals aged 18 years and over living in private households: 9003 interviews were completed, giving a response rate of 73.5%. HALS respondents have been tracked on the UK National Health Service registers regularly to provide reliable information about individual mortality. The deaths data were released in June 2005. This allows us to investigate survival up to April 2005: up to this date 97.8% of the original sample has been flagged and 27% of the respondents had died. The questionnaire was designed to provide comprehensive information about risky behaviours. In particular, the survey data contain retrospective information on smoking. Self-reported variables describing age at starting smoking, current smoking status and how long ago the respondent stopped smoking are used to derive two duration variables which can be used to study the hazard of starting smoking and the hazard of quitting smoking.

Fig. 1 illustrates the study time and the survival times for the respondents of the HALS. The vertical lines indicate the time when the survey questionnaire was administered (1984–1985) and the time of the follow-up (April 2005), which provides information about deaths. Individual characteristics, including smoking behaviours, are observed only at the survey time in 1984 (we do not make use of the smaller second wave of the HALS nor of the earlier releases of the deaths data). The study time is the calendar time period that each respondent spent in the study since the time of entry, at birth, indicated by a full circle. Dots that are closer to the y-axis represent older individuals; dots that are closer to the vertical line indicating the time of the interview represent younger individuals. For respondents who died by April 2005 the survival time is the period of time elapsed from the survey time to death (D). Respondents who are still alive at the follow-up (A) have a censored survival time. Respondent types i and j never smoked, so they do not have survival times for smoking. Respondent types k–n started smoking (S). In particular, k and l are current smokers at the time of the interview (they might have stopped smoking at some point in the future), so they have a complete survival time for starting to use tobacco but a censored survival time for quitting smoking (C), whereas m and n are ex-smokers (Q) and have a complete survival time for both starting and quitting smoking.

To analyse initiation of smoking we define a time variable *starting* which represents the number of years elapsed before someone starts to smoke. For consistency, individuals who claimed to be current smokers but whose age at starting was 0 have been eliminated. The variable starting is equal to the age at starting for those who started smoking; it is right censored at the age at the survey time for those who had not started by then. This means that each individual is assumed to become at risk of starting at birth. A binary indicator, *start*, indicates whether or not an individual started to smoke.

Cessation of smoking is described by the variable *sm_years* which indicates the number of years that a person smoked. For current smokers sm_years is right censored at the time of the interview. A binary indicator, *quit*, indicates whether or not a smoker had stopped smoking completely.

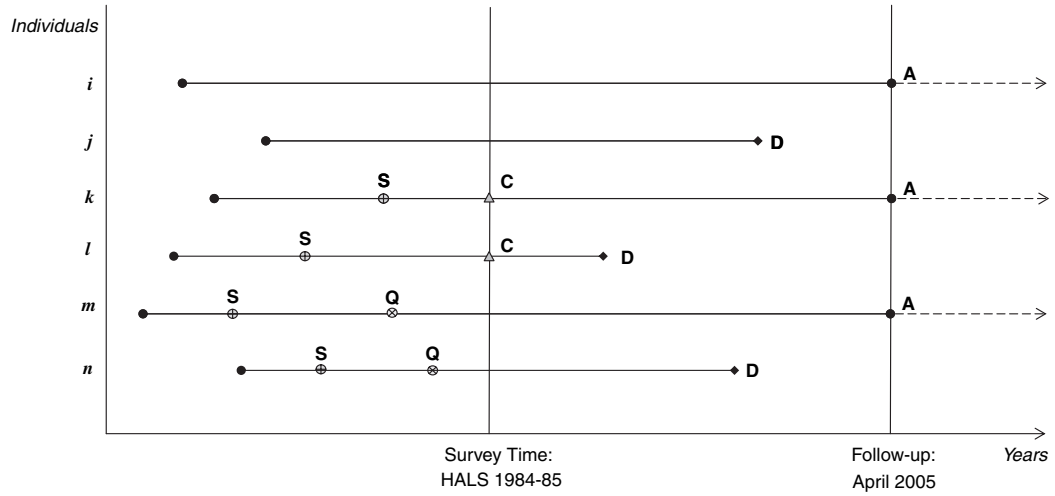

Fig. 1. Study time and survival time in the HALS

As shown by Fig. 1, the HALS data give us the scope to investigate the hazard of death. A complete duration is observed for those who died before the follow-up period, whereas an incomplete duration is associated with individuals who were still alive in April 2005. We construct a time variable *lifespan* which is equal to the age at death for those who died by April 2005, and to the age at the time of the follow-up, for those who were still alive. This allows us to measure length of survival from birth, conditional on survival up to the time of the survey, in which case the distribution of the survival time is said to be left truncated.

We restrict the analysis to individuals aged 40 years old or older at the time of the initial survey: only 1% of the sample died before age 40 years. This allows us to avoid confounding mortality with accidents, injuries or a genetic predisposition towards early death that is not related to smoking. Furthermore, new starts are usually rare for individuals in their 40s or over. In fact, the proportion of HALS respondents who reported an age of starting above 40 years is very low (1.2%).

For the purpose of our analysis, this subsample of individuals from the HALS has been reduced by listwise deletion of cases due to item non-response in the variables of interest. Differences between the means of relevant variables in the original sample of individuals aged 40 years old or older and the sample that was used for analysis are not statistically significant. Our sample slightly over-represents males, married individuals and full-time workers (about 2% more individuals in these categories). Widows are slightly under-represented (1.8% fewer widows), as well as retired (2%) and deaths (2%). The mean age is slightly lower (0.8 years) as well as age at starting (0.1 years). The sample that was used for analysis consists of 4572 individuals, of whom about 45% are males, and the mean age is 58 years. Those who started smoking at some point in their life account for the 62% of the sample of whom 49% had stopped smoking at the time of the HALS interview (1984–1985). The hazards of initiation of smoking and cessation are analysed by focusing on the subsample of smokers, which includes 2828 individuals. On average, current and ex-smokers in the sample have smoked for around 32 years. About 47% of smokers had died by April 2005, which was more than in the full sample (43%), and the mean lifespan is about 73 years, which is shorter than in the full sample (74 years). Complete summary statistics are reported in Table 1.

Table 1. Definitions of variables and summary statistics†

Name of variable	*Definition of variable*	*Full sample (N = 4572), mean*	*Only smokers (N = 2828), mean*
Survival time variables and censoring indicators			
starting	Number of years elapsed before starting	33.40 (21.63)	18.01 (6.20)
sm_years	Number of years a person smoked	—	32.12 (14.07)
lifespan	Number of years lived by April 2005	74.02 (9.64)	73.28 (9.27)
start	1 if started smoking before the HALS; 0 otherwise	0.62	—
quit	1 if quit smoking before the HALS; 0 otherwise	—	0.49
death	1 if has died by April 2005; 0 alive	0.43	0.47
Observed characteristics			
sc1	1 if professional or managerial; 0 otherwise	0.30	0.26
sc2	1 if skilled (manual or not) or armed service; 0 otherwise	0.47	0.49
sc3	1 if partly skilled, unskilled, unclassified or never occupied; 0 otherwise	0.23	0.25
degree	1 if university degree; 0 otherwise	0.12	0.10
other edu	1 if other vocational or professional qualifications; 0 otherwise	0.05	0.06
hvqA	1 if higher vocational qualifications or A level or equivalent; 0 otherwise	0.12	0.11
O-cse	1 if O level or Certificate of Secondary Education; 0 otherwise	0.09	0.09
no edu	1 if no qualification; 0 otherwise	0.63	0.65
married	1 if married; 0 otherwise	0.75	0.77
widow	1 if widow; 0 otherwise	0.13	0.11
sepdiv	1 if separated or divorced; 0 otherwise	0.06	0.06
single	1 if never married; 0 otherwise	0.06	0.06
full time	1 if full-time worker; 0 otherwise	0.35	0.37
part time	1 if part-time worker; 0 otherwise	0.13	0.11
unemployed	1 if the individual is unemployed; 0 otherwise	0.03	0.04
sick	1 if absent from work due to sickness; 0 otherwise	0.03	0.04
retired	1 if retired; 0 otherwise	0.36	0.35
housekeeper	1 if housekeeper; 0 otherwise	0.10	0.08
rural	1 if lives in the countryside; 0 otherwise	0.21	0.20
suburb	1 if lives in the suburbs of the city; 0 otherwise	0.46	0.46
urban	1 if lives in a urban area; 0 otherwise	0.33	0.34
household size	Number of other people in the household	1.60 (1.25)	1.61 (1.24)
mother smoked	1 if only mother smoked; 0 otherwise	0.05	0.05
father smoked	1 if only father smoked; 0 otherwise	0.58	0.59
both smoked	1 if both parents smoked; 0 otherwise	0.23	0.25
others smoked	1 if anyone else in house smoked regularly; 0 otherwise	0.35	0.39
male	1 if male; 0 otherwise	0.45	0.55
age	Age in years	58.08 (11.78)	57.90 (11.32)
bc10	1 if born before 1920; 0 otherwise	0.28	0.27
bc20	1 if born between 1920 and 1929; 0 otherwise	0.27	0.30
bc30	1 if born between 1930 and 1939; 0 otherwise	0.28	0.26
bc40	1 if born between 1940 and 1949; 0 otherwise	0.17	0.17
sy23	1 if year of starting is 1923 or earlier; 0 otherwise	—	0.05
sy2437	1 if year of starting is between 1924 and 1937; 0 otherwise	—	0.22
sy3844	1 if year of starting is between 1938 and 1944; 0 otherwise	—	0.23
sy4554	1 if year of starting is between 1945 and 1954; 0 otherwise	—	0.25
sy55	1 if year of starting is 1955 or later; 0 otherwise	—	0.25

†Standard deviations are given in parentheses.

3. Model for smoking and mortality

This section presents a simultaneous hazards model for mortality and smoking, which takes into account individual unobservable heterogeneity and controls for potential sample selection bias.

3.1. Hazard models

For initiation of smoking we use a simplified version of the split-population model, which was first used to model smoking by Douglas and Hariharan (1994), which can be viewed as a two-part specification of the duration model. The first part of the model is a probit for the probability that an individual will eventually start smoking:

$$\Pr(s=1|\mathbf{x}_1)=\Phi(\beta_1'\mathbf{x}_1) \tag{1}$$

where s is a binary indicator for starting and $\mathbf{x}$ is a set of observed exogenous variables. The second part is a duration model, which is applied only to the starters in the sample, which follows a log-logistic distribution with density

$$f(t_1|s=1,\mathbf{x}_2)=\frac{\phi^{1/\gamma_1}t_1^{1/\gamma_1}-1}{\gamma_1\{1+(\phi t_1)^{1/\gamma_1}\}^2} \tag{2}$$

where t_1 is the duration variable starting for initiation of smoking, $\phi=\exp(-\beta_2'\mathbf{x}_2)$ and γ_1 are the parametric component and the duration dependence parameters of the log-logistic model. The two-part duration model has already been shown to provide a good fit of the HALS data on initation of smoking in Forster and Jones (2001): this assumes conditional independence between the components of the model (the probability of starting and the age of starting) given x_2. Estimation of the structural model for initiation of smoking relies on the assumption of independence conditional on the latent factors.

For the cessation of smoking and mortality duration models we have chosen the most appropriate parameterizations on the basis of information criteria (Akaike's information criterion AIC and the Bayesian information criterion BIC), Cox–Snell residuals and comparison with a piecewise constant exponential model. The smoking cessation hazard model follows the Weibull distribution and can be estimated in the subsample of those who have eventually started smoking. Their contribution to the sample likelihood is

$$h(t_2|s=1,t_1,\mathbf{x}_3)^q\,S(t_2|s=1,t_1,\mathbf{x}_3)=(\psi\gamma_2 t_2^{\gamma_2-1})^q\exp(-\psi t_2^{\gamma_2}) \tag{3}$$

where t_2 is the duration variable sm_years, q is the censoring indicator for quitting, $\psi=\exp(-\delta_1 t_1-\beta_3'\mathbf{x}_3)$ and γ_2 are the parametric component and the duration dependence parameter. Former smokers ($q=1$), for whom a complete spell of smoking is observed, contribute with both the hazard and the survival functions, $h(t_2)\,S(t_2)$, whereas current smokers ($q=0$), who have a censored spell, contribute only with the survival function $S(t_2)$. The coefficient δ_1 measures the causal relationship between age at starting and duration of smoking. ϕ and ψ are negative functions of covariates because the smoking duration models are parameterized in the accelerated failure time (AFT) metric: coefficient estimates should be interpreted in terms of acceleration (or deceleration) of time to failure.

For the mortality hazard model we use the Gompertz distribution and account for left truncation of the duration variable in the individual contribution to the sample likelihood:

$$h(t_3|s,q,\mathbf{x}_4)^d\,\frac{S(t_3|s,q,\mathbf{x}_4)}{S(\tau|s,q,\mathbf{x}_4)}=\{\varphi\exp(\gamma_3 t_3)\}^d\,\frac{\exp[-(\varphi/\gamma_3)\{\exp(\gamma_3 t_3)-1\}]}{\exp[-(\varphi/\gamma_3)\{\exp(\gamma_3\tau)-1\}]} \tag{4}$$

where t_3 is lifespan for length of life, d is the censoring indicator for dying, $\varphi_i = \exp(\delta_2 s + \delta_3 q + \beta_4' \mathbf{x}_4)$ and γ_3 is the dependence duration parameter. The hazard of dying is observed for everyone in the sample: those who have a complete spell ($d = 1$) are represented by $h(t_3)S(t_3)/S(\tau)$, whereas those who are still alive at the time of the follow-up are represented by the left-truncated survival function $S(t_3)/S(\tau)$ where τ is the truncation variable, age at the time of the first interview. The specification of the model allows us to distinguish between never, current and former smokers: the coefficients δ_2 and $\delta_2 + \delta_3$ measure respectively the causal effect of being a current smoker or a former smoker on the mortality hazard. Never smokers are the reference category and are identified by setting $s = 0$ and $q = 0$. The Gompertz model can only be estimated by using a proportional hazard (PH) specification.

3.2. Endogeneity issues

Our model is a multiple-hazard regression model for equations (1)–(4) and has triangular form. This reflects the chronology of events, as starting smoking must precede quitting and quitting smoking can precede death but not vice versa. Smoking variables are potential endogenous regressors in the cessation of smoking and mortality equations in the presence of selection bias due to unobservables.

Unobservable heterogeneity is an issue in the estimation of the causal effect of health-related behaviours on health outcomes and hazard rates. Individual-specific unobservable (or unmeasured) factors might influence the decision to smoke, the age at starting, the number of years spent smoking and the individual (human) survival. The recent literature shows mixed evidence regarding the role of unobservable heterogeneity in the relationship between smoking and mortality. Adda and Lechene (2001, 2004) found that individuals with lower life expectancy, as well as those in poor health, select into smoking. Less healthy individuals would therefore be more likely to start smoking at an early age, and also less likely to give up smoking. This might occur when individuals' beliefs about their life expectancy have the effect of lowering the opportunity cost of smoking. Balia and Jones (2008) found that frailer individuals tend to adopt healthy behaviours and select out of smoking (either they never start or, if they started, they quit). This can be explained on the basis of heterogeneity in the way that people internalize the negative effects of tobacco consumption on health, as stressed by the literature on nicotine addiction. Unobservable factors also influence initiation of smoking and cessation. Hence, age at starting is potentially endogenous in the hazard of quitting smoking. van Ours (2006) showed evidence of a correlation between unobservable factors that influence initiation and cessation. In this case the unobserved factors are interpreted as an individual's preference for experimentation, such that a higher propensity to experiment accelerates both the time to starting and the time to quitting.

Latent factor models have been employed to account for endogeneity of regressors in simultaneous equations models and selection bias due to unmeasured variables. For example, many studies of utilization of healthcare take into account the endogeneity of the insurance status that is chosen by the patient and estimate models for binary, continuous or count outcome variables and endogenous binary regressors (see, for example, Goldman (1995), Mello *et al.* (2002) and Deb and Trivedi (2006)).

The duration analysis literature widely uses mixture models to deal with this issue (see, for example, van den Berg (2001)). In particular, frailty models include unobservable heterogeneity as a multiplicative random effect in the conditional hazard function. In mixture models, unobservable heterogeneity is integrated out either by specifying a (potentially incorrect) parametric distribution and evaluating the likelihood by quadrature or simulation methods, or

by approximating the heterogeneity distribution by discrete mass points (see, for example, Heckman and Singer (1984)). In particular, Mroz (1999) proposed a discrete factor approximation in simultaneous equation models to estimate the effect of a binary regressor on a continuous outcome. In the context of quality of care in hospitals, Picone *et al.* (2003) used a discrete factor model, where intensity of treatment and length of stay are treated as endogenous, and estimated a model for a binary dependent variable and two continuous endogenous variables. In the smoking literature, Gilleskie and Strumpf (2005) used a latent factor specification in a model for smoking behaviour where endogenous smoking history and unobservable individual heterogeneity are taken into consideration. Similar applications can be found also in van Ours (2003, 2004, 2006). Studying dynamics in the use of drugs and wage effects, he estimated bivariate duration models for the duration of non-use of two drugs and, more recently, for starting and quitting drug use. Any parametric assumption on heterogeneity is relaxed and the joint model is estimated in a semiparametric framework. The advantage of using this methodology is that no parametric assumptions are required about the distribution of the unobserved heterogeneity, whereas the standard maximum likelihood estimator relies on joint normality or other parametric assumptions.

3.3. Latent factor specification

We fit a latent factor model for the joint distribution, which is made up of the system of equations (1)–(4). This is based on the assumption that the error process depends on common latent factors that influence both smoking behaviours and mortality:

$$\varepsilon_{ij} = l_j + \omega_{ij} \qquad j = 1, \ldots, 4, \tag{5}$$

where ε_{ij} is the overall error for individual i in equation j, ω_{ij} is the independent component of this error and l_j is the latent factor; the latent factors $l_1, \ldots, l_4$ are associated and the same for each individual. Given independence of the ω_{ij}, the joint distribution of the errors can be written as

$$f(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4) = \int_{-\infty}^{\infty} f(\varepsilon_1|l_1)\, f(\varepsilon_2|l_2)\, f(\varepsilon_3|l_3)\, f(\varepsilon_4|l_4)\, g(l_1)\, \mathrm{d}l_1\, g(l_2)\, \mathrm{d}l_2\, g(l_3)\, \mathrm{d}l_3\, g(l_4)\, \mathrm{d}l_4. \tag{6}$$

This allows us to integrate the unobservables out of the model, treating them as random factors following Heckman and Singer (1984) and Mroz (1999). To evaluate the integral in equation (6), which does not exist in closed form, we use a semiparametric technique. We estimate a discrete latent factors model (DLFM), which offers an alternative to parametric approaches and has the advantage of reducing the bias in identification of the distribution of the latent factors when they are non-normal (see Mroz (1999)). The DLFM is based on a finite density estimator that approximates the unknown distribution of l_j by using a step function based on K location mass points, η_k:

$$\Pr(l_j = \eta_k) = \pi_k \qquad j = 1, \ldots, 4,\ k = 1, \ldots, K. \tag{7}$$

It follows that the individual contribution to the sample likelihood in the DLFM is given by

$$L_i = \sum_{k=1}^{K} \pi_k\, f_k(\cdot) \tag{8}$$

where π_k are mixing probabilities, each $\pi_k \geqslant 0$ and $\Sigma_{k=1}^{K} \pi_k = 1$. The component distribution $f_k(\cdot)$ in the mixture hazard model depends on the mass points η_k and is defined as

$$f_k(\cdot) = \{1 - \Pr(s=1|\mathbf{x}_1, \eta_k)\}^{1-s} \{f(t_1|\mathbf{x}_2, \eta_k) \Pr(s=1|\mathbf{x}_1, \eta_k)\}^{s} \{h(t_2|t_1, \mathbf{x}_3, \eta_k)^q \times S(t_2|t_1, \mathbf{x}_3, \eta_k)\}^{s} h(t_3|s, q, \mathbf{x}_4, \eta_k)^d \frac{S(t_3|s, q, \mathbf{x}_4, \eta_k)}{S(\tau|s, q, \mathbf{x}_4, \eta_k)}. \quad (9)$$

We assume that the latent factors in expression (5) have a factor loading specification such as

$$l_j = \rho_j u + \alpha_j v \quad (10)$$

where u and v are Bernoulli random variables which take value 1 with probability p_1 and p_2, and value 0 with probability $1 - p_1$ and $1 - p_2$ respectively. Given the factor loading specification of the latent factors, for each equation there are four location mass points ($K = 4$) because this is the number of possible combinations of u and v. Factor loadings measure the effects of the common factors and can be interpreted as coefficients for the omitted unobserved variable. Location mass points are the point of support for the discrete distribution. It follows that $\pi_1 = \Pr(u=0, v=0) = (1-p_1)(1-p_2)$ is the probability that $l_j = 0$. Analogously, $l_j = \rho_j$ with probability $\pi_2 = \Pr(u=1, v=0) = p_1(1-p_2)$, $l_j = \alpha_j$ with $\pi_3 = \Pr(u=0, v=1) = (1-p_1)p_2$ and $l_j = \rho_j + \alpha_j$ with $\pi_4 = \Pr(u=1, v=1) = p_1 p_2$. Expression (8) can be expressed in terms of summations over u and v:

$$L_i = \sum_{u=0}^{1} \sum_{v=0}^{1} p_1^u (1-p_1)^{1-u} p_2^v (1-p_2)^{1-v} \{1 - \Pr(s=1|\mathbf{x}_1, \rho_1 u + \alpha_1 v)\}^{1-s} \times \{f(t_1|\mathbf{x}_2, \rho_2 u + \alpha_2 v) \Pr(s=1|\mathbf{x}_1, \rho_1 u + \alpha_1 v)\}^{s} \{h(t_2|t_1, \mathbf{x}_3, \rho_3 u + \alpha_3 v)^q \times S(t_2|t_1, \mathbf{x}_3, \rho_3 u + \alpha_3 v)\}^{s} h(t_3|s, q, \mathbf{x}_4, \rho_4 u + \alpha_4 v)^d \frac{S(t_3|s, q, \mathbf{x}_4, \rho_4 u + \alpha_4 v)}{S(\tau|s, q, \mathbf{x}_4, \rho_4 u + \alpha_4 v)} \quad (11)$$

where the first part of the formula is for the probability mass functions of the random factors u and v.

The distribution of the latent factors is not identified without further assumptions. The location of l_j is arbitrary when each equation has an intercept, and the scale of l_j is also indeterminate. Therefore, identification requires a normalization that implies restricting the support of l_j (see Mroz (1999)). This implies restricting the range of mass points. We do this by fixing two mass points to 0 and 1 and through the assumption that u and v are Bernoulli random variables: $\eta_1 \equiv (u=0, v=0)$, $\eta_2 \equiv (u=1, v=0)$, $\eta_3 \equiv (u=0, v=1)$ and $\eta_4 \equiv (u=1, v=1)$. Identification of all parameters also requires parameterizing the probabilities p_1 and p_2, from which π_k are derived, by using a logistic distribution (other distributions could be used for this purpose): $p_j = \exp(\zeta_j)/\{1 + \exp(\zeta_j)\}$, where $j = 1, 2$. The ζ_j are additional parameters to estimate together with the factor loadings (ρ_j, α_j) and allow the mixing probabilities π_k to be recovered. Semiparametric identification of the distribution of unobservable heterogeneity in models with treatment times and a factor loading specification of the unobservable heterogeneity is discussed in Heckman and Navarro (2007).

This version of the DLFM has the advantage of being very general and allows testing restrictions on the l_j: for example, one might suppose that unobservables that influence the probability of starting have a common effect on the age at starting, thus imposing $l_1 = l_2$.

3.4. Measuring association due to unobservables

We measure the association between the smoking probability, the hazards of starting and quitting and the hazard of dying due to unobservable heterogenity. The variance–covariance matrix

for the common latent factors depends on the distribution of the random factors u and v as well as on the factor loadings. The diagonal elements of the matrix are the variances of each latent factor, $V_{(l_j)} = \rho_j^2 V_{(u)} + \alpha_j^2 V_{(v)}$, whereas the off-diagonal elements describe covariances between the latent factors $\text{cov}(l_j, l_m) = \rho_j \rho_m V_{(u)} + \alpha_j \alpha_m V_{(v)}$ for each $j \neq m$ where $V_{(u)} = p_1(1 - p_1)$ and $V_{(v)} = p_2(1 - p_2)$ are the variances of u and v. The cross-products of the factor loadings (ρ and α) determine the sign of the association between the latent factors. Correlation coefficients provide more intuitive measures of association than simple covariances, since they measure the strength of the association between the latent factors. They are computed as $\text{corr}(l_j, l_m) = \text{cov}(l_j, l_m)/\text{sd}_{(l_j)}\text{sd}_{(l_m)}$ where $\text{sd}_{(l_j,m)}$ are the standard deviations of the latent factors. These correlations measure the association between response variables which is caused by unobserved third factors.

3.5. Specification

In principle, the model is identified solely by functional form of the components of the mixture, where the mixture itself is identified by the factor loading specification. Non-linearity of each equation guarantees identification of the multiple-hazard models. However, estimators that rely on functional form for identification may be poorly identified and highly sensitive to misspecification, so stronger identification assumptions may be required to achieve reliable estimates. These identification assumptions often take the form of exclusion restrictions, that imply that the excluded variables do not have a direct effect on the outcome of interest. A comparison of various versions of the model, with and without exclusion restrictions, shows that estimation of the main parameters of interest is robust to the choice of the exclusion restrictions. Conditioning on exogenous and endogenous covariates and allowing for dependence between the hazard functions, the DLFM implies the parameterization

$$\Pr(s = 1) = \Phi(\beta_1' \mathbf{w} + \theta_1' \mathbf{z}_1 + \rho_1 u + \alpha_1 v), \tag{12}$$

$$\phi = \exp(-\beta_2' \mathbf{w} - \theta_2' \mathbf{z}_2 - \rho_2 u - \alpha_2 v), \tag{13}$$

$$\psi = \exp(-\delta_1 t_1 - \beta_3' \mathbf{w} - \theta_3' \mathbf{z}_3 - \rho_3 u - \alpha_3 v), \tag{14}$$

$$\varphi = \exp(\delta_2 s + \delta_3 q + \beta_4' \mathbf{w} + \rho_4 u + \alpha_4 v) \tag{15}$$

where $\mathbf{w}$ is a matrix of exogenous variables that affect all hazards and $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$ are matrices of variables that have a direct effect on smoking behaviours but do not affect mortality.

Ideally, analysis of the dynamics of smoking requires information that is recorded at the time when the individual started and quit smoking (for example the price of tobacco, family situation, social context, peer influences and so on would be desirable). This is not available in the data set that we use, however. Here, initiation of smoking and cessation depend on demographics, socio-economic controls (social class based on occupation according to the Registrar General's social class classification, level of education and occupational status), geographical variables and marital status at 1984–1985, under the assumption that they reflect past social and economic background. These variables are all included in $\mathbf{w}$.

Smoking behaviour of parents or other member of the household has a direct effect on child smoking attitudes and intertemporal preferences for health. Dummy variables for mother's, father's and both parents' smoking behaviour in the past years are included in $\mathbf{z}_1$ and $\mathbf{z}_2$ but are excluded from $\mathbf{z}_3$ as they are assumed to have a direct effect on age at starting but not on

cessation of smoking. Regular smoking behaviour of anyone else in the household at the time of the survey, instead, should influence the time to quit smoking and goes, therefore, in $\mathbf{z}_3$.

To capture time trend effects in smoking hazard equations we include dummy variables for calendar year of starting. In fact, initiation of smoking and cessation might be influenced by the calendar year of starting. For example, in the UK, 1954 was the year when a key report on the health risks of smoking was published (Doll and Hill, 1954). In the same year, the Minister of Health reported on the findings of a government-approved scientific committee about the relationship between smoking and lung cancer (Minister of Health-Britannic Government, 1954). This would have influenced awareness of the consequences on health of smoking.

Birth cohort dummy variables are used to capture the effect of age. However, these only appear in the probit for initiation of smoking and the mortality hazard models: this avoids problems of collinearity which may arise, including in the same model calendar year of starting and birth cohort dummy variables as well as age of starting.

4. Results

The empirical results that are presented here follow exploratory analysis on the subsample of smokers that we have done using a more restrictive one latent factor model estimated by applying Mroz's (1999) approach to duration data. The model was also estimated with the maximum simulated likelihood method and Gauss–Hermite quadrature, assuming a normal distribution for the heterogeneity, to check robustness of the model to the choice of a continuous or discrete distribution for the unobservable heterogeneity.

Parameter estimates of the two-part model for smoking initiation are shown in the second and fourth columns of Table 2: the second column refers to the probability of starting; the fourth refers to the duration model for age at starting in the subsample of smokers. Results show the substantial effect of parental smoking variables on age at initiation of smoking: the time to starting accelerates for individuals whose father or mother, or both parents, used to smoke. The coefficients of the dummy variables for calendar year of starting show a positive time trend in age at starting: the age at starting rises monotonically as we move from the earliest years of starting (1924–1937) until the time of the survey. The duration dependence parameter γ_1 is positive and smaller than 1, predicting first a rise and then a monotone decrease in the baseline hazard as the time spent non-smoking increases.

The sixth column of Table 2 reports results for the smoking cessation model. The coefficient of starting describes the effect of changes in age at starting on the hazard of quitting. This implies that smokers who started at younger ages, *ceteris paribus*, tend to spend more years smoking. Since starting is measured on the logarithmic scale, the AFT coefficient can be directly transformed in the PH coefficient to obtain the elasticity of the hazard of quitting with respect to age at starting, which is about 0.54. Shorter durations of smoking are predicted for men and individuals with highly skilled occupations, whereas longer durations are predicted for the low educated, unemployed and not-married individuals. As expected, the presence of other smokers in the household decelerates the time to quitting. We find that time spent smoking decreases monotonically with year of starting and this effect is substantial if we consider calendar years of starting which are subsequent to 1954 when the information dissemination policy on smoking was introduced. The estimate of the duration dependence parameter γ_2 suggests that the quitting hazard rate increases at a decreasing rate with years spent smoking. A comparison with the parameter estimates from the single-hazard model for cessation of smoking shows that the elasticity of age at starting is overestimated when unobservable heterogeneity is not included in the model.

Table 2. Results from the DLFM

Variable	*Smoking initiation*				*Smoking cessation, hazard of quitting, Weibull model (AFT metric)*		*Mortality, hazard of dying, Gompertz model (PH metric)*	
	Probability of smoking, probit model		*Hazard of starting, log-logistic model (AFT metric)*					
	Coefficient	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
start							1.117†	0.154
quit							−0.341†	0.071
starting					−0.279‡	0.109		
sc1	−0.352†	0.111	−0.009	0.011	−0.088‡	0.036	−0.126§	0.066
sc3	0.188§	0.103	−0.004	0.011	0.025	0.036	0.075	0.058
degree	−0.178	0.186	0.022	0.020	−0.086	0.063	−0.065	0.128
hvqA	0.088	0.181	0.037§	0.019	0.024	0.063	−0.209	0.130
no edu	0.393‡	0.165	0.006	0.016	0.089§	0.053	0.052	0.102
other edu	0.264	0.238	0.028	0.023	0.135§	0.079	0.052	0.141
part time	0.365‡	0.156	0.074†	0.016	−0.008	0.056	−0.042	0.110
unemployed	0.473	0.290	−0.041§	0.023	0.247†	0.096	0.478†	0.151
sick	1.078†	0.367	0.064†	0.024	0.108	0.079	0.722†	0.125
retired	0.900†	0.225	0.284†	0.016	0.052	0.058	−0.178§	0.100
housekeeper	0.323‡	0.164	0.017	0.018	0.029	0.067	0.250‡	0.127
rural	−0.393†	0.127	−0.009	0.012	−0.062	0.040	−0.061	0.068
suburb	−0.165§	0.099	0.000	0.010	−0.057§	0.033	−0.060	0.054
household size	−0.145†	0.053	−0.042†	0.005	−0.024	0.017	0.020	0.030
male	1.426†	0.201	−0.025‡	0.012	−0.159†	0.039	0.289†	0.066
widow	−0.308‡	0.143	0.044†	0.016	0.092§	0.049	0.107	0.071
sepdiv	0.005	0.192	−0.065†	0.020	0.355†	0.084	−0.020	0.129
single	−0.499†	0.186	−0.005	0.019	0.148‡	0.065	0.224‡	0.094
mother smoked	0.981†	0.238	−0.084†	0.024				
father smoked	0.745†	0.139	−0.051†	0.014				
both smoked	1.125†	0.183	−0.098†	0.016				
others smoked					0.423†	0.035		
sy2437			0.210†	0.022	−0.131‡	0.064		
sy3844			0.433†	0.024	−0.251†	0.086		
sy4554			0.568†	0.027	−0.259‡	0.101		
sy55			0.711†	0.029	−0.601†	0.118		
bc20	1.021†	0.187					−0.348†	0.083
bc30	0.978†	0.254					−0.620†	0.127
bc40	1.473†	0.338					−0.832†	0.185
constant	−2.031†	0.435	2.323†	0.045	4.917†	0.232	−10.523†	0.383
Duration dependence parameters								
γ_1			0.112†	0.004				
γ_2					1.947†	0.048		
γ_3						0.094†	0.004	
Log-likelihood	−25148.465							
N	4572							

†Significant at the 1% level.
‡Significant at the 5% level.
§Significant at the 10% level.

The penultimate column of Table 2 refers to the mortality hazard model. The estimated duration dependence γ_3 is positive and can be interpreted as a genuine life cycle effect: the hazard of dying increases with aging with a rotated L-shape. The estimated coefficients show that the mortality hazard is significantly higher for men, unemployed, housekeepers, individuals on sick leave and singles. The hazard increases for current smokers and the hazard rate ratio, which is calculated as the exponential of the coefficient, is 3.06 relative to never smokers. The hazard of dying is lower for those who quit smoking relative to those who are currently smoking, and higher relative to never smokers. In particular, we find that the hazard rate ratio is about 2.2 relative to never smokers. We compare these results with those from the single-hazard model and find that when unobservable heterogeneity is ignored the coefficient of start is underestimated, and the coefficient of quit as well as the coefficients of the socio-economic variables and gender are overestimated.

4.1. Unobservable heterogeneity and latent classes

Table 3 reports estimates for the parameters which describe the distribution of l_j, most of which are statistically significant. The mixing proportions and the location mass points highlight differences between heterogeneous population subgroups and allow us to distinguish between four latent classes.

All latent classes are to be interpreted in comparison with the class that is described by the mass point normalized to 0 that, in our model, is associated with the highest mixing probability (56%). The second mass point shows that about 11% of the sample are drawn from a

Table 3. Unobservable heterogeneity—estimated parameters from the DLFM

Parameter	*Smoking initiation*				*Smoking cessation, hazard of quitting (AFT metric)*		*Mortality, hazard of dying (PH metric)*	
	Probability of smoking		*Hazard of starting (AFT metric)*					
	Coefficient	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*	*Coefficient*	*Standard error*
Mass points								
η_1	0		0		0		0	
η_2: ρ_j	−8.851	25.605	−0.939†	0.055	−0.056	0.224	0.395†	0.235
η_3: α_j	6.683	25.603	0.176†	0.023	−0.118	0.099	−0.716†	0.164
η_4: $\rho_j + \alpha_j$	−2.018†	0.375	−0.763†	0.053	−0.174	0.180	−0.321	0.236
u- and v-probabilities								
ζ_1	−1.650†	0.191						
ζ_2	−0.685†	0.231						
p_1	0.161†	0.026						
p_2	0.335†	0.051						
Mixing probabilities								
π_1					0.558†	0.051		
π_2					0.107†	0.017		
π_3					0.281†	0.042		
π_4					0.054†	0.014		

†Significiant at the 1% level.

Table 4. Correlation coefficients—corr (l_j, l_m)

	l_1	l_2	l_3	l_4
l_1	1			
l_2	0.855	1		
l_3	−0.413	0.118	1	
l_4	−0.927	−0.599	0.724	1

subgroup of individuals who are more likely to die and less likely to smoke, but if they smoke they start sooner in life and quit sooner. The third mass point has a mixing probability of 28% and describes a latent class of individuals who are less likely to die, more likely to smoke, but start smoking later in life and smoke for a shorter time. The last latent class is captured by the fourth mass point which has the lowest mixing probability (about 5%): individuals in this class are less likely to die, less likely to smoke and start and quit smoking sooner.

For a more intuitive interpretation of the role of the unobservable heterogeneity in our model we look at correlations between latent factors, which are reported in Table 4. The correlation coefficients should be interpreted as the effects of unobservables on increasing (or decreasing) the probability of starting and the hazard of mortality, and accelerating (or decelerating) the time to starting and time to quitting. The correlations $\text{corr}(l_1, l_2)$ and $\text{corr}(l_1, l_3)$, between the probability of smoking and the smoking hazards, are positive and negative respectively. This indicates that unobservable factors which increase the likelihood of starting smoking also increase the age at starting and decrease the number of years spent smoking (i.e. smokers start late and quit early). $\text{corr}(l_2, l_3)$ can possibly be interpreted, according to van Ours (2006), as a preference for evidence on individual preferences for experimentation driving smoking behaviour: smokers who want to experiment with cigarettes start soonest and quit early. The correlations between latent factors in the mortality and smoking initiation equations are negative. $\text{corr}(l_1, l_4)$ is close to -1 and can be interpreted in terms of unobservable frailty which would increase the hazard of dying and lower, at the same time, the probability of smoking. This gives evidence on selection of frailer individuals into non-smoking behaviour. In contrast, $\text{corr}(l_2, l_4)$ gives evidence on selection of frailer smokers into early smoking. This seems to be supported by $\text{corr}(l_3, l_4)$: unobservables that increase the hazard of dying also increase the duration of smoking (i.e. unobservable frailty drives selection of smokers into smoking longer). These results could also be interpreted in terms of the opportunity cost of smoking, the latter being lower for people with low life expectancy.

5. Using the model to predict the distribution of outcomes

In this section we present a post-estimation analysis of the role of parental smoking as a factor influencing inequality of opportunity in smoking-related mortality. Our aim is to isolate the partial effect of changes in parental smoking on the individual's own hazard rates for smoking and mortality, thus shedding some light on the intergenerational transmission of smoking. This follows the recent literature in health economics (see, for example, Rosa Dias (2009) and Trannoy *et al.* (2010)).

We use the estimated model to predict the median age at starting, duration of smoking and lifespan for hypothetical individual types. These predictions use the estimated coefficients from the DLFM and focus on the change in median survival times in the absence of parental smoking. Characteristics other than parental smoking and the individuals' own smoking are fixed at

specific values. To simplify the presentation of results we consider four socio-economic types. The four individual types, which are shown in Table 5, differ on the basis of socio-economic status, as approximated by the sc1- and sc3-variables, and gender, and are equal in all other characteristics: they are married, live in a city, have high education level, have a full-time job, year of birth is pre 1920, year of smoking initiation is post 1954 (other observed characteristics are set at their sample mean value). The predictions from the model work recursively: changes in parental smoking affect the probability of being a smoker and the age of smoking; these then affect the probability of quitting and the number of years smoked; the probabilities of starting and quitting smoking, in turn, have a direct effect on mortality.

Irrespective of socio-economic status and gender, individuals whose parents did not smoke have a higher predicted median age at starting, a lower predicted median time spent smoking and a higher predicted median lifespan. The highest median age at starting is predicted for women and whose parents never smoked (about 20 years), and the lowest is for men whose parents both smoked (about 15 years). Overall, parental smoking lowers the median age at starting by about 4 years. Parental smoking also increases the median time spent smoking for each individual type by about 14 years. The highest median smoking duration is predicted for individuals in sc3, whose parents smoked (34.2 and 39.8 years).

Table 5 also reports the median lifespan by smoking status (current, former and never smoker) and for the whole sample (in which case the median lifespan is averaged across the three types of smoker). The exercise predicts that the median lifespan is longer for those who never smoked. For each individual type, former smokers have a higher median lifespan than current smokers. The shortest median lifespan (67.6 years) is predicted for men in sc3 and who are current smokers. Our analysis highlights a large difference in median survival between current and never smokers. Other things being equal, this difference is about 12 years. We find a smaller difference in median survival between current and former smokers: this is about 3.6 years irrespective of parental smoking. These magnitudes are in line with the epidemiological literature such as Doll *et al.* (2004). Overall, we find that the indirect effect of parental smoking on median lifespan

Table 5. Simulations of predicted median survival times for individual types

Coefficient	*Results for parents who*		*Results for parents who*	
	Smoked	*Did not smoke*	*Smoked*	*Did not smoke*
	Type 1: male, sc1		*Type 2: female, sc1*	
starting	15.2	19.2	15.6	19.7
sm_years	30.5	18.7	35.6	21.8
lifespan	72.2	80.8	79.4	84.7
current smokers	69.7		72.8	
former smokers	73.4		76.4	
never smokers	81.6		84.7	
	Type 3: male, sc3		*Type 4: female, sc3*	
starting	15.3	19.3	15.7	19.8
sm_years	34.2	21.0	39.8	24.4
lifespan	69.6	77.4	75.3	82.4
current smokers	67.6		70.7	
former smokers	71.2		74.3	
never smokers	79.5		82.6	

is negative for each individual type considered: calculations show a reduction in their children's median lifespan of about 7 years for all individuals.

5.1. Role of parental smoking in inequality of opportunities in smoking-related mortality

We extend the analysis of the model to measure the partial effects of parental smoking and smoking on inequality of opportunity in mortality. This implies working directly with the distribution of mortality. For the purpose of the analysis we use both the distribution of predicted survival and the distribution of predicted median lifespan and simulations of these quantities. We predict two scenarios: the best circumstances and the best effort. Under the first scenario everyone is given the best circumstances, with respect to parental smoking, by setting parental smoking variables to 0 and then using the model to predict their smoking and lifespan. Under the second scenario everyone is given the level of smoking that corresponds to the highest level of effort, whatever their circumstances, so the smoking variables are set to 0. Results are shown in Table 6. We calculate overall inequality in mortality and explore inequality of opportunities by assessing generalized Lorenz curve (GLC) dominance.

Overall inequality in health (or, in our case, mortality) can be investigated by using the Gini coefficient G, which is an index borrowed from the literature on income inequality (Wagstaff *et al.*, 1991). The Gini coefficient is twice the area between the Lorenz curve and the line of complete equality, where the Lorenz curve plots cumulative shares of health against relative rank in the distribution of health. The Gini coefficient is a measure of relative inequality and does not address the potential for an 'equity–efficiency trade-off' between the average health of the population and levels of health inequality (Wagstaff, 2002). This can be done by looking at generalized Lorenz dominance and using an achievement index. In our case, the GLC multiplies the Lorenz curve by the average of lifespan or survivor probability: the vertical height of the curve shows the mean level of lifespan and the curvature reflects inequality in mortality. GLCs allow generalized dominance analysis: the dominating GLCs indicate distributions which are more efficient and less unequal (i.e. a higher level of welfare). The corresponding achievement index is Sen's (1976) welfare index $\mu(1-G)$, which captures both the mean of the distribution (μ) and the level of inequality, so that both concerns are captured in a single index.

Table 6 shows the mean value, the Gini coefficient for overall inequality and Sen's welfare index for the distributions under the two hypothetical scenarios best circumstances and best effort and the baseline distribution which is predicted by using actual sample values for parental

Table 6. Distributional analysis: simulation results

Scenario	*Average predicted survivor probability:* $\bar{S}(t)$	*Gini G*	*Sen's welfare index* $\bar{S}(t)(1-G)$
Actual data	0.67	0.209	0.53
Best circumstances	0.72	0.178	0.59
Best effort	0.79	0.131	0.69
	Average predicted median lifespan: $\bar{t}$ *(years)*	*Gini G*	*Sen's welfare index* $\bar{t}(1-G)$ *(years)*
Actual data	79.5	0.041	76.2
Best circumstances	82.2	0.039	79.0
Best effort	86.0	0.027	84.7

and individual smoking. The top panel of Table 6 shows that the average predicted survivor probability is higher if respondents never smoked and if their parents did not smoke. The Gini coefficient is about 15% lower in the best circumstances scenario (0.178), than in the case of predictions from the baseline model (0.209), meaning that inequality would be lower if all parents were non-smokers. Inequality is 37% lower when nobody smokes ('best effort'), where the Gini coefficient is 0.131. The same exercise is replicated in the bottom panel of Table 6 for predicted median lifespan and results from these predictions mirror those for the average survivor probability. The median lifespan is higher in the best circumstances and best effort scenarios (about 82.2 and 86.0 years respectively) than in the baseline model predictions (about 79.5 years). Inequality is about 5% lower in 'best circumstances' and 34% in 'best effort', where the Gini coefficient is 0.039 and 0.027 respectively.

The relative difference between the best circumstances and the baseline model predictions measures the gain in health due to not having parents who smoke (i.e. due to that specific set of circumstances): this represents a gain in health of about 7% in terms of average survivor probability and about 3% in terms of average median lifespan. Similarly, we can derive the gain in health due to not smoking (i.e. due to adopting the best level of effort, whatever the circumstances) as the relative difference between the 'best effort' and the baseline model predictions. The gain in health, in this case, is about 18% in terms of average survivor probability and 5% in terms of predicted median lifespan.

In the top panel of Table 6, the best circumstances and best effort scenarios have Sen indexes of 0.59 and 0.69 respectively, meaning that the unequal distributions, when parental smoking is set to 0 and when individuals have never smoked, are equivalent to equally distributed survivor probabilities of 59% and 69%, compared with 53% in the baseline data. In the bottom panel, Sen's index suggests an equivalence between the median lifespan of 76 years in the actual data and 79 years if all parents were non-smokers and 84.7 years if nobody smoked. Fig. 2 shows that the GLCs for survivor probability and median lifespan that are derived from the baseline data are always dominated by the GLCs for the hypothetical scenarios.

6. Conclusions

We use the British HALS to investigate the relationship between smoking and individual mortality and study the role of parental smoking in inequality of opportunity in smoking-related mortality.

A recursive system of multiple-hazard regressions (starting, quitting and dying) is estimated. This allows for a mixture of hazard functions depending on observed individual characteristics and latent factors representing unobservable time invariant heterogeneity. A DLFM is used, with the advantage of not relying on any parametric assumption on the latent factors.

We find that the earlier individuals start smoking the less likely they are to quit, and that dynamics in tobacco consumption should, therefore, drive policies that aim at reducing smoking. We also find that smoking status matters: compared with never smokers, both current and former smokers have a higher mortality hazard.

Predictions from the model, based on the DLFM estimates, show that parental smoking decreases the predicted median age at starting and increases the median time spent smoking, independently of gender and socio-economic status. A substantial gap in predicted median lifespan is found between individuals with different smoking status. This favours never smokers and, to a lower extent, former smokers relative to current smokers. Our study of inequality of opportunities in predicted survivor probability and median lifespan shows that, on average, the children of smokers live shorter lives. The GLCs in the predicted scenarios, i.e. when neither

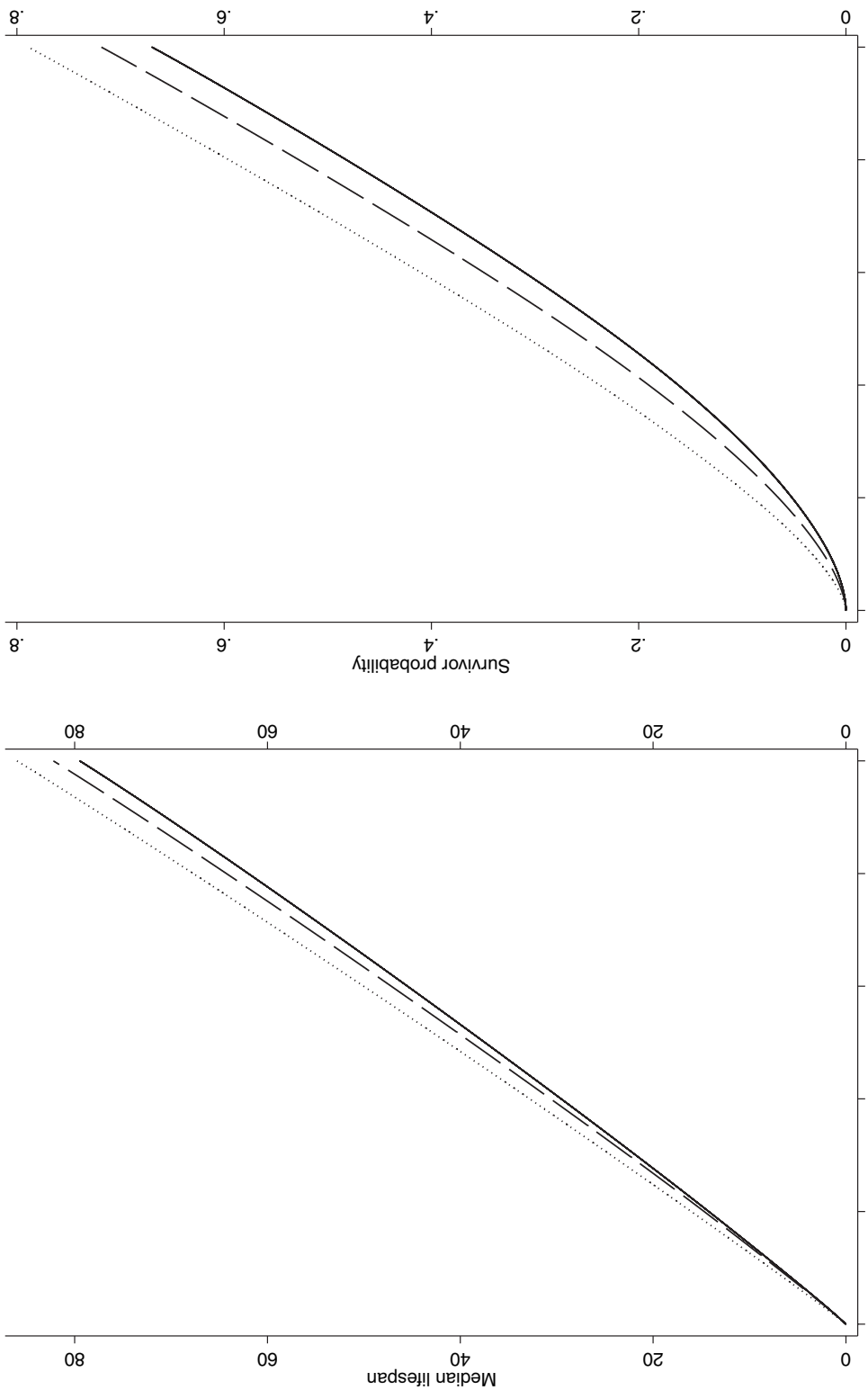

Fig. 2. GLCs: ——, current data; — —, best circumstances; · · · · · · , best effort

parents smoked (the best circumstance) or when all individuals have never smoked (best effort), dominate the GLCs that are derived from the model using the actual sample characteristics. Furthermore, our model predicts substantial gains in health associated with parental non-smoking behaviour and, mainly, to own non-smoking habits on both survivor probability and median lifespan. The results of our analysis may be relevant for health policy makers whose aim is effectively to reduce inequality in smoking-related mortality. The findings suggest that preventing the early initiation of smoking has an effect beyond the targeted individuals and has the long-term benefit of reducing inequality of opportunity for their children and subsequent generations.

Acknowledgements

Data from the HALS were supplied by the Economic and Social Research Council Data Archive. We thank participants in the Centro Richerche Economiche Nord Sud Workshop in Health Economics (2008) and in the Italian Health Economics Association 13th annual conference (2008), Nigel Rice, Maarten Lindeboom, Donna Gilleskie, Anirban Basu, Christopher Auld and other participants in the first Annual Health Econometrics Workshop meeting in Chicago (2009) for their helpful comments. We gratefully acknowledge funding from the Economic and Social Research Council under grant RES-060-25-0045.

References

Adda, J. and Lechene, V. (2001) Smoking and endogenous mortality: does heterogeneity in life expectancy explain differences in smoking behavior? *Discussion Paper 77*. Department of Economics, University of Oxford, Oxford.

Adda, J. and Lechene, V. (2004) On the identification of the effect of smoking on mortality. *Working Paper CWP13/04*. Centre for Microdata Methods and Practice, Institute for Fiscal Studies, London.

Auld, M. C. (2006) Using observational data to identify the effects of health-related behavior. In *Elgar Companion to Health Economics* (ed. A. Jones). Cheltenham: Elgar.

Balia, S. and Jones, A. M. (2008) Mortality, lifestyle and socio-economic status. *J. Hlth Econ.*, **27**, 1–26.

van den Berg, G. J. (2001) Duration models: specification, identification and multiple durations. In *Handbook of Econometrics*, vol. 5 (ed. Z. Griliches), pp. 3381–3460. Cambridge: Harvard University Press.

Cox, B., Blaxter, M., Buckle, A., Fenner, N., Golding, J., Gore, M., Huppert, F., Nickson, J., Roth, M., Stark, J., Wadsworth, M. and Whichelow, M. (1987) *The Health and Lifestyle Survey*. London: Health Promotion Research Trust.

Cox, B., Huppert, F. and Whichelow, M. (1993) *The Health and Lifestyle Survey: Seven Years On*. Aldershot: Dartmouth Press.

Deb, P. and Trivedi, P. K. (2006) Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: application to health care utilization. *Econmetr. J.*, **9**, 307–331.

Doll, R. and Hill, A. B. (1954) The mortality of doctors in relation to their smoking habits. *Br. Med. J.*, **228**, 1451–1455.

Doll, R., Peto, R., Boreham, J. and Sutherland, I. (2004) Mortality in relation to smoking: 50 years' observations on male British doctors. *Br. Med. J.*, **328**, 15–19.

Douglas, S. and Hariharan, G. (1994) The hazard of starting smoking: estimates from a split population duration model. *J. Hlth Econ.*, **13**, 213–230.

Forster, M. and Jones, A. M. (2001) The role of tobacco taxes in starting and quitting smoking: duration analysis of British data. *J. R. Statist. Soc.* A, **164**, 517–547.

Gilleskie, D. and Strumpf, K. L. (2005) The behavioral dynamics of youth smoking. *J. Hum. Resour.*, **40**, 822–866.

Goldman, D. (1995) Managed care as a public cost-containment mechanism. *Rand J. Econ.*, **26**, 277–295.

Grossman, M. (1972) On the concept of health capital and the demand for health. *J. Polit. Econ.*, **80**, 223–255.

Heckman, J. J. and Navarro, S. (2007) Dynamic discrete choice and dynamic treatment effects. *J. Econmetr.*, **136**, 341–396.

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

Keyes, M., Legrand, L. N., Iacono, W. G. and McGue, M. (2008) Parental smoking and adolescent problem behavior: an adoption study of general and specific effects. *Am. J. Psyhiatr.*, **165**, 1338–1344.

Kunst, A., Giskes, K. and Mackenbach, J. (2004) Socio-economic inequalities in smoking in the European Union: applying an equity lens to tobacco control policy. Department of Public Health, Erasmus Medical Center, Rotterdam.

Mello, M. M., Stern, S. and Norton, E. (2002) Do Medicare HMO still reduce health service use after controlling for selection bias? *Hlth Econ.*, **11**, 323–340.
Minister of Health-Britannic Government (1954) Memorandum on tobacco smoking and cancer of the lung to the cabinet home affairs committee. *Technical Report CAB 129–66*. London.
Mroz, T. A. (1999) Discrete factor approximations in simultaneous equation models: estimating the impact of a dummy endogenous variable on a continuous outcome. *J. Econmetr.*, **92**, 233–274.
van Ours, J. C. (2003) Is cannabis a stepping-stone for cocaine? *J. Hlth Econ.*, **22**, 539–554.
van Ours, J. C. (2004) A pint a day raises a man's pay; but smoking blows that gain away. *J. Hlth Econ.*, **23**, 863–886.
van Ours, J. C. (2006) Dynamics in the use of drugs. *Hlth Econ.*, **15**, 1283–1294.
Peto, R., Lopez, A. D., Boreham, J. and Thun, M. (2006) *Mortality from Smoking in Developed Countries 1950-2000*. Oxford: Oxford University Press.
Picone, G., Sloan, F., Chou, S. and Taylor, D. (2003) Does higher hospital cost imply higher quality of care? *Rev. Econ. Statist.*, **85**, 51–62.
Roemer, J. E. (1998) *Equality of Opportunity*. Cambridge: Harvard University Press.
Roemer, J. E. (2002) Equality of opportunity: a progress report. *Socl Choice Welf.*, **19**, 455–471.
Rosa Dias, P. (2009) Inequality of opportunity in health: evidence from a UK cohort study. *Hlth Econ.*, **18**, 1057–1074.
Rosa Dias, P. (2010) Modelling opportunity in health under partial observability of circumstances. *Hlth Econ.*, **19**, 252–254.
Rosa Dias, P. and Jones, A. M. (2007) Giving equality of opportunity a fair innings. *Hlth Econ.*, **16**, 109–112.
Sen, A. K. (1976) Real national income. *Rev. Econ. Stud.*, **43**, 19–39.
Shaap, M. and Kunst, A. (2009) Monitoring of socio-economic inequalities in smoking: learning from the experiences of recent scientific studies. *Publ. Hlth*, **123**, 103–109.
Trannoy, P., Tubeuf, S., Jusot, F. and Devaus, M. (2010) Inequality in opportunities in health in France: a first pass. *Hlth Econ.*, **19**, 921–938.
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. *J. Natn. Cancer Inst.*, **96**, 99–106.
Wagstaff, A. (2002) Inequality aversion, health inequalities and health achievement. *J. Hlth Econ.*, **21**, 627–641.
Wagstaff, A., Paci, P. and van Doorslaer, E. (1991) On the measurements of inequalities in health. *Socl Sci. Med.*, **33**, 545–557.