HEDG Working Paper 05/02

Mortality, lifestyle and socio-economic status

Silvia Balia
Andrew M Jones

June 2005

ISSN 1751-1976

york.ac.uk/res/herc/hedgwp

Mortality, Lifestyle and Socio-Economic Status

Silvia Balia* and Andrew M. Jones†
Department of Economics and Related Studies
University of York, York YO10 5DD, UK‡

June 2005

Abstract

This paper uses the *British Health and Lifestyle Survey* (1984-1985) data and the longitudinal follow-up of May 2003 to investigate the determinants of premature mortality risk in Great Britain and the contribution of lifestyle choices to socio-economic inequality in health. A behavioural model, which relates premature mortality to a set of observable and unobservable factors, is considered. We focus on unobservable individual heterogeneity and endogeneity affecting the mortality equation. A maximum simulated likelihood (MSL) approach for a multivariate probit (MVP) is used to estimate a recursive system of equations for mortality, morbidity and lifestyles. In order to detect inequality in the distribution of health within the population and to calculate the contribution of socio-economic factors, we compute the Gini coefficient for overall health inequality. A decomposition analysis for predicted mortality shows that, after allowing for endogeneity, lifestyles contribute strongly to inequality in mortality, reducing the direct role of socio-economic status. This contradicts the view, which is widely held in epidemiology, that lifestyles make a relatively minor contribution to observed socio-economic gradients in health.

JEL codes I1 C0
Keywords: Mortality; Lifestyle; Socio-economic status; Multivariate Probit; Simulation-based inference; Health Inequality.

E-mail address: sb231@york.ac.uk
†*E-mail address*: amj1@york.ac.uk. *Fax*: 0044-1904-433759
‡The authors wish to thank Paul Contoyannis, Fabrice Etilé, Martin Forster, Angel López and Nigel Rice for their suggestions and comments. Data from the Health and Lifestyle Survey (HALS) were supplied by the ESRC Data Archive. Any errors in the analysis and interpretation of the data presented in this paper are the responsibility of the authors.

Health inequalities have been of growing interest for the economics literature in recent years (see e.g., Deaton, 2003; Smith, 1999). Inequalities in health, for example between social groups, are partly explained by differences in lifestyle and living conditions, and lifestyle can vary between groups depending on economic circumstances (see Wagstaff, 1986). To better explain inequalities in health, it is appropriate to use a behavioural model, which contains socio-economic characteristics but also individual health decisions. Such a model will be helpful for the policy-maker, whose purpose is to improve the overall health status of the population and to reduce inequalities in health[1]. We specify a joint model of mortality, morbidity and lifestyle choices.

The importance of inequality in health stems from evidence that substantial differences in health do exist across individuals. The best way to investigate the determinants of health status in different groups of the population is to work with individual data: the British Health and Lifestyle Survey (HALS) gives us the scope for an analysis at the individual level. We exploit a longitudinal follow-up that records deaths of respondents to the original HALS study. We use data from the fourth revision of the deaths data, released in May 2003, to estimate the recursive model of mortality, morbidity and lifestyles, which is specified as a multivariate probit model and estimated by maximum simulated likelihood (MSL).

Our empirical findings contradict the view, that is widely held in the epidemiological literature, that differences in lifestyle make a relatively minor contribution to the observed socio-economic gradient in health.

Following Contoyannis and Jones (2004), we refer to lifestyle as a set of behaviours which can influence health. We assume that individuals choose these health-related behaviours. Following Fleurbaey (2004), our approach recognizes that equal opportunities for health depend on the distinction between

[1]Much of the literature is concerned with the debate about the best normative framework to achieve equity and efficiency in health. (see Hurley, 2000, for a review of the literature).

what the individual chooses and what he has no control over. On one hand, lifestyles might reflect individual preferences, on the other hand, they might reflect the constraints of economic, environmental and personal circumstances, including social origin. However, as Roemer (1998) suggested, if we look at a particular social group, there is still space for individual responsibility in the variation of the lifestyles relative to the typical behaviour of the group.

As well as contextual factors, the presence of potential risky behaviours should be considered when dealing with individual choice. People can make an unhealthy risky decision because the cost of being fully informed about the health consequences is too high. There might be a problem of incomplete or incorrect information. Moreover, the rate at which people discount the future can influence choices, especially when the decision is made much earlier than when it will have its effect on health. Fuchs (1982) suggests that the individual rate of time preference has an important role in individual health decisions. When the discount rate is low, people tend to invest more in both education and behaviours enhancing health. Hence, lifestyle and education are not the only two variables influencing health, but the rate of time preference should be considered as a potential "third variable". Information problems and time preferences are likely to be unobservable to the researcher.

The plan of the paper is as follows. The next section surveys the previous economic literature in this area. Section 1 presents the economic framework and the microeconomic concepts underlying the empirical model. Section 2 describes the Health and Lifestyle Survey dataset. Section 3 describes the variables of main interest in the analysis. Section 4 gives a simple description of the sample, considering socio-economic status, lifestyle and mortality. Section 5 describes the estimation strategy and the main results of our model. Section 6 presents the Gini measure of overall inequality in health and shows the results of the decomposition analysis. The last section concludes.

1. Previous Economic Literature

Over the last thirty years, the literature has used the concept of a health production technology through which individuals produce and modify their own health status. Health is reckoned to be both a consumption commodity and the outcome of a production process, which depends on individual behaviours. If we want to identify factors that influence good health, we should recognize the important role played by individual behaviours.

Economic theory recognizes human capital as a determinant of investment and growth, and several studies consider health not only as a good that enhances individual satisfaction, but also as a capital stock. In Becker's (1965) theory of the allocation of time, households consume and produce commodities, using goods and time as inputs. Investments in human capital might be measured in terms of market goods and of the opportunity cost of time lost for competing and alternative uses. Following this idea, Grossman (1972) emphasized the fundamental distinction between human capital and health: health can increase the possibility of consuming market goods[2]. People invest in health to reduce spells of illness and be able to consume more. Becker and Grossman's works provided the framework for further studies, which have developed the concept of the health production function, using the time of the consumer, health-related behaviours and environmental factors as inputs.

Empirical analysis faces problems with the identification of the effects that socio-economic characteristics and health-related behaviours have on health. Three serious econometric problems arise: heterogeneity, endogeneity and selection-bias. Unobservable factors might exist, leading to biased estimates and to spurious relationships with the dependent variable. Regressors are not always exogenously determined, but rather correlated with the error term. If

[2]In particular, although health is not a tradable good, it affects the quantity of time available to individuals for working.

the sample is not representative of the underlying population, on account of self-selection [3], then bias will arise.

Many empirical analyses of the determinants of health have relied on single equation instrumental variable approaches, such as 2SLS and GMM, to address the potential biases (see Auster et al., 1969; Mullahy and Portney, 1990; Mullahy and Sindelar, 1996). Our approach is different and extends the work of Contoyannis and Jones (2004). A wider concept of lifestyle is used: six, not just one, health-related behaviours are considered. They are described by reduced form equations and appear as potential endogenous regressors in the health equations. A multivariate probit model, whose specification is justified by the underlying economic theory, is implemented to allow for unobservable heterogeneity. The structure of the empirical model will be illustrated below.

2. The Economic Framework

The purpose of this paper is to investigate whether or not lifestyle choices have a direct and significant influence on the risk of mortality and to assess their contribution to the observed socio-economic gradient of health. It is well documented that individual socio-economic status is correlated with health and that health inequalities in the population are associated with socio-economic differentials (see van Doorslaer and Koolman, 2004). However, lifestyles might capture part of the variation in health that is usually attributable to changes in the socio-economic status. We will measure both lifestyle-related inequality and socio-economic related inequality in health.

[3] Grossman and Joyce (1990), in research on the first infant health production function, controlled for self-selection in the resolution of pregnancy as live births or induced abortions and in the use of prenatal medical care services.

2.1. *Measuring the Contribution of Lifestyle to Health Inequality*

Our exercise contributes to show the inadequacy of past epidemiological studies on health inequalities. They have usually used measures such as odds ratios or Agresti's Alpha to quantify the relationship between social position, work environment, education and health[4]. These measures assume a fixed ordering of social position and often do not consider intermediate categories. To evaluate the effect of including lifestyle variables on the social class gradient Borg and Kirstensen (2000) used a measure that compares odds ratios for the lowest social class relative to the highest for a model that includes lifestyles and a model that excludes them[5]. In the economic literature a similar indicator is the measure of range (see Contoyannis and Jones, 2004). It could be constructed to compare the variation of social class and education gradients, in terms of partial effects on mortality, between a model that incorporates lifestyles and a model that excludes lifestyles. Comparing the differences in the impacts of social class and education on mortality, particularly between extreme classes, in the two models, is a way to calculate the change in a range measure of inequality in health. A reduction of the range once lifestyles are added to the mortality equation would be expected. However the range is limited for at least three reasons: it fails to consider the intermediate classes (even if the gap between extreme groups remains unchanged, inequality might be bigger in the middle groups), it does not take account of the size of the groups, and finally it can be interpreted only if the underlying distribution has a clear monotonic gradient.

This paper tries to improve the analysis by computing and decomposing the Gini coefficient of total health inequality in order to assess the contribution

[4]A good review of the indicators of social position used in the epidemiological literature and methods used to measure inequality in health is offered by Manor et al. (1997).

[5]They measure $\frac{OR_I - OR_E}{OR_E - 1}$, where OR_I is calculated for the model that includes lifestyles, and OR_E is calculated for the models that includes lifestyles and other control variables.

of socio-economic variables to the overall health inequality.

The literature on income distribution and income inequality has been of help to health researchers, who have adapted income inequality measurements to fit the distribution of health (see van Doorslaer and Koolman, 2004). Socio-economic inequalities in health and pure inequality in health can be easily measured. The former can be thought of as a subset of the latter, where the population is ranked on the basis of individual income or social position and not of health. Both the concentration index (CI) of income-related inequality and the Gini coefficient of health inequality (G_H) could be used in this framework, however the particular nature of our data favours the Gini coefficient[6]. As Lerman and Yitzhaki (1989) show, the Gini can be expressed by:

$$G_H = \frac{2}{\bar{y}} cov[y, F(y)]. \tag{1}$$

where $F(y)$ is the cumulative distribution of health and $\bar{y}$ is the mean of health. The estimator of $F(y)$ in a random sample is the rank of y divided by the sample size.

2.1.1. *Decomposition of Total Health Inequality*

The Gini measure of overall health inequality *per se* does not capture the socio-economic dimension of inequality. We are interested in revealing the contri-

[6]The British HALS did not collect a continuous measure for income. Apart from bands of income, which suffer a high rate of item non-response, the only available information about the economic position of the individual is given by the social class classification. Since the indicator of social class is an ordinal categorical variable, it is not possible to generate the concentration curve measuring the degree of socio-economic inequalities, by plotting the cumulative proportion of the population ranked by social class (x-axis) against the cumulative proportion of health (y-axis). Problems arising in the generation of the fractional rank, lead to unreliable estimates of the covariance between the measure of health and the fractional rank. The routine that generates the rank is very straightforward and is available in Stata through the user-written command –glcurve7– or through the command –egen rank (), unique–. These commands require the user to sort the dataset by the ranking variable. It can be shown that if this variable does not uniquely identify each observation, but only groups of observation, then the associated fractional rank will generate a different ordering each time.

bution of each determinant of mortality to inequality. Wagstaff et al. (2001) stress the importance of "unpacking" the causes of socio-economic inequalities and use a regression-based decomposition method. Inequalities depend both on the direct impact that the various determinants of health (e.g., lifestyles, parental factors, geography, income, education, ethnicity) have on the health outcome and on the distribution of these determinants across socio-economic groups. Morris et al. (2003) decompose both the CI and the Gini coefficient for the use of health care in England. van Doorslaer and Jones (2003) decompose both indexes, using self-assessed health (SAH) and the Health Utility Index (HUI).

For the linear regression model

$$y_i = \sum_k \beta_k x_{ki} + \varepsilon_i. \tag{2}$$

the Gini can be written as an addictively decomposable form (see Wagstaff et al., 2003):

$$G_H = \sum_k \left(\frac{\beta_k \bar{x}_k}{\bar{y}_i} \right) C_k + \frac{GC_\varepsilon}{\bar{y}_i} = \sum_k \eta_k C_k + \frac{GC_\varepsilon}{\bar{y}_i}. \tag{3}$$

The first component of (3) is the explained part of inequality: it is a weighted sum of the CI of the regressors, where the weights are the elasticity, η_k, of mortality with respect to each regressor x_k. The second component is the generalized concentration index for the error term, which can be computed as a residual. The C_k are the health-related CI of each regressor. The larger η_k and C_k are, the bigger is the importance of x_k in accounting for inequality in health. If C_k is small and the elasticity is large, then regressor x_k is important to explain mortality but not to explain inequality in mortality. This kind of decomposition analysis provides the motivation of our behavioural model.

2.2. *The Behavioural Model*

In this section we present a simple behavioural model. Following Grossman (1972), our model relies on key assumptions: health is both a consumption good, demanded by the individual to enhance well-being and utility, and a fundamental commodity, produced by health-related behaviours and other inputs.

Health-related behaviours are particular individual investment choices, that explain part of the variation in the distribution of health in the population. Nevertheless, other factors that influence an individual's decisions, although known to the individual, are hidden to the researcher who does not have a complete knowledge of the decision process. These unobservable factors make the desired consumption level differ for each individual. Contoyannis and Jones (2004) proposed a model for the health production function that controls for individual heterogeneity. This provides the starting point for our analysis.

We try to disentangle health as a consumption good from health as the outcome of a production process. We assume that the individual chooses the optimal level of the demand for health given a time and budget constraint and given the trade-off with other consumption goods that enhance his utility. The individual is a rational and forward-looking economic agent. He maximizes his lifetime utility and knows the marginal productivity of investing in health-related behaviors as well as all the parameters of the decision process.

Recent work by Adda and Lechene (2001), Chang (2005) and Carbone et al. (2005) contributes to the definition of a behavioural model based on expected lifetime utility. In particular, the latter two papers focus on the demand for health investments. Carbone et al. (2005) emphasize the role of individuals' adaptation to their own health status and the existence of alternative beneficial behaviours that can be undertaken to offset the health damage of smoking.

We consider a set of six healthy behaviours among which the individual can

choose. He can decide to undertake either the full set of healthy behaviours or none, or only some of them. We use a dynamic programming approach to define the economic problem. At the initial period the individual decides the optimal behaviour that maximizes lifetime utility, where date of death is uncertain. In discrete time, the problem can be written as:

$$\max_{C} \sum_{t=0}^{\infty} \beta^t \pi_t \cdot u(\mathbf{C}_t, H_t; \mathbf{X}_U, \boldsymbol{\mu}_U). \tag{4}$$

The instantaneous utility function, $u(\mathbf{C}_t, H_t; \mathbf{X}_U, \boldsymbol{\mu}_U)$, depends on a vector of lifestyles that affect health, $\mathbf{C}_t$, and health, H_t, conditional on exogenous variables, $\mathbf{X}_U$, and a vector of unobservable factors, $\boldsymbol{\mu}_U$, that influence personal preferences. The vector of lifestyles is the choice variable and health represents the state variable. Utility is discounted by $\beta^t \pi_t$, a discount factor that includes both intertemporal preferences, β^t, and the probability of surviving until the next period π_t.

The individual's problem consists in maximizing equation (4) subject to a money budget constraint and a time constraint:

$$\sum_{l=1}^{L} p_{lt}\mathbf{C}_{lt} \leq y_t + \omega_t \, WT_t \qquad \text{where,} \quad l = 1, ..., L$$
$$\sum_{l=1}^{L} \tau_{lt}\mathbf{C}_{lt} = T - WT_t.$$

where y_t is exogenous income, ω_t is the wage rate, τ_{lt} is the amount of time necessary to consume a unit of the commmodity $\mathbf{C}_{lt}$, T is the total time available and WT_t represents the hours of labour and L is the number of lifestyles of interest. Taking $WT_t = \tau_{lt}\mathbf{C}_{lt} - T$ and combining the two constraints above,

gives the full income constraint[7]:

$$\sum_{l=1}^{L}(p_{lt} + \omega\tau_{lt})\mathbf{C}_{lt} \leq y_t + \omega T. \tag{5}$$

The problem described by equations (4) and (5) is augmented by the following transition equation describing the evolution of health:

$$H_{t+1} = f(\mathbf{C}_t; \mathbf{X}_H, \boldsymbol{\mu}_H) + (1 - \delta) \cdot H_t. \tag{6}$$

Each individual starts with a given endowment H_0 that depreciates over time at rate δ until death. Investment in health is captured by the household production function, $f(\mathbf{C}_t; \mathbf{X}_H, \boldsymbol{\mu}_H)$, which is concave and increasing in consumption of healthy lifestyles $\mathbf{C}_t$, conditional on a vector of exogenous determinants of health, $\mathbf{X}_H$, and a vector of unobservable factors influencing health $\boldsymbol{\mu}_H$. The problem can be written in terms of Bellman's equation:

$$V = \max_{C_t} \; u(\mathbf{C}_t, H_t) + E\,V(H_{t+1}). \tag{7}$$

Where the expected value of indirect utility depends on the probability of surviving, π_{t+1}, until the next period, which is assumed to be influenced directly by future health and health investment decisions:

$$\pi_{t+1} = \pi_{t+1}(\mathbf{C}_t, H_{t+1}; \mathbf{X}_M, \boldsymbol{\mu}_M). \tag{8}$$

The maximum utility achievable is given by the following value function:

$$V = \max_{C_t} \; u(\mathbf{C}_t, H_t; \mathbf{X}_U, \boldsymbol{\mu}_U) + \beta^{t+1} \cdot \pi_{t+1}(\mathbf{C}_t, H_{t+1}; \mathbf{X}_M, \boldsymbol{\mu}_M) \cdot V(H_{t+1}). \tag{9}$$

[7]Allowing for an intertemporal budget constraint, with a possibility of borrowing and saving, would be a unnecessary complication for our analysis of health and lifestyles. Our approach follows Adda and Lechene (2001).

subject to the constraint in equation (5). Future utility clearly depends on past consumption decisions: at each period utility is updated with the optimal levels of C from the utility maximization problem in the previous period.

We consider survival in each period as stochastic. We are interested in the probability of being dead by the time of the longitudinal follow-up of the HALS in May 2003, which represents the final measured health outcome for those individuals who are alive at the time of the first interview of the survey. In this framework the risk of mortality is defined as a function of the optimal level of consumption goods, investment in health-related behaviours and health at the time of the first interview.

The individual faces a trade-off between goods that maximize his direct satisfaction and behaviours that improve health. If the individual decides to improve his health, he can reduce the consumption of goods believed to be detrimental for health, and consume more goods which have beneficial effects on health. Cigarettes, alcohol and high-calorie foods produce an intrinsic pleasure, but also might have long-term negative impact on health. Individual tastes, the rate of time preference, and expectations about the probability of survival should influence the pattern of intertemporal consumption. We assume that these elements are hidden to the researcher.

Solving the maximization problem gives the demand functions for lifestyles and health, that may be defined by the following reduced form equations:

$$\mathbf{C}_l = f_l(\mathbf{X}, \boldsymbol{\mu}), \tag{10}$$

$$H = f_H(\mathbf{X}, \boldsymbol{\mu}). \tag{11}$$

where X is the vector of all observable exogenous variables in the model; and $\boldsymbol{\mu}$ includes unobservable factors which influence both the individual utility function ($\boldsymbol{\mu}_U$), the health outcome ($\boldsymbol{\mu}_H$) and the risk of mortality ($\boldsymbol{\mu}_M$).

We estimate the behavioural model by means of a recursive model for the mortality equation, the health equation and the lifestyle equations. In this framework we focus on a structural form equation for deaths to estimate the risk of mortality and a structural equation for health[8]:

$$H = h(\mathbf{C}_l, \mathbf{X}_H, \boldsymbol{\mu}_H), \tag{12}$$

$$M = \pi(\mathbf{C}_l, H, \mathbf{X}_M, \boldsymbol{\mu}_M). \tag{13}$$

We compare a model for equations (10), (11) and (13), [Model I], where health is in its reduced form, with a model where equation (11) is substituted by the structural representation of health given by equation (12), [Model II].

The indicators of mortality, health and lifestyles are observed binary variables, where health is measured by individual self-reported health. The choice to undertake some health-related behaviours depends on unobserved latent factors, including genetics, past experience, uncertainty about the future, taste differentials and rate of time preference. Also the risk of mortality and the health production function depend on unobervables. We model the empirical recursive system taking into account the unobservable heterogeneity and considering health and lifestyles as endogenous regressors in the mortality equation.

The empirical recursive system can be estimated jointly by Full Information Maximum Likelihood (FIML). This specification allows us to control for unobservable heterogeneity in the population allowing for correlation between the mortality risk, health and lifestyle choices. However, complications in the maximization of the likelihood arise on account of binary endogenous regressors, especially if there are more than three dependent variables. The error

[8]These are simply restatements of equations (6) and (8).

term in each reduced form equation depends on $\boldsymbol{\mu}$; assuming that random components in equations (10), (11) or (12), and (13) have a multivariate normal distribution, the integrals in the likelihood function have no closed-form. The MSL approach permits computation of the model and is used to estimate a multivariate probit model.

3. Data

In this paper, data from the first wave of the Health and Lifestyle Survey (HALS1) are used to measure lifestyle. They were collected between Autumn 1984 and Summer 1985, in two home visits (the second one by a nurse). The questionnaire was designed and piloted by a study team at the University of Cambridge School of Clinical Medicine and funded by the Health Promotion Research Trust. The sample design permits inferences about the British population, aged 18 and over in 1984-85.

We use a binary indicator of death as our measure of final health outcome. This allow us to measure health outcomes as recently as 2003. The accuracy of the mortality data offsets concerns about measurement errors, reporting bias or state-dependence in self-reported measures of health.

Most of the 9003 individuals interviewed in HALS1, have been *flagged* on the NHS Central Register. In May 2003 the fourth deaths revision and the first cancer revision were completed[9]. The flagging process was quite lengthy because it required several checks in order to be sure that the flagging registrations were related to the person previously interviewed.

As reported in Table 1, 97.8% of the sample has been flagged. Deaths account for some 24% of the original sample. The length of the follow-up period, nineteen years, and the checks on the official registers of deaths ensure

[9]For further information see the Working Manual for HALS available at the UK Data Archive Health and Lifestyle Survey - University of Cambridge Clinical School (2003).

the reliability of mortality data in HALS.

We propose an analysis of the relationship between death and individual characteristics, measured nineteen years before. Although the analysis covers a relatively long follow-up period, increased risk of mortality may reflect the cumulative effect of poor health (people in the poorest health status are more likely to die). Hence the probability of death is studied to explain to what extent initial conditions (measured in 1984) determine subsequent health.

4. Variables and Sample

The lifestyle variables indicate whether the individual is a non-smoker, a prudent consumer of alcohol, eats breakfast, sleeps the "optimal" numbers of hours, is not obese, and did sufficient physical activity in the last fortnight[10]:

1. Smoking is defined in terms of number of cigarettes smoked per day, using an indicator for current smokers who smoke one or more cigarettes per day.

2. Drinking is measured by a binary variable which indicates prudent alcohol consumption. The indicator is gender specific and is based on the number of drinks consumed in the past seven days before the interview[11].

3. Breakfast is an indicator of diet: we assume that eating breakfast within one hour of waking is a healthy behaviour.

[10]We use lifestyle indicators close to the categories of healthy behaviours found in epidemiological and health economic studies, such as Belloc (1973), Belloc and Breslow (1972), Kenkel (1995). These are based on the Alameda County survey carried out in California in 1965.

[11]Our indicator of alcohol consumption cannot discriminate between different styles of drinking: it does not capture differences among people who are in an abstinence period, and could be heavy drinkers, and those who are completely non-drinkers. Hence, the interpretation of the impact of this variable on mortality can be difficult if these two drinking styles are likely to have different effects. In general, the interpretation of the impact of alcohol consumption on health is not easy because there is evidence that moderate consumption gives some positive effects on health.

Table 1
Flagging Status in May 2003

Flagging Status	frequency	%
On file[a]	6506	72.26
Not NHS registered[b]	86	0.96
Deceased[c]	2171	24.11
Reported dead to HALS not on NHS Register[d]	1	0.01
Embarked - abroad[e]	43	0.48
Not yet flagged[f]	196	2.18

Notes:
[a]Currently alive and flagged on the NHS Register
[b]But not known to be dead
[c]Known dead and death certificate information recorded on file
[d]May be alive
[e]Identified on NHS Register but currently out of country
[f]Not currently flagged for various reasons (no name etc.)

4. Sleep patterns have been recognized as potential determinant of health status. Hence, we created an indicator which splits the sample into two groups according to their sleep pattern[12].

5. The indicator for obesity is measured using the Body Mass Index (BMI) reported at the nurse visit and it is gender specific [13].

6. Since sporting activities are known to be healthy and to help people in stress or depression, we also use an indicator of physical activity, measured in the last fortnight. It is created for each individual by summing the time involved in each of fourteen types of exercise.

Health is measured by an indicator of self-assessed health. Respondents to the HALS in 1984 were asked to rank they health status with respect to people of their own age. We have transformed the categorical indicator of SAH in a binary variable that takes value one if individual perceived health is excellent

[12]We define the optimal sleeping level as sleeping between seven and nine hours per night. Inadequate sleeping may reflect physical and psychological problems and may be an indicator of stress, possibly attributable to effects of lack of social support, autonomy or control. This kind of stress has been identified as an important source of the socio-economic gradient in health. (see, e.g., Marmot et al., 1984, 1991, 1997; Marmot, 2004)

[13]Height, which is a component of the BMI, is also included as a continuous exogenous variable in the econometric model, because it is known to be a good predictor of mortality and morbidity risks and captures heterogeneity in initial endowments.

or good, and zero if it is fair or poor.

The other variables describing socio-economic characteristics, geographical position, marital status, housing are reported in Table A.2 in the appendix.

Our purpose is to investigate premature mortality in the adult population: we use a sample limited to 3655 individuals aged 40 and over, from the original sample of 9003 observations[14]. All observations containing missing values have been dropped. Both descriptive statistics and econometric analysis are conditioned upon this restriction.

5. Descriptive Analysis

5.1. *Lifestyle and Socio-Economic Status*

A simple descriptive analysis is reported in Table A.2 in the appendix, which presents sample means for the most relevant variables that describe the sample. 70% of the individuals interviewed in 1984 declared to have a good or excellent health status relative to people of their own age. The sample comprises 46% men and 54% women and is made up of individuals whose behaviours are mostly healthy. 88 and 85% of the sample is prudent in the consumption of alcohol and is not obese. Only 30% of individuals are smokers, while 32% of them devote time to physical activities, 71% usually eat breakfast and 58% sleep a healthy number of hours.

The social class classification is derived from the Registar General's Social Class (RGSC), based on occupation. For our purposes, the most convenient way to aggregate individuals in the sample is to collapse social classes in three macro groups: a top class (sc1) including students, professional, managerial and intermediate workers, a middle class (sc2) including skilled workers and

[14]Persons who are younger than 40 are more likely to change their educational qualification over time, but the cross-sectional nature of the analysis does not allow to control for this change. The interpretation of the effect of education on mortality risk could be biased.

armed services, and a bottom class (sc3) including partly skilled and unskilled. Individuals are largely concentrated in sc2; the extreme classes are smaller. Around 61% of the sample does not have formal educational qualifications, and only 13% has a university degree.

Table 2 shows the mean values of some variables of interest in our analysis. The statistics are reported for sub-samples, according to the intensity of healthy behaviours. Since splitting the sample in different groups on the basis of every combination of lifestyle choices would have been prohibitive, the table only presents the characteristics of three sub-groups[15]. Although sixty-four (2^6) possible combinations of health-related behaviours can be analyzed, we are only interested in the probability of having a totally healthy lifestyle $Pr(6)$, the probability of following a totally unhealthy lifestyle, one or two healthy behaviours $Pr(0/1/2)$, and the probability of three, four, or five healthy behaviours $Pr(3/4/5)$[16].

We find $Pr(0/1/2) = 0.064$, $Pr(3/4/5) = 0.867$ and $Pr(6) = 0.069$. These probabilities, multiplied by the full sample size, give the expected frequency of each group, reported between brackets in the table. The expected frequencies for the most unhealthy and the most healthy groups are underestimated. The expected frequency for the group of persons who follow three, four or five healthy practices is estimated to be 1.08 times bigger than observed frequency.

The subdivision of the sample shows that the number of deaths decreases as we move from the more unhealthy group to the healthiest lifestyle. Social position, work environment and education can have a strong role in the deter-

[15]We do not report in the table the case in which individuals have a totally unhealthy lifestyle because the size of this group is too small to make inferences.

[16]These probabilities are computed using the following formulas: $Pr(6) = \prod_{i=1}^{6} P(Y_i = 1) = \{1,1,1,1,1,1\}$, $Pr(0/1/2) = \prod_{i=1}^{6} P(Y_i = 0) + \sum_{i=1}^{6} P(Y_i = 1) \prod_{k \neq i} P(Y_k = 0) + \sum_{i=1}^{6} \sum_{k>i} P(Y_i = 1) P(Y_k = 1) \prod_{j \neq k \neq i} P(Y_j = 0)$, $Pr(3/4/5) = 1 - Pr(0) - Pr(1/2) - Pr(6)$, where $Pr(0) = \prod_{i=1}^{6} P(Y_i = 0) = \{0,0,0,0,0,0\}$. Since $Pr(0)$ is very small (0.0005), people having either no healthy behaviour, or at least one or two, compose the most unhealthy group.

Table 2
Variable means by sub-samples defined by number of a priori *"healthy behaviours"*

variable	full sample N=3655	0/1/2 N=372 (*Exp*=233.92)	3/4/5 N=2932 (*Exp*=3168.88)	6 N=351 (*Exp*=252.20)
Health				
death	0.359	0.401	0.374	0.191
sah	0.703	0.605	0.696	0.863
Social Class				
sc1	0.316	0.215	0.306	0.500
sc2	0.467	0.497	0.473	0.380
sc3	0.218	0.290	0.220	0.123
Education Level				
degree	0.125	0.078	0.118	0.237
hvqA	0.125	0.094	0.124	0.165
O-cse	0.094	0.059	0.093	0.143
no edu.	0.608	0.716	0.619	0.402
other edu.	0.047	0.054	0.046	0.054
Occupational Status				
full time - student	0.364	0.462	0.343	0.436
part time	0.132	0.116	0.124	0.217
unemployed	0.03	0.054	0.03	0.009
sick	0.033	0.054	0.033	0.011
retired	0.339	0.191	0.372	0.219
housekeeper	0.102	0.124	0.098	0.108
shift worker	0.057	0.102	0.055	0.031
Gender and Age				
male	0.455	0.505	0.451	0.442
age	57.468	53.889	58.412	53.380

mination of one's health status and, consequently, of the odds of dying. The healthier individual behaviours are, the bigger the proportion of persons belonging to the higher social classes. The number of individuals in the bottom classes decreases moving from the most unhealthy lifestyles to the healthiest. A strong association of schooling with health-related choices is shown by the table: people have more healthy behaviours if they are more educated. Individuals with no educational qualifications have more unhealthy lifestyles. The role of schooling on health has been emphasized in Grossman (1972) and investigated in Kenkel's empirical work (1991).

This evidence leads to the same conclusion as Contoyannis and Jones's (2004) study of SAH : health-related behaviours are not randomly distributed but cluster together in certain categories of individuals, and the relationship

with social class and education must be taken into account. However we still need evidence about the extent to which these factors influence the health outcome. We do not know if their impact on health is subject to change depending on individual propensities to behave in a healthy way.

5.2. *Deaths and Socio-Economic Status*

A crude way to see if mortality varies with the characteristics of the population is to use the simple death rate[17]. Table 3 records death rates for some variables of interest. It is not surprising that the highest social classes have lower mortality rates. The death rate increases from the highest social class (sc1) to the lowest classes and it is almost double in the lowest two classes with respect to the highest. About 43 individuals die for every 100 persons in the population in the sc3 class. Only 27 individuals, for every 100, die in the highest social class. Premature mortality (death before age 65), is higher among people who are unskilled.

Table 3 also shows the association between mortality and education. The role that a person can play in the labour market depends on their educational qualifications. The attained level of education is, in such a way, related to socio-economic status, income, housing, health status and health inequalities. Kenkel (1991) finds that better educated persons are also more likely to have a good knowledge of what a person should do to be more healthy. The death rate is higher for people who have no qualifications and is about twice the death rate of those who obtained a University degree or an O level.

[17]The death rate is calculated in each class as the percentage of deaths in that class. The number of deaths in a certain category, times 100, divided by the total number of individuals in that category.

Table 3
Percentage death rate in different socio-economic groups

variable	death rate %
sc1	26.89
sc2	38.86
sc3	42.71
degree	24.73
hvqA	22.59
O-cse	22.32
no edu.	43.10
male	42.91
female	30.09
sah	31.08

6. Econometric Methods and Results

This section illustrates our estimation strategy. We describe the econometric approach that we adopt to obtain consistent estimates of the causal effect of the health-related behaviours on health. This method allow us to control for unobservable heterogeneity across the population and endogeneity of the behaviours, which is reckoned a common methodological problem in the empirical literature. We show that the inclusion of lifestyles, even under a very restrictive assumption of exogeneity, is crucial in explaining socio-economic inequalities in mortality. Alternative models for mortality are presented.

6.1. *A Multivariate Probit Model for Mortality*

Efficient and consistent estimation of the parameters in the health production function requires a model that takes account of the nature of the variables used. The multivariate probit model is appropriate because it considers unobservable heterogeneity. Our model consists of a recursive system of equations for lifestyles, morbidity and mortality. Its most important feature is that the random components of the lifestyle equations are allowed to be freely correlated with the random component of the mortality equation. If there are unobservable individual characteristics, influencing both individual's healthy

behaviours and their probability of death, the model is able to take them into account.

Endogeneity can arise with the inclusion of lifestyles as regressors in the mortality equation, due to potential correlation between the error terms in the lifestyle equations and the error term in the mortality equation. If endogeneity is proven to exist, then estimates from the univariate probit version of the mortality equation will be consistent[18].

The multivariate probit model estimates a set of probabilities depending on whether the i-th individual is dead or alive and has a more or less healthy lifestyle, according to the definition of healthy behaviours that we are using. There are 256 (2^8) combinations of successes and failures in our model, because we have eight equations and each response has got two possible outcomes.

We have a recursive system, which consists of structural equations for the health production functions and six reduced-form equations for lifestyles. In the main equation, the mortality equation, the dependent variable y_{id} is equal to one if the individual had died by May 2003 and zero if the individual was still alive. In the other equations the dependent variables, y_{ih} and y_{il}, take value one if perceived health is excellent or good and if a particular lifestyle is "healthy" (individuals do not smoke, do sleep well, and so on), and zero otherwise.

[18]We would like also to control for another issue, associated with unobservable heterogeneity: the potential measurement error in the indicators of lifestyle. One of the limitations of some epidemiological studies is that they fail not only to consider the problem of unobservable heterogeneity but also the problem of measurement errors. Lynch et al. (1996), and Lynch et al. (1997), shed light on the possibility of measurement error in risk factors (behavioural, biological, psychological, social risks) used in the analysis. To obtain consistent and efficient estimates, Wooldridge (2002, p. 470-478) suggests maximum likelihood estimation when, in a system of two equations, an explanatory variable indicating participation (for example, a binary variable for smoking or drinking) is measured with error. After normalization of the error term in the equation with measurement problems, the ML procedure is used to calculate the average partial effect of the mis-measured variable on the response. Then it should be straightforward to test if there is measurement error by a simple asymptotic t-test, or LR-test, on the null hypothesis that the correlation between the errors of each lifestyle equation and the error of the mortality equation is zero. However, in this context, this test has a limitation due to the difficulty in discriminating between mis-measurement and unobservable heterogeneity in the data.

The latent variables underlying each observed variable define the following equations:

$$\begin{aligned} y^*_{il} &= \alpha'_l \mathbf{W}_{im} + \beta'_l \mathbf{Z}_{is} + \gamma'_l \mathbf{I}_{ig} + \varepsilon_{il}, \qquad l = 1, \ldots, 6, \;\; m = 1, \ldots, M \\ y^*_{ih} &= \delta'_h \mathbf{Y}^L_{il} + \alpha'_h \mathbf{W}_{im} + \beta'_h \mathbf{Z}_{is} + \varepsilon_{ih}, \qquad s = 1, \ldots, S, \;\; g = 1, \ldots, G \\ y^*_{id} &= \delta'_d \mathbf{Y}^L_{il} + \vartheta'_d y^H_{ih} + \alpha'_d \mathbf{W}_{im} + \varepsilon_{id} \qquad i = 1, \ldots, n. \end{aligned} \tag{14}$$

where $\mathbf{Y}^L_{il} = \{y_{i1}, y_{i2}, y_{i3}, y_{i4}, y_{i5}, y_{i6}\}$ is a vector of six lifestyles. Each lifestyle and self-assessed health are observed by the researcher as:

$$y_{il,h} = \begin{cases} 1 & \text{if } y^*_{il,h} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Using the matrix notation our recursive system of eight equations can be simplified as follows:

$$\mathbf{Y}_{[(l+2)\times 1]} = \begin{pmatrix} \mathbf{\Gamma}_{[(l+2)\times g]} \\ \mathbf{B}_{[(l+2)\times s]} \\ \mathbf{A}_{[(l+2)\times m]} \\ \mathbf{\Delta}_{[(l+2)\times l]} \\ \mathbf{\Upsilon}_{[(l+2)\times 1]} \end{pmatrix}' \begin{pmatrix} \mathbf{I}_{(g\times 1)} \\ \mathbf{Z}_{(s\times 1)} \\ \mathbf{W}_{(m\times 1)} \\ y^H_{(1\times 1)} \end{pmatrix} + \mathbf{E}_{[(l+2)\times 1]}.$$

If $\mathbf{Y}_i$ is the random vector of the responses $\{\mathbf{Y}_l, \; \mathbf{Y}_d\}$, then the probability of observing a certain combination of responses on all seven variables, conditioned on parameters $\mathbf{\Theta} = \{\mathbf{\Gamma}, \mathbf{B}, \mathbf{A}, \mathbf{\Delta}, \mathbf{\Upsilon}\}$ and $\mathbf{\Omega}$ is given by[19]:

$$P(\, \boldsymbol{Y}_i = y_i \mid \mathbf{\Theta}, \mathbf{\Omega}) = \Phi_8(x_{i1d}, \ldots, x_{ilhd}). \tag{15}$$

[19]In this formulation $K_{il,h} = 2y_{il,h} - 1$, for each i, l = 1, ..., 6, and $d_{id} = 2y_{id} - 1$.

where Φ_8 is the 8-dimensional multivariate standard normal distribution, $x_{ilhd} = d_{id}\mathbf{K}_{il,h}\mathbf{\Theta}'\mathbf{X}$, $\mathbf{X} = \{\mathbf{I}, \mathbf{Z}, \mathbf{W}, \mathbf{Y}^L, \mathbf{Y}^H\}$, and the matrix Ω has values of 1 on the leading diagonal and correlations between the error terms of the seven equations as off-diagonal elements. The errors terms of the latent equations have a multivariate normal distribution: $\varepsilon_i \sim MVN(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \{\rho_{jk}\}$ is the correlation matrix for the j-dimensional multivariate normal, obtained considering the Choleski decomposition of the covariance matrix for the errors: $\mathbf{\Sigma} = \mathbf{C}ee'\mathbf{C}'$, where e are independent standard normal random variables.

Chib and Greenberg (1997) suggest the use of the correlation matrix *for identifiability reasons*. The variances of the epsilon must be equal to one and the off-diagonal elements are symmetric. The parameters and the elements of $\mathbf{\Omega}$ are not likelihood identified together, whereas the $J(J\text{-}1)/2$ parameters of $\mathbf{\Sigma}$ can be identified.

Conventionally identification of recursive multivariate probit models has been based on exclusion restrictions. According to Schmidt (1981), simultaneous probit models suffer from identification problems. Imposing restrictions allows us to estimate a unique outcome of the latent y^*_{id}, for any value of the regressors and for any error term. Exclusion restrictions are imposed in order to estimate all parameters. In this case, given the triangular matrix of coefficients in the recursive model, the $\mathbf{W}$ matrix in the mortality equation is just a portion of the matrix of regressors in the lifestyles and health equations.

Our model is a simple recursive model of the type model 6 in Maddala (1983, p. 117-138)[20]. If the random components of the latent equations are not independent and the matrix of regressors in the primary equation (mortality equation) includes all the regressors of the secondary equations (self-assessed health and lifestyle equations), then Maddala argues that the parameters in the equation of y^*_{id} are not identified. The identification restriction is that

[20]The responses (y_{id}, y^H_{ih} and $\mathbf{Y}_{il}$) are observed as dichotomous variables, and y^*_{id} depends on y^H_{ih} and $\mathbf{Y}_{il}$.

at least one variable in the secondary equations is not included in the deaths equations. Following this approach, we would need to exclude the regressors $\mathbf{Z}_{is}$ and $\mathbf{I}_{ig}$ from the mortality equation because they only indirectly influence the risk of dying, and the regressors $\mathbf{I}_{ig}$ from the health equation because they have an indirect impact on health. However, more recent research by Wilde (2000) shows that Maddala's approach to identification of multiple equation probit models is valid for a simple constant-only model. Wilde shows that, given the full rank of the regressor matrix, it is only necessary to have varying exogenous regressors to avoid identification problems and exclusion restrictions are not required. In our empirical analysis we carry out a sensitivity analysis, comparing a model with no exclusion restrictions to a range of models with different identifying assumptions.

Evaluating the likelihood function, raises computational problems due to the fact that unobservable factors are jointly normal distributed. The log-likelihood function for our model has the form:

$$L = \sum_{i} \log \Phi_8(x_{i1d}, \ldots, x_{ilhd}). \tag{16}$$

Problems arise due to the numerical computation of multidimensional integrals. Here, the multivariate probit model is estimated in Stata using a GHK (Geweke-Hajivassilou-Keane) simulator for probabilities and a MSL procedure.

The GHK simulator exploits the Choleski decomposition of the covariance matrix, so that the joint probability originally based on unobservables can be written as the product of univariate conditional probabilities where the epsilon's are substituted by error terms, $u_i \sim \Phi_8(0, I_8)$, independent from each other by construction.

Although this simulation technique presents several advantages, the evaluation of the log-likelihood also requires another important stratagem to reduce

simulation bias. The simulated ML procedure using GHK at each iteration is numerically intensive. Indeed, even though $\hat{P}_n(\theta)$ is unbiased for $P_n(\theta)$, $ln\ \hat{P}_n(\theta)$ is not unbiased for $ln\ P_n(\theta)$. Only if the number of draws R grows at a rate that is faster than $\sqrt{N}$, the MSL estimator, which maximises the score function after plugging in the simulated probability, $\vartheta ln\check{P}_n(\theta)/\vartheta(\theta)$, is asymptotically consistent and efficient.

The MSL approach used in the Multivariate Probit Model is preferable to the alternative procedure of separate ML estimation of the univariate probit models for the mortality equation and the lifestyle equations. The latter does not account for the correlation between the error terms, but rests on the assumption of exogeneity of the lifestyle covariates. Maddala (1983, p. 123) finds that if the error terms are not independent, the probit ML approach gives inconsistent estimates of the parameters. Maddala (1983) and Knapp and Seaks (1998) show that the log-likelihood function to be maximized in the multivariate probit model is equal to the sum of the log-likelihood functions obtained by the separate ML probit models when the restriction of independence of the errors is true. They propose two alternatives to the Hausman test for exogeneity of a dummy variable: a z test and a LR test[21]. The latter considers the likelihoods for the separate ML probit models as identical to equation (16) if the restriction $\rho_{jk} = 0$ holds. However, the LR test is not easy to calculate for each null of exogeneity because of the high number of marginal probabilities in the log-likelihood function. We use the statistic $z = \frac{\hat{\rho}}{S.E.(\hat{\rho})}$ for testing $H_o : \rho_{jk} = 0$. If the errors are independent, the MSL estimation is equivalent to the separate ML probit estimations. Hence, it is sufficient to test the unique restriction $\rho_{jk} = 0$, using the asymptotic standard errors provided by the MSL estimation.

[21] An application can be found in Brown et al.'s paper (2005), which, analyzing the propensity to employment, considers the potential endogeneity of a binary variable indicating diabetes.

6.2. *Probit Models*

Following Wilde's (2000) result on identification of multiple equation probit models, we tried to estimate a multivariate model where each equation has the same regressor matrix. However, probably due to either the high number of endogenous regressors or the nature of our data, the MSL of the multivariate probit does not converge to a global maximum. Therefore, relying on Schmidt (1981) and Maddala's (1983) approach, we decided to change the specification of the model by setting some exclusion restrictions.

We compare four different sets of exclusion restrictions and use information criteria to balance statistical fit of the model and sufficient parsimony in the parametrization. We exclude piecewise, from the mortality equation, those variables that are assumed to influence directly individual's decisions about health-related behaviours and perceived health, but do not have a direct effect on longevity.

The best set of exclusion restrictions has been chosen by comparing the fit of the models. For a more robust specification we compare a model that includes reduced form equations for lifestyles and morbidity represented by equations (10) (11) and (13) [model I], to a model with a structural equation for health, represented by equations (10) (12) and (13) [model II]. Changes in the exclusion restrictions lead only to marginal variation in the coefficients, in both models I and II. A direct comparison of the two models suggests that, as it should be, the estimates are stable in spite of the use of a reduced form equation for the health equation. Statistical criteria suggest a quadratic polynomial of age and support the exclusion of parental variables, which are supposed to influence directly lifestyle habits rather then longevity (they indicate if parents used to smoke or drink regularly), geographical variables, marital status, house ownership and size of household, which should, analogously, have a direct effect

on SAH[22].

In particular, we believe that differences in morbidity between people living in different geographical regions of the UK could be explained by different access to health care services and regional variations in the quality of health care sector. Environmental pollution is also likely to vary by regions and to influence morbidity. We reckon that marital status has a direct influence on health and health-related behaviours rather than on mortality risk. In Kenkel (1995), marital status is excluded from the health equation because it is considered a determinant in the input demand equation. Married people show lower death rates because they exhibit a better health status, they have more positive health attitudes, are more likely to be wealthier and are linked to tighter-knit social support networks[23]. House ownership and number of individuals in the households are considered to have only a direct effect on lifestyle and morbidity but not on mortality risk.

The Likelihood Ratio (LR), the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), support this model specification[24]. We use a second order polynomial to define the probability of death as a smooth and flexible function of age, since variations in the other covariates can be attributed to age. Indeed, we need to separate variations across different ages and variations of other variables on the risk of mortality. Information criteria suggest that a cubic representation would not improve the fit of the model. We also performed a RESET test which suggests that the mortality equation

[22]The preferred specification of the model corresponds to the set of exclusion restriction called IV4. The variables excluded from the mortality equation are: mother smoked, father smoked, both smoked, others smoked, father drinker, mother drinker, wales, north, nwest, yorks, wmids, emids, anglia, swest, london and scotland, widow, divorced, separated, single, house owner, household size. Tables with full results are available from the authors upon request.

[23]See chapter 10 in the Handbook of Population, Poston and Micklin (2005).

[24]The AIC is calculated as $(-2logL+2p)$ and the BIC is calculated, following the Schwarz formula, as $(-2logL+log(N)p)$ where p is the number of the parameters and N is the sample size.

with these exclusion restrictions is not misspecified [25].

To refine the specification of the recursive system, we analyze the matrix of correlation coefficients from the multivariate probit. Table 4 shows that non smoker, breakfast and sleep well, which are statistically significant explanatory variables of mortality in the multivariate probit, are also highly correlated to the risk of death. We propose a reduced recursive system for mortality where non smoking, breakfast behaviour and sleeping patterns are endogenous and the remaining lifestyles are exogenous.

The reduced multivariate probit model has been estimated both following Wilde (2000) and setting exclusion restrictions. Table 5 reports the coefficient estimates of the variables of interest obtained from five different specifications and from both univariate and multivariate probit models. The last column of Table 5 shows the estimates from the model with no exclusion restrictions. The impact of eating breakfast on the risk of mortality becomes positive in the multivariate probit, but that variable is no longer statistically significant. Sleeping patterns become statistically significant and their impact on mortality is much higher in the multivariate probit, as well as the impact of non smoking and self-assessed health. The impact of the socio-economic variables is lower than in the exogenous model. The univariate probit is likely to pick up the effect of the unobervables. The other columns of the table compare four different sets of exclusion restrictions. We find only marginal variation in the coefficients moving from one specification to another. Table 6 reports the statistical fit of these specifications. The LR test, and the information criteria (AIC and BIC) favour the IV4 specification, that is the same exclusion restriction set preferred in the general model where all lifestyles are endogenous.

[25] Wooldridge (2002, p.124-125), advises not to use the RESET to test the presence of omitted variables, because it can only test whether or not the expected value of the dependent variable conditional on the set of regressors is linear in the regressors. The χ^2 test statistic, computed for the mortality equation that includes exogenous sah and lifestyles, is equal to 1.23, with p-value above the conventional significance level ($p = 0.2671$).

Table 4

Correlation coefficients from the multivariate probit - general model

equation	deaths	sah	non smoker	breakfast	sleep well	prudent drinker	non obese	exercise
death	1.000							
sah	0.278	1.000						
non smoker	0.334^{**}	0.112	1.000					
breakfast	-0.281^{*}	-0.165	0.270^{**}	1.000				
sleep well	0.381^{**}	0.059	0.031	0.124^{**}	1.000			
prudent drinker	0.040	0.000	0.183^{**}	0.268^{**}	0.049	1.000		
non obese	$-0.223^{\dagger}$	-0.354^{*}	-0.227^{**}	0.077^{*}	0.040	-0.045	1.000	
exercise	-0.089	-0.014	0.095^{**}	0.073^{*}	-0.017	-0.078^{*}	0.127^{**}	1.000

Notes: Based on the specification IV4 of the general model where all the lifestyles regressors are endogeneous.
Significance levels: $\dagger$: 10% $*$: 5% $**$: 1%

Table 5

Selected coefficient estimates for the mortality equation in the reduced model, under exogeneity and endogeneity

	IV1		IV2		IV3		IV4		no IV	
	Exog[§]	MVP[§§]	Exog	MVP	Exog	MVP	Exog	MVP	Exog	MVP
sah	-0.300^{**}	$-0.614^{\dagger}$	-0.300^{**}	$-0.581^{\dagger}$	-0.302^{**}	$-0.602^{\dagger}$	-0.301^{**}	$-0.555^{\dagger}$	-0.298^{**}	$-0.625^{\dagger}$
non smoker	-0.354^{**}	-0.859^{**}	-0.354^{**}	-0.875^{**}	-0.354^{**}	-0.850^{**}	-0.354^{**}	-0.868^{**}	-0.321^{**}	-0.864^{**}
eat breakfast	-0.155^{**}	$0.409^{\dagger}$	-0.156^{**}	$0.401^{\dagger}$	-0.155^{**}	0.415^{*}	-0.155^{**}	0.415^{*}	-0.167^{**}	0.310
sleep well	-0.077	-0.734^{**}	-0.077	-0.681^{**}	-0.080	-0.750^{**}	-0.079	-0.681^{**}	-0.075	-0.717^{*}
prudent drinker	-0.116	-0.098	-0.113	-0.096	-0.116	-0.099	-0.116	-0.098	-0.088	-0.087
non obese	-0.173^{*}	$-0.121^{\dagger}$	-0.174^{*}	$-0.127^{\dagger}$	-0.177^{*}	$-0.122^{\dagger}$	-0.177^{*}	$-0.128^{\dagger}$	-0.166^{*}	$-0.125^{\dagger}$
exercise	-0.089	-0.050	-0.090	0.060	-0.090	-0.053	-0.090	0.060	-0.091	-0.051
sc1	$-0.124^{\dagger}$	-0.061	$-0.124^{\dagger}$	-0.065	$-0.129^{\dagger}$	-0.064	$-0.129^{\dagger}$	-0.069	$-0.116^{\dagger}$	-0.058
sc3	-0.018	-0.042	-0.018	-0.038	-0.017	-0.041	-0.016	-0.036	-0.024	-0.043
degree	0.028	0.046	0.029	0.045	0.035	0.047	0.035	0.046	0.056	0.051
hvqA	-0.085	-0.056	-0.084	-0.057	-0.083	-0.055	-0.083	-0.056	-0.062	-0.055
no edu.	0.087	0.057	0.087	0.060	0.090	0.058	0.090	0.063	0.099	0.050
other edu.	-0.089	-0.090	-0.087	-0.090	-0.082	-0.084	-0.082	-0.084	-0.076	-0.101
logL (Exog)	-1574.487		-1574.603		-1575.440		-1575.447		-1565.019	
logL (MVP)		-9924.136		-9924.7268		-9924.5268		-9925.1967		-9914.791

Notes: [§]Exog: single equation probit with exogeous lifestyles; [§§]MVP: multivariate probit

The sets of exclusion restrictions are as follows. IV1 = parental and geographical; IV2 = IV1 + household and tenure; IV3 = IV1 + marital status; IV4 = IV2 + marital status; no IV = no exclusion restrictions.

Significance levels: $\dagger$: 10% $*$: 5% $**$: 1%

Table 6
Statistics for the reduced model assuming endogeneity

	IV1	IV2	IV3	IV4	no IV
AIC (Akaike)	20260.272	20257.454	20253.054	20250.393	20285.582
BIC (Schwarz)	21538.265	21523.039	21506.232	21491.166	21700.060
		IV2 - IV1	IV3 - IV1	IV4 - IV1	IV4 - no IV
LR test:		$\chi^2_2 = 1.182$	$\chi^2_4 = 0.782$	$\chi^2_6 = 2.122$	$\chi^2_{28} = 20.812$

The main results of the model are reported in Table 7, which compares the average partial effects in the multivariate probit and the univariate probits[26]. The reference individual in the mortality equation is female, a full-time skilled worker, with O level educational qualifications or equivalent, white European, and lives in an inner city.

In the exogenous model all the lifestyle indicators have a negative sign. But only non smoking, eating breakfast and obesity are statistically significant, with a higher negative impact of non smoking on mortality. Healthy behaviours reduce the probability of death. The model predicts a higher probability of death for men, white Europeans, unemployed, people absent from work because of illness and housekeepers. People in the top social class and people that do not live in the city are less likely to die with respect to the reference individual.

Looking at the average partial effects of the socio-economic variables on the probability of dying for alternative models of mortality, we find a quite clear social class gradient in mortality: the probability of dying decreases, as expected, moving from the top class to the bottom class. Unsurprisingly, the models that assume exogenous lifestyles predict a lower impact of social classes

[26]The MSL estimates of the coefficients and their statistical significance for our system of equations are computed by the software Stata, using the command *mvprobit* created by Cappellari and Jenkins. For more details about the algorithm see Cappellari and Jenkins (2003). Sample standard deviations, that measure variation across individuals in the partial effects, are reported along with the average partial effects in Table 7. Table A.1 in the appendix reports the full results from the multivariate probit model. Coefficient estimates from the single equation probit are available from the authors upon request.

on mortality than in the model that excludes lifestyles. It is more difficult to find the same clear gradient in the distribution of the partial effects of educational qualifications. The partial effects of the other regressors tend to be stable or to become slightly smaller when lifestyles are included in the model, even assuming exogeneity. This suggests that the researcher should consider the close connection between health-related behaviours and mortality: variations in the the socio-economic gradient might be associated to differences in lifestyle. Indeed, the important role of unobservable factors, such as individual preference differentials, in the decision process of how healthy their life should be, has to be considered.

In the multivariate probit the probability of death decreases for those who have excellent or good health, do not smoke, sleep well, are not obese and are shift workers; it also increases for men, white Europeans, the housekeepers and the unemployed. The risk of mortality increases with age. With respect to the single equation probit social class, absence from work due to sickness and type of area are no longer statistically significant in the multivariate probit.

We are interested in the average partial effects of lifestyles on mortality. The partial effects of non smoking, sleeping patterns and eating breakfast, are bigger than in the univariate probit. In particular, non smoking and sleeping patterns have the highest impact on mortality risk with respect to the other lifestyles. Non obesity is still statistically significant but with a lower impact on mortality in the multivariate probit. The multivariate probit also suggests that non obese and fit people are more likely to have good or excellent health. However, once endogeneity is controlled for, breakfast has a positive impact on the probability of death[27]. According to previous evidence in the empirical

[27]It is worth noting that breakfast has a positive sign in the SAH equation. In particular, individuals who eat breakfast are predicted to be more likely to die early, while in the SAH equation eating breakfast is not statistically significant but seems to predict a higher probability to be in excellent or good health. A counter-intuitive sign of the partial effects of the nutrition variables in the health equation has been found in other research (see Kenkel, 1995; Contoyannis and Jones, 2004). This might suggest that breakfast is picking up the

Table 7
Average partial effects (APE) *in alternative models for mortality*

variable	Excl[§]		Exog		MVP[§§]	
	APE	S.D.	APE	S.D.	APE	S.D.
sah			−0.075	0.036	−0.143	0.064
non smoker			−0.087	0.042	−0.220	0.096
breakfast			−0.038	0.019	0.101	0.052
sleep well			−0.019	0.010	−0.175	0.077
prudent drinker			−0.029	0.014	−0.024	0.012
non obese			−0.043	0.021	−0.032	0.016
exercise			−0.022	0.011	−0.015	0.007
sc1	−0.045	0.021	−0.031	0.016	−0.017	0.009
sc3	0.006	0.003	−0.004	0.002	−0.009	0.005
degree	0.001	0.001	0.008	0.004	0.011	0.006
hvqA	−0.020	0.010	−0.020	0.010	−0.013	0.007
no edu.	0.040	0.018	0.022	0.011	0.016	0.008
other edu.	−0.005	0.002	−0.020	0.010	−0.020	0.011
part time	0.033	0.015	0.036	0.018	0.040	0.020
unemployed	0.102	0.043	0.072	0.034	0.065	0.032
sick	0.227	0.085	0.156	0.068	0.072	0.034
retired	0.023	0.010	0.016	0.008	−0.004	0.002
housekeeper	0.077	0.034	0.061	0.030	0.060	0.030
shift worker	−0.047	0.023	−0.062	0.033	−0.072	0.038
rural	−0.053	0.025	−0.038	0.020	−0.025	0.013
suburb	−0.026	0.012	−0.018	0.009	−0.020	0.010
ethwheur	0.078	0.040	0.092	0.051	0.082	0.044
height	0.002	0.001	0.002	0.001	0.003	0.001
male	0.111	0.049	0.107	0.052	0.096	0.047
age	0.009	0.004	0.009	0.004	0.001	0.001
age2	0.009	0.004	0.010	0.005	0.014	0.007
N	3655		3655		3655	
Log-likelihood	-1628.1314		-1575.4469		-14535.445	
χ^2	$\chi^2_{(19)}$ =1516.99		$\chi^2_{(26)}$ =1622.36		$\chi^2_{(313)}$ = 3813.39	

Notes: [§]Excl: single equation probit without lifestyle variables as regressors; [§§]Multivariate probit for the reduced model, IV4.

literature, prudent drinking has a negative influence on the risk of mortality: moderate alcohol consumption may have a positive effect on health.

Table 8 shows that the null of exogeneity is highly rejected only for non smoking and sleeping well. Unsurprisingly, the correlation between the mortality equations and the error terms of the non smoker and sleep well equations is positive, meaning that unobserved factors that increase the probability of being a non smoker and sleeping well, also make someone more likely to die earlier. Smoking and bad sleeping habits might be associate to depression and

effect of unobservables on mortality risk.

psychological distress that arise from the inability of the individual to control over his life and unsatisfactory social support. Recent research shows that people who are supported and participate in social networks are more likely to have better health (see Marmot, 2004). In the case of breakfast, the correlation coefficient has, as expected, a negative sign. This is consistent with the lower estimates in the univariate probit. The exogeneity assumption generates downward biased estimates of the causal effects of the behaviours. Accounting for endogeneity of the regressors, we see that there is a statistically significant effect of unobserved factors both on the mortality risk and the probability of some behaviours.

Employing the concept of frailty to interpret the correlation coefficients, we could say that, on one hand, frailer individuals tend to select into, for example, non-smoking and that, on the other hand, those who have a bad health history are more likely to die prematurely. Frailer people seem to be more prone to adopt healthier lifestyles (ie., they do not smoke or they quit smoking) than persons with a better health endowment[28]. Then, the unobserved effect, captured by the multivariate probit, would imply that non-smokers are also more likely to die early. The univariate probit model incorporates this positive effect as a bias of the causal effect of the healthy lifestyle, reducing its negative influence on mortality.

[28] Adda and Lechene (2004), tried to overcome the weaknesses of the medical, epidemiological and economic literature dealing with the health production function and the problem of individual heterogeneity and endogeneity, using a tobacco-free morbidity score in a duration model on data from the Swedish Survey on living condition (ULF). Assuming an *a priori* correlation between smoking and mortality, they found, according to previous evidence (see Adda and Lechene, 2001), that people with poor underlying health, and consequently lower life expectancy, do select into smoking. Contrarily, our model suggests that people in poor health do not select into smoking and that non-smokers are less likely to die even if they are frailer.

Table 8
Correlation coefficients from the multivariate probit - reduced model IV4

equation	deaths	sah	non smoker	breakfast	sleep well
death	1.000				
sah	0.239	1.000			
non smoker	0.321**	0.079	1.000		
breakfast	−0.239†	−0.167	0.271**	1.000	
sleep well	0.374*	0.038	0.031	0.122	1.000

Significance levels: † : 10% * : 5% ** : 1%

7. Results from a Decomposition Analysis of Total Health Inequality

In this section we compute the Gini coefficient to give a robust measure of health inequality using equation (1) presented in section 3.

Since our mortality indicator is a binary variable, we use predicted mortality, that is the linear index for death predicted from probit models of the mortality equation presented in section 6, to analyze total health inequality. van Doorslaer and Jones (2003), dealing with an ordered categorical dependent variable for SAH from the Canadian National Population Health Survey, used the predictions from ordered probit or interval regressions. The advantage of using predictions is that they allow us to give a different value of health (in this case mortality) to each individual in the sample. Individuals in the sample are ranked by their predicted mortality. The estimated linear index allows a sufficient degree of individual variation in the measure of mortality but it associates to each individual either positive or negative values of mortality. The index is transformed in order to guarantee a positive support and to ensure that the Gini coefficient is positive[29]. The Gini coefficient for the excluded and the exogenous mortality equation, is 0.296 and 0.274 respectively. The measure of pure inequality in health, irrespective of any socio-economic dimension,

[29]The definition of the new dependent variable requires that the linear index $x\hat{b}$ is transformed in $x\hat{b}^*$, where $x\hat{b}^* = x\hat{b} - min(x\hat{b})$, which satisfies the condition $x\hat{b}^* \geq 0$. van Doorslaer and Jones (2003) note that the percentage contributions in the regression-based decomposition analysis are invariant to linear transformation of y.

Table 9
Percentage contributions to overall inequality in mortality

variable	Excl	Exog	MVP
sah		3.426	8.323
non smoker		1.373	9.634
breakfast		-0.229	3.086
sleep well		0.538	11.856
prudent drinker		-0.034	-0.024
non obese		0.279	0.062
exercise		1.207	0.686
lifestyles		3.134	25.301
social class	2.087	1.187	0.431
education	3.102	1.653	0.877
occupational status	3.636	2.333	-0.926
area	1.046	0.698	0.416
ethnicity	0.489	0.535	0.365
height	0.130	0.205	0.175
sex	5.006	4.509	3.418
age	84.503	82.322	62.156

is a bit smaller if health-related behaviours are considered as exogenous determinants of the health outcome. The multivariate probit model predicts a lower level of inequality, around 0.234.

We decompose overall health inequality using equation (3). Our dependent variable, predicted mortality, is additive in the regressors and it permits us to estimate only the deterministic part of the decomposition equation.

Table 9 presents the components of (3) as percentage contributions. The most important contribution to overall health inequality is attributable to age. Apart from age, gender, education and social class make relatively important contributions. These contributions turn out to be smaller if lifestyles are included in the model. Lifestyles contribute 3.13% and perceived health status contributes 3.4%.

Controlling for potential endogeneity of the lifestyle variables by means of the multivariate probit, gives very strong results in terms of the contribution of socio-economic variables to overall inequality in mortality. Social class' contribution diminishes by around 21%, falling from 2.09 without lifestyles to 0.43; education's contribution is reduced by 28%, falling from 3.10 to 0.88. The

contribution of age falls from 84.5% to 62%. The extent of the contribution of lifestyles is very large in the endogenous model, which predicts a 25% contribution to overall inequality, due in particular to the variables sleep well and non smoking[30].

The Gini coefficient of total health inequality and the estimation of the contribution of socio-economic variables to its variation, captures the effect of including health-related behaviours in the mortality equation. Even though the Gini coefficient does not measure the socio-economic dimension of inequalities in health directly, they do shed light on the nature of inequality by means of the decomposition approach.

8. Conclusion

We use the British Health and Lifestyle Survey (HALS, 1984-1985) data and the longitudinal follow-up of May 2003 to investigate the determinants of premature mortality risk in Great Britain.

We propose a simple behavioural model where the economic agent maximizes his lifetime utility. A value function is used to relate future utility to survival probability. Health investments decisions are assumed to influence longevity.

We relate the risk of mortality to a set of observable and unobservable factors. Observable factors influencing mortality are perceived health, socio-economic and demographic characteristics, ethnicity, type of area and individual health-related behaviours. Individuals' choices about their lifestyle may induce variations in health status and affect premature mortality. We assume that the relationship between the socio-economic environment and premature mortality is mediated by lifestyles. In order to assess the impact of lifestyles,

[30]Full results for the decomposition analysis are available from the authors upon request.

we estimate probit models and compare models without lifestyles and models which include them.

The main econometric issue that arises in our analysis is unobservable individual heterogeneity and endogeneity of the discrete explanatory variables that affect the mortality equation. Factors hidden to the researcher, like the rate of time preference, biological or genetic characteristics and past experiences, may influence individual demand for health and for health inputs. We propose a MSL approach to estimate a recursive system of equations for deaths, health and lifestyles, in order to correct for heterogeneity and potential endogeneity of self-assessed health and lifestyles. The multivariate probit model allows us to test whether unobservable characteristics influencing lifestyle also affect premature mortality. This model is then compared with a model without lifestyles and with a model that includes exogenous lifestyle variables.

The main economic concern is to detect inequality in the distribution of health within the population and to understand to what extent differences in social and economic characteristics contribute to inequality. We focus mainly on social class and education differences in the sample. We are critical of a crude measure of inequality used by epidemiologists. A more robust measure of health inequality is offered by the Gini coefficient for overall health inequality. We are able to decompose the Gini for predicted mortality and to compute health-related concentration indexes for all factors influencing mortality, including social class and education. We find that lifestyles, in particular smoking and sleep pattern, strongly contribute to inequality in mortality, reducing the relative contribution of socio-economic factors and of ageing.

Appendix A

Table A.1
Multivariate Probit for the reduced model (R = 50)

variable	**Eq. 1** *deaths*		**Eq. 2** *sah*		**Eq. 3** *nsmoker*		**Eq. 4** *breakfast*		**Eq. 5** *sleepgd*	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
sah	−0.555†	(0.326)								
non	−0.868**	(0.173)	0.054	(0.190)						
breakfast	0.414*	(0.211)	0.314	(0.272)						
sleep	−0.681**	(0.246)	0.080	(0.309)						
prudent	−0.098	(0.072)	−0.021	(0.075)						
non	−0.128†	(0.068)	0.180**	(0.064)						
exercise	−0.060	(0.058)	0.253**	(0.052)						
sc1	−0.069	(0.067)	0.194**	(0.061)	0.142*	(0.060)	0.050	(0.059)	0.033	(0.054)
sc3	−0.036	(0.064)	−0.139*	(0.058)	−0.084	(0.059)	−0.045	(0.059)	0.021	(0.055)
degree	0.046	(0.116)	0.124	(0.109)	0.113	(0.108)	0.183†	(0.106)	0.039	(0.095)
hvqA	−0.056	(0.113)	0.063	(0.104)	−0.044	(0.102)	−0.050	(0.100)	−0.047	(0.093)
no	0.063	(0.096)	−0.142	(0.088)	−0.215*	(0.084)	−0.275**	(0.084)	−0.073	(0.078)
other	−0.084	(0.142)	−0.080	(0.133)	−0.358**	(0.128)	−0.219†	(0.129)	0.051	(0.121)
part	0.164†	(0.099)	−0.059	(0.093)	−0.103	(0.083)	0.162*	(0.082)	0.318**	(0.078)
unemployed	0.259†	(0.140)	−0.179	(0.139)	−0.379**	(0.132)	−0.293*	(0.130)	0.230†	(0.131)
sick	0.284	(0.220)	−1.556**	(0.156)	−0.256†	(0.131)	−0.205	(0.127)	−0.207†	(0.124)
retired	−0.018	(0.095)	−0.248**	(0.091)	−0.272**	(0.089)	0.238**	(0.089)	0.145†	(0.081)
housekeeper	0.242*	(0.106)	−0.272**	(0.094)	−0.121	(0.091)	−0.048	(0.088)	0.196*	(0.085)
shift	−0.300*	(0.123)	0.007	(0.110)	−0.173†	(0.099)	−0.152	(0.096)	−0.281**	(0.094)
rural	−0.100	(0.072)	0.106	(0.070)	0.092	(0.069)	0.121†	(0.068)	0.056	(0.064)
suburb	−0.082	(0.057)	0.046	(0.056)	0.021	(0.054)	0.168**	(0.054)	−0.044	(0.051)
ethwheur	0.349†	(0.201)	0.459**	(0.167)	0.075	(0.163)	0.474**	(0.152)	0.354*	(0.153)
height	0.010	(0.009)	−0.004	(0.009)	0.010	(0.009)	0.016†	(0.009)	0.003	(0.008)
male	0.387**	(0.076)	−0.004	(0.072)	−0.250**	(0.072)	−0.115	(0.072)	0.048	(0.066)
age	0.004	(0.026)	−0.054**	(0.021)	−0.050*	(0.023)	0.024	(0.021)	−0.024	(0.019)
age2	0.058**	(0.021)	0.046**	(0.017)	0.063**	(0.019)	−0.009	(0.017)	0.010	(0.015)
widow			−0.173*	(0.081)	−0.225**	(0.084)	−0.187*	(0.085)	−0.197**	(0.074)
divorced			−0.125	(0.128)	−0.370**	(0.120)	−0.269*	(0.118)	−0.255*	(0.112)
separated			−0.010	(0.187)	−0.233	(0.175)	−0.335†	(0.174)	−0.332*	(0.164)
single			0.009	(0.111)	−0.157	(0.110)	−0.094	(0.110)	−0.191†	(0.100)
wales			−0.271*	(0.108)	−0.200†	(0.110)	0.117	(0.108)	0.032	(0.101)
north			−0.009	(0.114)	−0.341**	(0.104)	0.165	(0.105)	−0.144	(0.097)
nwest			−0.210*	(0.089)	−0.331**	(0.084)	0.239**	(0.085)	0.115	(0.078)
yorks			−0.100	(0.096)	−0.249**	(0.095)	−0.004	(0.093)	0.004	(0.086)
wmids			−0.089	(0.100)	−0.249*	(0.098)	0.006	(0.094)	−0.072	(0.089)
emids			−0.006	(0.102)	−0.085	(0.102)	−0.068	(0.096)	−0.084	(0.090)
anglia			−0.147	(0.127)	−0.083	(0.129)	−0.020	(0.126)	−0.094	(0.115)
swest			−0.164†	(0.096)	−0.085	(0.099)	0.065	(0.095)	0.001	(0.087)
london			−0.062	(0.097)	−0.176†	(0.096)	0.073	(0.094)	−0.035	(0.085)
scot			−0.239*	(0.098)	−0.375**	(0.092)	0.212*	(0.093)	−0.162†	(0.084)
house			−0.192	(0.144)	−0.108	(0.137)	−0.137	(0.133)	−0.508**	(0.128)
household			0.007	(0.026)	0.053*	(0.025)	−0.030	(0.024)	−0.019	(0.023)
other					−0.709**	(0.051)	−0.342**	(0.050)	−0.039	(0.048)
mother					−0.442**	(0.146)	−0.183	(0.141)	0.252†	(0.138)
father					−0.192*	(0.077)	0.071	(0.073)	0.087	(0.067)
both					−0.282**	(0.088)	−0.125	(0.084)	0.053	(0.078)
father					−0.044*	(0.021)	−0.047*	(0.021)	−0.019	(0.019)
mother					−0.046†	(0.025)	−0.020	(0.025)	0.006	(0.024)
cons	−2.533*	(1.052)	1.884*	(0.943)	1.785†	(0.914)	−1.544†	(0.874)	1.213	(0.808)

N	3655
Log-likelihood	-9925.197
$\chi^2_{(190)}$	2890.80

Significance levels: † : 10% ∗ : 5% ∗∗ : 1%

Table A.2
Variable Definitions and Summary Statistics

Variable Name	Variable Definition	Mean	S.D.
death	1 if has died at May 2003, 0 alive	0.359	0.480
sah	1 if self-assessed health is excellent or good, 0 if fair or poor	0.703	0.457
Lifestyle			
non smoker	1 if does not smoke, 0 if current smoker	0.700	0.459
breakfast	1 if does a healthy breakfast, 0 otherwise	0.707	0.455
sleep well	1 if sleeps between 7 and 9 hours, 0 otherwise	0.583	0.493
prudent drinker	1 if consume alcohol prudently, 0 otherwise	0.88	0.325
non obese	1 if is not obese, 0 otherwise	0.853	0.354
exercise	1 if did physical exercise in the last fortnight, 0 otherwise	0.323	0.468
Social Class			
sc1	1 if professional/student or managerial/intermediate, 0 otherwise	0.316	0.465
sc2	1 if skilled or armed service, 0 otherwise	0.467	0.499
sc3	1 if partly skilled, unskilled, unclass. or partner never occupied, 0 otherwise	0.218	0.413
Education Level			
degree	1 if University degree, 0 otherwise	0.125	0.331
hvqA	1 if higher vocational qualifications or A level or equivalent, 0 otherwise	0.125	0.331
O-cse	1 if O level/CSE, 0 otherwise	0.094	0.292
no edu.	1 if no qualification, 0 otherwise	0.608	0.488
other edu.	1 if other vocational/professional qualifications, 0 otherwise	0.047	0.213
Marital Status			
married	1 if married, 0 otherwise	0.761	0.427
widow	1 if widow, 0 otherwise	0.128	0.334
divorced	1 if divorced, 0 otherwise	0.038	0.192
separated	1 if separated, 0 otherwise	0.017	0.128
single	1 if single, 0 otherwise	0.057	0.231
Occupational Status			
shift worker	1 if shift worker, 0 otherwise	0.057	0.232
full time	1 if full time worker or student, 0 otherwise	0.364	0.481
part time	1 if part time worker, 0 otherwise	0.132	0.338
unemployed	1 if the individual unemployed, 0 otherwise	0.030	0.171
sick	1 if absent from work due to sickness, 0 otherwise	0.033	0.179

continued on next page

Table A.2 – *continued from previous page*

Variable Name	Variable Definition	Mean	S.D.
retired	1 if retired, 0 otherwise	0.339	0.473
housekeeper	1 if housekeeper, 0 otherwise	0.102	0.303
Geographical			
wales	1 if lives in Wales, 0 otherwise	0.058	0.233
north	1 if lives in North, 0 otherwise	0.065	0.247
nwest	1 if lives in North West, 0 otherwise	0.128	0.334
yorks	1 if lives in Yorkshire, 0 otherwise	0.086	0.281
wmids	1 if lives in West Midlands, 0 otherwise	0.080	0.272
emids	1 if lives in East Midlands, 0 otherwise	0.077	0.266
anglia	1 if lives in East Anglia, 0 otherwise	0.04	0.196
swest	1 if lives in South West, 0 otherwise	0.088	0.284
london	1 if lives in London, 0 otherwise	0.094	0.292
scotland	1 if lives in Scotland, 0 otherwise	0.097	0.295
Area			
rural	1 if lives in the countryside, 0 otherwise	0.219	0.414
suburb	1 if lives in the suburbs of the city, 0 otherwise	0.472	0.499
Ethnicity			
ethwheur	1 if White European, 0 otherwise	0.979	0.143
Physical			
male	1 if male, 0 otherwise	0.455	0.498
height	height in inches	65.950	3.703
age	age in years	57.468	11.673
age2	$\text{age}^2/100$	34.388	14.076
Tenure			
house owner	1 if own house, 0 otherwise	0.966	0.182
Household			
household size	number of other people in the house	1.651	1.272
others smoked	1 if anyone else in house smoked, 0 otherwise	0.351	0.477
Parental			
mother smoked	1 if only mother smoked, 0 otherwise	0.031	0.173
father smoked	1 if only father smoked, 0 otherwise	0.595	0.491
both smoked	1 if both parents smoked, 0 otherwise	0.246	0.431
father drinker	father, non to heavy drinker (0-4)	1.891	1.200
mother drinker	mother, non to heavy drinker (0-4)	0.912	0.981

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.

Adda, J. and Lechene, V. (2004). 'On The Identification Of The Effect Of Smoking On Mortality'. CeMMAP working papers CWP13/04. Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Auster, R., Levenson, I., and Sarachek, D. (1969). 'The production of health: an exploratory study'. *Journal of Human Resources*, vol. 4, pp. 411–436.

Becker, G. (1965). 'A Theory of the Allocation of Time'. *ECONOMIC JOURNAL*, vol. 75(299), pp. 493–517.

Belloc, N. B. (1973). 'Relationship of Health Practices and Mortality'. *Preventive Medicine*, vol. 2, pp. 67–81.

Belloc, N. B. and Breslow, L. (1972). 'Relationship of Physical Health Status and Health Practices'. *Preventive Medicine*, vol. 1, pp. 409–421.

Borg, V. and Kirstensen, T. S. (2000). 'Social class and self-rated health: can the gradient be explained by differences in life style or work environment?'. *Social Science and Medicine*, vol. 51, pp. 1019–1030.

Brown, H. S., Pagán, J. A., and Bastida, E. (2005). 'The Impact of Diabetes on Employment: Genetic IVs in a Bivariate Probit'. *Health Economics*. Forthcoming.

Cappellari, L. and Jenkins, S. P. (2003). 'Multivariate Probit regression using simulated maximum likelihood'. *The Stata Journal*, vol. 3(3), pp. 278–294.

Carbone, J. C., Kverndokk, S., and Rogeberg, O. J. (2005). 'Smoking, health, risk, and perception'. *Journal of Health Economics.* forthcoming.

Chang, F.-R. (2005). 'A theory of health investment under competing mortality risks'. *Journal of Health Economics*, vol. 24, pp. 449–463.

Chib, S. and Greenberg, E. (1997). 'Analysis of Multivariate Probit'. *Biometrika*, vol. 85(2), pp. 347–361.

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

Deaton, A. (2003). 'Health, Inequality and Economic Development'. *Journal of Economic Literature.*, vol. 41(1), pp. 113–158.

Fleurbaey, M. (2004). 'Health, Equity and Social Welfare'. Working Paper. Université de Pau et des Pays de l'Adour.

Fuchs, V. R. (1982). 'Economic aspect of health'. In V. R. Fuchs, e., editor, *Time preferences and health: An explanatory study.* University of Chicago Press for NBER: Chicago.

Grossman, M. (1972). 'On the Concept of Health Capital and the Demand for Health'. *The Journal of Political Economy*, vol. 80(2), pp. 223–255.

Grossman, M. and Joyce, T. J. (1990). 'Unobservables, Pregnancy Resolutions, and Birth Weight Production Functions in New York City'. *The Journal of Political Economy*, vol. 98(5), pp. 983–1007.

Health and Lifestyle Survey - University of Cambridge Clinical School (2003). 'Deaths (4^{st} update) and Cancer (1^{st} Update - 2^{nd} listing) May 2003 - Working Manual'. University of Cambridge, available at the web site http://www.data-archive.ac.uk/.

Hurley, J. (2000). 'An overview of the normative economics of the health sector'. *in Culyer, A.J., and J. P. Newhouse (eds.) Handbook of Health Economics, Elsevier.*

Kenkel, S. D. (1991). 'Health Behavior, Health Knowledge, and Schooling'. *Journal of Political Economy*, vol. 99(2), pp. 287–305.

Kenkel, S. D. (1995). 'Should you eat breakfast? estimates from health production functions'. *Health Economics*, vol. 4, pp. 15–25.

Knapp, L. G. and Seaks, T. G. (1998). 'A Hausman test for a dummy variable in probit'. *Applied Economics Letters*, vol. 5, pp. 321–323.

Lerman, R. I. and Yitzhaki, S. (1989). 'Improving the accuracy of estimates of gini coefficients'. *Journal of Econometrics*, vol. 42, pp. 43–47.

Lynch, J. W., Kaplan, G. A., Cohen, R. D., Tuoomilehto, J., and Salomen, J. T. (1996). 'Do Cardiovascular Risk Factors Explain the Relation between Socioeconomic Status, Risk of All-Cause Mortality, Cardiovascualr Mortality, and Acute Myocardial Infraction?'. *American Journal of Epidemiology*, vol. 144(10), pp. 934–941.

Lynch, J. W., Kaplan, G. A., and Salomen., J. T. (1997). 'Why do poor people behave poorly?'. *Social Science and Medicine*, vol. 44(6), pp. 809–819.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge Unversity Press edition.

Manor, O., Mattheus, S., and Power, C. (1997). 'Comparing measure of health inequality'. *Social Science and Medicine*, vol. 45(5), pp. 761–771.

Marmot, M., Ryff, C. D., Bumpass, L. L., Shipley, M., and Marks, N. F. (1997). 'Social inequalities in health: next questions and converging evidence'. *Social Science Medicine*, vol. 44(6), pp. 901–910.

Marmot, M. G. (2004). *Status Syndrome.* London: Bloomsbury.

Marmot, M. G., Shipley, M. J., and Rose, G. (1984). 'Inequalities in health - specific explanation of a general pattern?'. *The Lancet*, vol. 323(8384), pp. 1003–1006.

Marmot, M. G., 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*, vol. 337(8754), pp. 1387–1393.

Morris, S., Sutton, M., and Gravelle, H. (2003). 'Inequity and inequality in the use of health care in england: an empirical investigation'. *CHE Technical Paper Series 27.*

Mullahy, J. and Portney, P. (1990). 'Air Pollution, Cigarette Smoking, and the Production of Respiratory Health'. *Journal of Health Economics*, vol. 9, pp. 193–205.

Mullahy, J. and Sindelar, J. (1996). 'Employment, Unemployment, and Prob-

lem Drinking'. *Journal of Health Economics*, vol. 15, pp. 409–434.

Poston, D. L. and Micklin, M. (2005). *Handbook of Population*. Handbooks of Sociology and Social Research Series. Kluwer Academic Publishers.

Roemer, J. E. (1998). *Equality of Opportunity*. Harvard U. Press.

Schmidt, P. (1981). 'Constraints on the Parameters in Simultaneous Tobit and Probit Models'. In *Structural Analysis of Discrete Data and Econometric Applications*. by Manski, C. F. and D. L. McFadden, The MIT Press, Cambridge.

Smith, J. P. (1999). 'Healthy Bodies and Thick Wallets: The Dual Relation between Health and Economic Status'. *The Journal of Economic Perspective*, vol. 13(2), pp. 145–166.

van Doorslaer, E. and Jones, A. M. (2003). 'Inequalities in self-reported health: validation of a new approach to measurement'. *Journal of Health Economics*, vol. 22, pp. 61–87.

van Doorslaer, E. and Koolman, X. (2004). 'Explaining the differences in income-related health inequalities across european countries'. *Health Economics*, vol. 13(7), pp. 609–628.

Wagstaff, A. (1986). 'The demand for health: theory and applications'. *Journal of Epidemiology and Community Health*, vol. 40, pp. 1–11.

Wagstaff, A., Paci, P., and Joshi, H. (2001). 'Causes of inequalities in health. Who you are? Where you live? Or who your parents are?'. Policy Research Working Paper 2713, World Bank.

Wagstaff, A., van Doorslaer, E., and Watanabe, N. (2003). 'On decomposing the causes of health sector inequalities with an application to malnutrition inequalities in Vietnam'. *Journal of Econometrics*, vol. 112(1), pp. 207–223.

Wilde, J. (2000). 'Identification of multiple equation probit models with endogenous dummy variables'. *Economic Letters*, vol. 69(1), pp. 309–312.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Massachussetts.