HEDG Health, Econometrics and Data Group
THE UNIVERSITY of York

HEDG Working Paper 07/18

Panel data methods and applications to health economics

Andrew M. Jones

August 2007

Panel Data Methods and Applications to Health Economics

Andrew M. Jones
University of York

Prepared for *The Palgrave Handbook of Econometrics Volume II: Applied Econometrics*, Terence C. Mills & Kerry Patterson (eds), Basingstoke: Palgrave MacMillan.

Acknowledgements: This chapter draws on joint work with Teresa Bago d'Uva, Silvia Balia, Paul Contoyannis, Cristina Hernández Quevedo, Xander Koolman, José Labeaga, Miaw-chwen Lee, Roberto Leon Gonzalez, Nigel Rice, Stefanie Schurer, Eddy van Doorslaer and John Wildman and by Casey Quinn. I am grateful for comments from Giorgia Marini, Pedro Rosa Dias and João Santos Silva.

Abstract

Much of the empirical analysis done by health economists seeks to estimate the impact of specific health policies and the greatest challenge for successful applied work is to find appropriate sources of variation to identify the treatment effects of interest. Estimation can be prone to selection bias, when the assignment to treatments is associated with the potential outcomes of the treatment. Overcoming this bias requires variation in the assignment of treatments that is independent of the outcomes. One source of independent variation comes from randomised controlled experiments. But, in practice, most economic studies have to draw on non-experimental data. Many studies seek to use variation across time and events that takes the form of a quasi-experimental design, or "natural experiment", that mimics the features of a genuine experiment. This chapter reviews the data and methods that are used in applied health economics with a particular emphasis on the use of panel data. The focus is on nonlinear models and methods that can accommodate unobserved heterogeneity. These include conditional estimators, maximum simulated likelihood, Bayesian MCMC, finite mixtures and copulas.

Contents

1. Introduction

A common thread that runs through this chapter is the "evaluation problem": is it possible to identify the impact of policies from empirical data? The focus of the chapter is on individual-level longitudinal data, so consider an "outcome" y_{it}, for individual i at time t. The treatment effect of interest is:

$$TE_{it} = \Delta_{it} = y_{it}^1 - y_{it}^0 \tag{1}$$

where *1* denotes treatment and *0* denotes control[1]. The pure treatment effect cannot be identified because the counterfactual can never be observed: each individual is either treated or untreated at a particular point in time so only one of the potential outcomes can be observed. The outcome that is actually observed can be written in terms of the potential outcomes:

$$y_{it} = y_{it}^0 + d_{it}(y_{it}^1 - y_{it}^0) \tag{2}$$

where d_{it} is an indicator of treatment.

One response to the problem of defining a counterfactual is to concentrate on the average treatment effect (ATE), comparing the average outcomes between the treated and controls:

$$ATE = E\left(y_{it}^1 - y_{it}^0\right) \tag{3}$$

When there is heterogeneity in individual responses to the treatment, that may influence the assignment of treatment, for example when doctors select patients on the basis of their capacity to benefit, attention is likely to focus on the average treatment effect on the treated (ATT) rather than the ATE:

$$ATT = E\left(y_{it}^1 - y_{it}^0 \mid d_{it} = 1\right) \tag{4}$$

This is the average effect of treatment for those individuals who would actually select into treatment.

Moving towards a regression framework, assume that the observed outcome under the two treatment regimes is given by the general regression models:

$$y_{it} = f_j(x_{it}, u_{jit}),\ j = 0,1 \tag{5}$$

[1] This chapter uses the "treatment-outcome" terminology that is commonplace in the evaluation literature. In practice many treatments are broad policy reforms associated with the financing and delivery of health care rather than specific clinical interventions. Treatment effects are defined here in terms of a binary treatment with just two regimes – the treated and the controls. In practice there may be multiple treatments and varying intensities of treatment.

The vector *x* includes observable factors that influence the outcome and that may influence the assignment of treatment (reflecting "selection on observables"). The *u* are unobservable factors that influence the outcomes and may influence the assignment of treatment ("selection on unobservables"). Formulating the problem in this way requires SUTVA (stable unit-treatment value assumption) to hold – an individual's potential outcomes and treatments are independent of others in the population, ruling out spill-over and general equilibrium effects. These spill-overs may be important in some health economics applications and the evaluation of treatment effects would then have to be designed to accommodate them (Chandra and Staiger (2007); Miguel and Kremer (2004)). Using linear functions for *f(.)*, gives a switching regression model:

$$y_{it} = x_{it}^{'}\beta_j + u_{jit},\ j = 0,1 \tag{6}$$

A simplification of this model, which assumes a homogeneous treatment effect so that only the intercept varies with treatment, gives the regression function:

$$y_{it} = x_{it}^{'}\beta + d_{it}\delta + u_{it} \tag{7}$$

In this case ATE=ATT=δ.

If unobserved factors (*u*) influence whether an individual is selected into the treatment group or how they respond to the treatment, this will lead to biased estimates of the treatment effect. A randomised experimental design may achieve the desired orthogonality of measured covariates (*x, d*) and unobservables (*u)*. However, econometric studies typically rely on observational data gathered in a non-experimental setting. One strategy is to rely on selection on observables: finding a sufficiently rich set of observable characteristics so that unobservables can be assumed to have no systematic influence on treatments. This approach includes matching estimators and inverse probability weighted estimators. In contrast the selection on unobservables strategy looks for factors that predict treatment but have no direct effect on outcomes and that can therefore be used to mimic random assignment of treatment. This approach includes using within-individual variation to allow for time invariant individual heterogeneity in panel data models (fixed effects) as well conventional instrumental variables (IV) estimators. It also includes multiple equation models in which equations for the treatment and outcome are estimated jointly by full information maximum likelihood (FIML). "Natural experiments" often lead to the use of difference-in-differences estimators, which combine selection of observables (by including *x* in the regression models) with selection on unobservables (by using differencing to control for time invariant heterogeneity).

Natural experiments are often also linked to instrumental variables (IV) estimation, which relies on instruments (*z*) that predict the assignment of treatment but do not have a direct effect on the outcome. When there is heterogeneity in the response to treatment the IV estimator identifies a local average treatment effect, or LATE (Imbens and Angrist (1994); McClellan et al. (1994)). This is the average treatment effect over the sub-group of the population that are induced to participate in the treatment by variation in the instrument. The fact that IV estimates only identify the LATE and that the results are therefore contingent on the set of instruments explains why different empirical studies can produce quite different estimates, even though they examine the same outcomes and treatments. Heterogeneity in treatment effects is likely to be widespread: for example, Auld (2006a) finds considerable heterogeneity in the treatment effect of local HIV

infection prevalence on risky sexual behaviour among gay men in the San Francisco Men's Health Study (SFMHS), with HIV prevalence having less impact among those at high risk.

Recent work by Heckman and Vytlacil has extended the analysis of local treatment effects by specifying a model for the assignment of treatment and using it to identify those individuals who are indifferent between treatments, given x and z (Heckman and Vytlacil (1999); Heckman and Vytlacil (2007); and see Basu et al. (2007) for an application to health data). This defines the marginal treatment effect (MTE): the treatment effect among those individuals at the margin. The MTE provides a building block for the LATE, ATT and ATE. It can be identified using Local-IV methods or by specifying multiple equation models with a common factor structure (e.g., Aakvik et al. (2005); Basu et al. (2007)).

For example in Aakvik et al. (2005) the treatment is a Norwegian vocational rehabilitation (VR) programme and the outcome is a binary measure of employment. Analysis is based on a 10% sample of all those who applied for VR in 1989. To define the treatment effects of interest Aakvik et al. (2005) specify a discrete choice model with a common factor structure. There is a switching regression for the binary indicator of employment under the two treatment regimes:

$$\begin{aligned} y_i^1 &= f_1(x_i, u_{1i}) = 1(x_i'\beta_1 \geq u_{1i}) \\ y_i^0 &= f_0(x_i, u_{0i}) = 1(x_i'\beta_0 \geq u_{0i}) \end{aligned} \tag{8}$$

Along with a latent variable model for the assignment of treatment:

$$\begin{aligned} &d_i = 1 \\ &if \\ &d_i^* = z_i'\beta_d - u_{di} > 0 \end{aligned} \tag{9}$$

The error terms are assumed to have a common factor structure:

$$\begin{aligned} u_{di} &= -\eta_i + \varepsilon_{di} \\ u_{1i} &= -\alpha_1\eta_i + \varepsilon_{1i} \\ u_{0i} &= -\alpha_0\eta_i + \varepsilon_{0i} \end{aligned} \tag{10}$$

Estimation is by FIML assuming that the error components are jointly normal.

Given this set up, the treatment effects of interest can be defined as follows:

$$MTE(x,u) = E(\Delta \mid x, d^* = 0) = E(\Delta \mid x, u_d = z_i'\beta_d) \tag{11}$$

$$ATE(x) = \int MTE(x,u)dF(u) = E(\Delta \mid x) \tag{12}$$

$$ATE = E(ATE(x)) = \int E(\Delta \mid x)dF(x) \tag{13}$$

$$ATT(x,u) = E(\Delta \mid x, d = 1) = E(\Delta \mid x, u_d < z_i'\beta_d) \tag{14}$$

Aakvik et al. (2005) do not use the concept of the LATE in their study but, based on the notation of their model, it could be expressed as:

$$LATE(x,z,\tilde{z}) = E\left(\Delta \mid x, z_i'\beta_d < u_d < \tilde{z}_i'\beta_d\right) \quad (15)$$

Where, for illustration, it is assumed that assignment to treatment is monotonically related to a single instrument that takes two values z and $\tilde{z}$, where $z_i'\beta_d < \tilde{z}_i'\beta_d$. The LATE defines the treatment effect for all those individuals who are induced into the treatment by the change in the instrument (see e.g., Basu et al. (2007)).

The nonlinear model is identified by functional form but an exclusion restriction is also imposed by including an instrument – the degree of rationing of VR places in the individual's locality – in z but not in x. The apparent positive impact of the VR programme is reversed when selection bias is taken into account and there is evidence of perverse cream-skimming, with those most likely to benefit being the least likely to be selected by the programme administrators.

A note on the scope of the chapter

This chapter takes the identification of treatment effects as its starting part and concentrates on microeconometric methods that can be used with longitudinal and other complex and multilevel datasets. Although the methods described in the chapter are widely used throughout applied econometrics, the applications reviewed here all relate to one specific area: health economics. The chapter follows an earlier review of the literature on "health econometrics" (Jones (2000)) and concentrates on studies that have appeared as peer-reviewed publications from 2000 onwards. The emphasis is on applications that use health and health care as outcomes. Less attention is devoted to the large number of studies of health-related behaviours such as diet, smoking, drinking and illicit drugs (see e.g., Cowell (2006); Dee et al. (2005); Forster and Jones (2001); Harris et al. (2006); Terza (2002); Van Ours (2006)) and to those that investigate the impact of health, health care and health insurance on labour market outcomes (see e.g., Askildsen et al. (2005); Au et al. (2005); Auld (2002); Bradley et al. (2005); Contoyannis and Rice (2001); Disney et al. (2006); French (2005); Hogelund and Holm (2006); Morris (2006); Morris (2007); Royalty and Abraham (2006); Stewart (2001); Van Ours (2004)) or labour outcomes for health care professionals (see e.g., Arulampalam et al. (2004); Holmas (2002); Frijters et al. (2006)). The scope does not include studies that use econometric techniques in the context of contingent valuation and discrete choice experiments, where random effects models are often applied (see e.g., Ryan et al. (2006)); multinomial models of the choice of insurance plans or health care providers (see e.g., Deb and Trivedi (2006a); Ho (2006); Sahn et al. (2003)); productivity analysis based on models of cost and production functions and estimation of stochastic frontier models (see e.g., Bradford et al. (2001); Burgess (2006); Dranove and Lindrooth (2003); Smith and Street (2005); Wilson and Carey (2004)); and in the context of cost-benefit and cost-effectiveness analysis, where econometric methods are starting to be used alongside methods from biostatistics and epidemiology (see e.g., Hoch et al. (2002); Willan et al. (2004); Briggs (2006)).

The focus is primarily on studies that use micro-level data derived from longitudinal, multilevel and other complex data structures. Relatively few cross section studies are discussed and the chapter does not attempt to review studies that use aggregate time series or panels and that apply pure time series methods (see e.g., Aakvik and Holmas

(2006); Abadie and Gay (2006); Chou (2007); Wang and Rettenmaier (2007); Garcia-Ferrer et al. (2007); Paton (2002); Leigh and Jencks (2007); Or et al. (2005); Ruhm (2003)). Analysis of longitudinal data often makes use of the methods of survival analysis (see e.g., Arulampalam et al. (2004); Chou (2002); Disney et al. (2006); Farsi and Ridder (2006); Forster and Jones (2001); Frijters et al. (2006); Harrison (2007); Holmas (2002); Kyle (2007); Picone et al. (2003a); Stewart (2001); Van Ours (2004); Van Ours (2006)) but these methods are not discussed in detail here.

The empirical findings of many of the studies are discussed but no attempt is made to provide a systematic synthesis of the empirical results. It is notable that meta-analyses of regression results are beginning to appear in the health economics literature. For example Gallet and List (2003) present a meta-analysis of the tobacco price elasticity and Gemmill et al. (2007) carry out a meta-regression of estimates of the price elasticity of prescription drugs.

On the whole the original sources for the econometric methods are not cited. These are reviewed in Volume 1 of this Handbook (particularly the chapters by Badi Baltagi, William Greene and Lung-fei Lee) and in other chapters in this Volume, in particular those by Colin Cameron, William Greene, Costas Meghir and Pravin Trivedi.

2. Identification strategies: finding relevant variation

The success of applied work depends on finding appropriate sources of variation to identify the effects of interest. Estimation of treatment effects can be prone to selection bias, where the assignment to treatments is associated with the potential outcomes of the treatment. Overcoming this selection bias requires variation in the assignment of treatments that is independent of the outcomes. One source of independent variation comes from randomised controlled experiments. While these are the norm in the evaluation of new clinical therapies, their use for the evaluation of social programmes remains rare (Gertler (2004); Kremer (2003); Miguel and Kremer (2004)). Most economic studies have to draw on non-experimental, or observational, data. This section presents a series of case-studies from the recent literature and describes how these have sought out relevant identifying information.

2.1 Randomised experiments

The "gold standard" methodology that is used to identify the efficacy and effectiveness of new medical technologies is the randomised clinical trial (RCT). Much of the work done by health economists to measure the cost-effectiveness of these technologies draws on data collected within RCTs to perform statistical analyses (Briggs (2006)) or to calibrate decision analytic models (Claxton et al. (2006)). Econometric methods are sometimes used in secondary analysis of such data to model costs and outcomes as functions of observable covariates (Willan et al. (2004))

Broader randomised social experiments are far less prevalent. One exception, that has played a highly influential role in the development of the health economics and that has driven many of the early developments in the use of econometrics in the field, is the RAND Health Insurance Experiment (Manning et al. (1987)). The RAND experiment

was designed to address the problem of self-selection in the choice of insurance plans. Participants were randomised between HMO and reimbursement plans and across plans with different levels of copayments and a plan with deductibles. The RAND study has had a strong influence, especially in the US. It has focused attention on the use of two-part or multi-part models to model health care utilisation and expenditures and on the choice of functional form to deal with heavily skewed data and the consequent problems of retransformation back to the 'natural scale' in the presence of individual heterogeneity (Manning (2006)). The RAND data are available over the internet and have been used to test more recent developments of econometric methods (Deb and Trivedi (2002); Bago d'Uva (2006); Gilleskie and Mroz (2004); Vera-Hernandez (2003)).

More recently randomised experiments have begun to play an influential role in research and policy in developing countries. This is exemplified by the studies of Gertler (2004) and Miguel and Kremer (2004). The Mexican government's PROGRESA programme, which was initiated in 1997, has received considerable attention and has influenced policy throughout Latin America. The programme relies on conditional cash transfers that are designed to influence the use of health and welfare services for children in poor families. It covers 2.6 million families in 50,000 rural villages. The programme focuses on health, hygiene and nutrition. It links substantial cash transfers, on average amounting to 20-30% of household income, to the use of prenatal care, well-baby care and immunization, nutrition monitoring and supplementation, preventive check-ups and participation in educational programmes. PROGRESA works by first selecting whole communities to participate in the scheme and then selecting households within those communities that satisfy the eligibility criteria to receive the benefits of the scheme. Financial constraints on the implementation of PROGRESA meant that its introduction was phased. To make the implementation equitable, communities were selected randomly to receive the benefits either immediately or with a delay. The random phasing provides researchers with an ideal opportunity to use a randomised design in the evaluation of the impact of the programme. Of the communities selected for the programme, 320 were randomly selected to receive the intervention in August-September 1998 with the remaining 185 delayed for two years. The communities in the control group were not informed that they would eventually receive the programme, reducing the scope for anticipation of treatment to influence the outcome.

Gertler (2004) focuses on health outcomes among children. These include self-reported morbidity, measured by illnesses in the past month, as reported by the child's mother, and objective measures including anthropometric measures of height and stunting and a biomarker for anaemia (haemoglobin levels). The analysis is restricted to those households, in both the treatment and control groups, that satisfy the eligibility criteria for PROGRESA. Although the data is randomised, multivariate regression models are used to control for observed covariates and reduce idiosyncratic variation. Individual and village random effects are included, the latter to allow for the clustered sampling. The results show significant improvements in both self-reported and objective measures of health and the impact increases with length of exposure to the programme. Gertler (2004) is careful to note that the comparison of treated and controls does not explain the mechanism behind this effect: for example, it is not possible to say whether an unconditional transfer would have had the same effect as the conditional one.

In their "worms" paper Miguel and Kremer (2004) analyse a randomised experiment to evaluate the impact of the Kenyan Primary School Deworming Project (PSDP) on hookworm infection rates and on school attendance. The programme included drug

therapy and public health education on avoiding hookworm infection, with the assignment of treatment randomly phased. Randomisation was done at the level of schools rather than individuals: one group of schools received treatment in 1998 and 1999, another group only in 1999 and a third group only in 2003. Data were collected in 1998 and 1999 so, in the Miguel and Kremer (2004) study, the first group are the treated and the second and third groups make up the controls. Miguel and Kremer (2004) argue that randomisation at the levels of schools is crucial in this context as it avoids biases created by spill-over effects of the deworming programme in reducing infection rates. They argue that an ideal prospective study would randomise treatments across pupils within schools, across schools within clusters and across these clusters. This multilevel variation in the assignment of treatments could then be used to estimate different levels of the treatment effect in the case where spill-overs are important.

2.2 Natural experiments

Health shocks

Almond (2006) makes inventive use of the 1918 influenza pandemic as a natural experiment to provide evidence in favour of the "fetal origins hypothesis". Cohorts that were in utero during the pandemic, between the Fall of 1918 and January 1919, are shown to have poorer outcomes: lower educational attainment, more disability, lower income, lower socioeconomic status and higher transfer payments. The pandemic has the potential to be used as a natural experiment: it was unanticipated, the period of exposure was short and the impact varied systematically across states. The study uses discontinuity across birth cohorts to identify the long-term effects, drawing on data from the 1960, 70 and 80 US Census microdata (which identify quarter of birth). Geographic variation is also exploited, based on the 'laggard' states where the epidemic had less pronounced long-term effects, although this does reduce the sample size available. This is a paper where simple graphical analysis tells the main story, although this is backed up by thorough statistical modelling.

Doyle (2005) makes innovative use of data on severe traffic accidents to measure variation in unanticipated health shocks and finds that, in the United States, the uninsured receive 20% less treatment and have a substantially higher mortality rate. The Crash Outcome Evaluation System (CODES) links police accident reports to hospital discharge data. This study uses data for Wisconsin covering 1992-97 with a sample of 28,236, ten per cent of whom were uninsured. Severe traffic accidents are assumed to be unanticipated at the time that insurance is taken out and the consequent use of health care is non-discretionary. Descriptive evidence suggests that the uninsured are riskier drivers and have worse health problems, creating a problem of selection bias. To deal with selection a control group is selected from those with medical insurance but without car insurance. Within-hospital variation and time effects are controlled for. The robustness of the findings is checked by using the sub-sample where both insured and uninsured individuals are injured in the same crash, this allows for different severities of accident (accident fixed effects). Also the sample of passengers is used to abstract from differences in the quality of driving between the treated and control groups. Robustness is also assessed by doing separate analyses by diagnoses and by medical procedures. In the light of these results, the lower levels of treatment for the uninsured are attributed to decisions made by providers in response to insurance status rather than differences in background characteristics of the patients. Similar issues are faced by Levitt and Porter

(2001) who address the problem of selection bias in the US Fatality Analysis Reporting System (FARS) which they use for an analysis of the effectiveness of seat belts and air bags. The problem arises because data are only included for fatal crashes. The use of safety devices influences the probability of survival and hence of inclusion in the sample. The identification strategy adopted to get around this problem is to use a sample based on crashes where someone in a different car dies. The aim is to make the sample selection independent of the observation's own treatment status and outcomes. In doing so they find that seat belts are more effective and air bags are less effective than previous evidence had suggested.

Economic shocks

Evans and Lien (2005) make use of the 1992 Port Authority Transit (PAT) strike in Allegheny County, Pennsylvania as a source of independent variation in access to prenatal care. Prenatal visits were affected most for black women and city residents (in Pittsburgh) and the results show that, for these groups, missing visits early in pregnancy had a detrimental effect but missing those later in the pregnancy did not. The main source of information is observational data from the 1990-94 US Natality Detail Files, which contain a census of births in each given year, taken from birth records. This is augmented by survey data that is used to assess the impact of the strike on access to prenatal care. A control group of counties that were not affected by the strike are selected on the basis of regression analyses. The use of prenatal care by women who were pregnant at the time of the strike is included in regression equations for birth weight, gestation, maternal weight gain and maternal smoking. Models are estimated by OLS and 2SLS, the latter using the strike as an instrument. These produce similar results, suggesting that selection bias is not a problem. The clearest effect of prenatal care is on maternal smoking. The robustness of the findings is tested by checking for general a decline in earnings or employment coincident with the strike and for evidence of increases in abortions or 'unwanted' births.

In Frijters et al. (2005) the reunification of Germany in 1990 provides a natural experiment to assess the causal effect of income on self-reported health satisfaction. A positive and statistically significant effect is found, but the effect is small. The increase in incomes for those in East Germany is used as a source of independent variation in income that is not contaminated by reverse causality from health. The suitably of this setting as a natural experiment is justified by the fact that the changes in income associated with the fall of the Berlin Wall are assumed to have been unanticipated, that the income transfers were large in magnitude (affecting the real value of savings, collectively bargained wages and pay in general) and that there was individual variation in the impact, with civil servants experiencing an immediate effect. This variation is exploited in an econometric framework that also allows for entry and attrition from the panel dataset and for inherent individual heterogeneity. The analysis uses longitudinal data from the German Socioeconomic Panel (GSOEP) from 1984 to 2002 for West Germans and for 1990 to 2002 for East Germans. Data for East Germans is not available prior to reunification, so separate models are estimated for East and West Germans and the natural experiment has to be used indirectly.

In 1996 a crisis in the public pension system in Russia meant that 14 of the 39 million state pensioners faced substantial arrears in their payments. Jensen and Richter (2004) exploit this pensions crisis as a natural experiment. Their findings show a doubling of

poverty rates, significant declines in calorie and protein intake, and reductions in the use of health services and medications. They also show evidence of attempts to mitigate the loss of pension income through work, sales of assets, borrowing and private transfers. Data from the Russian Longitudinal Monitoring Survey (RLMS) for 1995 and 1996 are used to assess the impact of the crisis. Identification stems from geographic variation in arrears which arises because decisions were regionally decentralised across oblasts and discretion was exercised within oblasts. The control group is made up of households who continued to receive their pensions. Estimation uses a difference-in-differences design, with the policy effect measured by an interaction between the post-1996 period and whether an individual's pension was in arrears. The identification strategy relies on the assignment of arrears not being associated with outcomes prior to the crisis. The paper attempts to assess the validity of this assumption and presents evidence of a common trend for treated and controls prior to the crisis.

In contrast to Jensen and Richter (2004), Duflo (2000) uses a positive economic shock associated with public pensions as a source of exogenous variation in income in her study of child health in South Africa. The end of the Apartheid era in the early 1990s lead to large increases in benefits for black Africans within the South African Old Age Pension System. Duflo's study uses cross section data collected during 1993 and faces a selection problem, as children living in households with pension recipients are more likely to be disadvantaged and to live in rural areas. Her identification strategy compares eligible and non-eligible households and those children exposed to the increased household pension income for all of their lives or for only a fraction of their lives. Outcomes are measured using height-for-age z-scores and there is evidence of an effect on child health and nutrition. This effect is entirely attributable to pensions received by women and the effect is strongest for girls.

Chay and Greenstone (2003) bring together a comprehensive set of data sources within a quasi-experimental research design to investigate the impact of atmospheric pollution on infant health in the US. They find significant effects of total suspended particles (TSPs) on infant mortality, mostly driven by deaths within one month of birth. There is heterogeneity in this effect with the impact on infant mortality rates being twice as large among blacks. Identification is based on geographic variation in the impact of the 1981-82 recession on levels of TSP, which is treated as a source of random variation. County level data from various sources are merged for the period 1978-84. First differenced (fixed effects) models are used and, unusually, differenced models with fixed effects for the trend (double differencing) are also used. The latter allows for heterogenity in trends. The analysis goes a step further by selecting neighbouring counties as controls. This uses non-manufacturing counties that either are or are not neighbours to manufacturing counties to try and isolate the effects of pollution from the socioeconomic effects of the recession.

Lindahl (2005) shows how lottery winnings can provide one source of exogenous variation in income, in an attempt to overcome the selection biases inherent in disentangling the socioeconomic gradient in health. There is a statistically significant effect of income on morbidity and mortality and the magnitude of this effect is largely unchanged when lottery winnings are used as an instrument, although the estimates are less precise. This income effect is not apparent for the sub-sample aged over 60. Data from the Swedish Level of Living Surveys (SLLS) for 1968, 74 and 81 are matched with register data on income and deaths up to 1997. Morbidity is measured by combining 48 symptoms into a standardised measure and mortality is measured as death within 5 or 10

years of the surveys. Lottery winnings are treated as a source of exogenous variation in income: assuming that the variation is independent of health. Models are estimated with lottery winnings included directly. Then OLS and instrumental variable estimates (using winnings as the instrument) are compared for the sample of individuals who are identified as "players". The magnitudes of the income effects are similar although standard errors are inflated when IV is used. A similar strategy is adopted by Gardner and Oswald (2007). They use data on the GHQ-12 measure of psychological well-being from the British Household Panel (BHPS) for 1996-2003 and compare those who received lottery winnings of between £1000 and £120,000 to two control groups, those with smaller wins and those with no wins. The study finds a statistically significant effect of 1.4 GHQ-12 points after two years (compared to the average drop of 5 points associated with widowhood). An important caveat is the small number of treated cases: there are only 137 observations with large lottery wins.

In Van den Berg et al. (2006) the state of the economy during infancy is shown to have long-term consequences for mortality rates in this inventive study of those born in the 19th Century in the Netherlands. The analysis finds a significant effect of the stage of the business cycle (boom or bust) at the time of birth on individuals' subsequent age of death. Data from the Historical Sample of the Netherlands (HSN), drawn from registers of births, marriages and deaths, covers 14,000 individuals born between 1812 and 1912 with follow-up to 2000. These data are merged with macroeconomic time series that are used to identify the phases of the business cycle. Macroeconomic conditions early in life are used as an instrument for socioeconomic conditions in infancy in order to avoid the problems of unobservable heterogeneity bias that plague cross section comparisons. The impact of early life conditions is analysed nonparametrically, comparing those born in booms and recessions, and through duration analysis of the individual mortality data.

Educational reforms

Lleras-Muney (2005) shows how historical changes in the US educational system can be used as a natural experiment, based on a discontinuity design, to identify the effect of education on adult health. The estimated effect is larger than previous studies have suggested, with the magnitude of the instrumental variables estimate of the local average treatment effect three times larger than the OLS estimate. The natural experiment is based on changes across states in compulsory schooling and child labour laws between 1915 and 1939. Identification stems from variation over states and across time in the age at which children had to enter school, the age at which they could leave school and get a work permit, and whether those with work permits had to continue in school part-time. Estimation uses a regression discontinuity design, which attributes any jumps associated with school leaving age to the policy effect. This is applied using a linear probability specification for deaths as a function of years of schooling. Synthetic cohorts are constructed from successive US censuses (1960, 70 & 80) to select those who were aged 14 between 1915 and 1939 and to follow-up subsequent mortality rates. These are synthetic in the sense that they do not follow the same individuals and are based on gender, birth cohort and state of birth.

Educational reforms are also used as a natural experiment in Arendt (2005). In this case the analysis focuses on Denmark and reforms in 1958, that removed formal tests before middle school, and 1975, that increased the compulsory minimum school leaving age. Data are taken from the Danish National Work Environment Cohort Study (WECS),

with two waves in 1990 and 1995, and covers workers aged 18-59 in 1990. The impact of years of schooling on outcomes later in life is estimated using two-stage conditional maximum likelihood (2SCML) estimates, allowing for a random individual effect, for self-reported health. Models are also estimated for body mass index and for an indicator of never having smoked. The latter is included for comparison as, for most people, it is determined while they are still in education. The impact of education on health is amplified when instruments are used but at the same time the standard errors are inflated so that exogeneity is not rejected. However tests suggest that there may be a problem of weak instruments, as the reforms have low explanatory power in the reduced form equations.

Health policies and reforms

Bleakley (2007) is a good example of combining a natural experiment with a long-term follow-up to explore the economic consequences of a public health intervention aimed at children. A programme aimed at the eradication of hookworm is shown to lead to a long-term gain in the income of beneficiaries. Areas with greater scope for benefiting, due to higher levels of hookworm infection, show greater contemporaneous increases in school enrolment and attendance and in literacy among children. The natural experiment is the Rockefeller Sanitary Commission's (RSC) funding of treatment and education programmes to eradicate hookworm in the Southern US which took place around 1910-1915. This policy intervention was implemented over a well-defined and relatively short period. Geographic differences in infection rates prior to the intervention can be used to formulate and identify a treatment/control design, estimated using difference-in-differences. Data on long-term consequences for the cohorts exposed to the eradication programme are obtained from the US Census available through the Integrated Public Use Micro Sample (IPUMS).

In 1966 the Ceausescu regime in Romania banned abortion and family planning. Birth rates doubled the following year. In Pop-Eleches (2006) this provides an interesting contrast to studies that have examined moves in the opposite direction in the United States. The raw data shows an improvement in educational attainment and labour market outcomes associated with the ban but these results are reversed by allowing for compositional changes in the type of families having children. The findings are explained by the fact that affluent urban women were more likely to have abortions and use contraception before the ban. Data are drawn from a 15% sample of the 1992 Romanian census and focus on children born between January and October 1967. There is a spike in births between July and October due to the ban, but all of these children entered school in the same year and experienced the same over-crowding effect. Although it is not labelled as such, the paper uses a discontinuity design estimating a simple difference equation that includes a dummy variable for the period after the policy. Additional covariates are included but there is no control group and identification relies on any sudden changes in outcomes for those born just before or just after the ban.

Lakdawalla et al. (2006) present a careful application of the method of instrumental variables, based on state-level variation in Medicaid eligibility in the US, which shows that an unintended consequence of highly active antiretroviral therapy (HAART) is to increase risky sexual behaviour among patients who are HIV+. Simple correlations show lower sexual activity among those who are HIV+, but this is because of the debilitating effects of the disease and does not show the causal effect of treatment. Panel data on

HIV+ patients in care are taken from the HIV Costs and Sevices Utilization Study (HCSUS) for the period 1996-98. The outcome of interest is the number of sex partners of the previous six months and the treatment is HAART, which is inferred from records of medications. Simple, unconditional, estimates do not show a difference in sexual activity. But when treatment is instrumented by variation in the eligibility rules for Medicaid across states a positive effect emerges. The validity of these instruments is checked by examining the reduced form association between Medicaid eligibility and sexual activity prior to the introduction of HAART in 1996. There is no association pre-1996 but there is post-1996.

2.3 Natural controls

Families

Auld and Sidhu (2005) present evidence that around a quarter of the association between schooling and health is attributable to variation in cognitive ability and that the causal effect of schooling on health is concentrated among those with low levels of education. Estimates that allow for both schooling and health to be influenced by a common "third factor" diminish the effect of schooling on health except for those individuals with no greater than high-school education. The models are estimated using the 1979 and 2000 US National Longitudinal Survey of Youth (NLSY). The validity of the estimated causal effects relies on the use of parental education as a source of independent variation, using variation in the individual's own education that is associated with their parent's educational achievements.

Siblings who are brought up together share common background characteristics which may be unobserved and also influence the treatments and outcomes of interest to researchers. Using within-sibling variation can control for these factors. Holmlund (2005) shows how variation within biological sisters can be used to assess the long-term consequences of teenage pregnancy for educational outcomes. The siblings approach and standard cross section methods produce similar results so long as heterogeneity within the family is controlled for. The potential for selection bias is that teenage mothers may have family backgrounds that would lead to poorer outcomes irrespective of an early pregnancy. Variation within biological sisters can be used to control for these 'family effects'. However within-sibling variation will not deal with heterogeneity within the family and the study controls for observable pre-motherhood school performance, measured by the grade point average (GPA) from primary school, to try and control for this. Data are taken from a 20 per cent sample of each cohort born in Sweden between 1974 and 1977, with the population register used to identify siblings.

Sibling fixed effects play a role in Currie and Stabile's (2006) study of the impact of Attention Deficit Hyperactivity Disorder (ADHD) on educational outcomes. Within-sibling variation is used to control for omitted variables at the level of the family. Data from the Canadian National Longitudinal Survey of Children and Youth (NLSCY) and the US National Longitudinal Survey of Youth (NLSY) are used. ADHD symptoms are based on parental reports and are recorded in 1994 in the Canadian data and between 1990 and 1994 in the US data. Educational outcomes include the repetition of grades, enrolment in special education, reading and maths tests and delinquency. These are measured in 1998 for Canada and 1998-2000 for the US. The study finds large effects, relative to chronic physical conditions, and for low levels of ADHD symptoms in cases

that would not usually receive treatment. The results for Canada and the US are similar to each other.

Twin Studies

Within-sibling variation can control for common factors relating to family background, upbringing and environment. But siblings are born at different times and they have different genes. Twin studies take the notion of natural controls a step further by removing the genetic variation (at least for monozygotic twins). In the context of research on birth outcomes, using twins means that the siblings share the same pregnancy and are born at the same time. This controls for unobservable characteristics of their mother and her behaviour and environment during the pregnancy. Almond et al. (2005) show that using variation in birth weight between twins leads to lower estimates of the impact of low birth weight (LBW), defined as less than 2500 grams, on short-run outcomes than is typically found in cross section studies. They find heterogeneity in the effects of LBW, suggesting a highly nonlinear relationship. Two identification strategies are adopted. The first exploits variation "within mothers" by comparing outcomes for heavier and lighter infants for all twins born in the US between 1983 and 2000. Using this within-variation should control for all observed and unobserved characteristics of the mother. The second exploits variation "between mothers" in a complementary analysis of maternal smoking and singleton births. The strategy here is to attribute the whole effect of smoking to LBW and compare it with the twins estimates. Data are drawn from two sources: linked birth and infant deaths data from the US National Center for Health Statistics (NCHS), covering the population of US twins, and data from hospital discharge abstracts from the Healthcare Cost and Utilization Project (HCUP) state inpatient database. There are some caveats to bear in mind with this study. Some useful descriptive analysis presented in the paper highlights the inherent differences between twins and singeltons (the latter are more healthy). This raises questions about external validity of analysis based on samples of twins rather than the general population. The study only uses short-run outcomes and may miss long-term consequences (see the studies by Behrman and Rosenzweig (2004) and Black et al. (2007) below). For fraternal twins, genetic differences may mean that changes in birth weight may be associated with changes in unobservables (there is evidence of a negative correlation of birth weight with congenital defects) and the fixed effects approach may overestimate the impact of birth weight. Also the data do not distinguish between monozygotic (identical) and dizygotic (fraternal) twins.

In Behrman and Rosenzweig (2004) variation between monozygotic twins provides a way of identifying the impact of birth weight on long-term outcomes such as measures of adult health, anthropometric measures and adult schooling and earnings. Increased birth weight, as measured on the birth certificate, increases schooling among adults and this effect is underestimated by 50 per cent when cross section variation is used to identify the effect. Data were collected through a survey mailed to monozygotic twins on the Minnesota Twins Registry, the largest birth certificate based registry in the US. The identification strategy assumes that difference in birth weight reflects random differences in nutrition in the womb that are uncorrelated with individual endowments and therefore avoids selection bias. The estimates allow for heterogeneous treatment effects and show an impact on labour market outcomes for low birth weight but not for high. The implications of the US results for World-wide health inequalities are explored at the end of the paper.

Black et al. (2007) use Norwegian registry data and, like Behrman and Rosenzweig (2004), use twins to investigate the impact of low birth weight on long-term socioeconomic outcomes rather than just short-run outcomes. Within-twins fixed effects estimates are shown to be significant and similar to standard least squares estimates for long-run outcomes, such as height, IQ, earnings and education, while the estimates for short-run outcomes are smaller for the twins data, as suggested by Almond et al. (2005). The analysis is made possible by the richness of the data. This uses personal identifiers to link all Norwegian births between 1967-97, as recorded in the birth registry, with other registry data for those aged 16-74 in the period 1986-2002. The register data is augmented with military records and a survey of twins that identifies zygocity. Within-twin variation is used to capture unobservable socioeconomic and genetic factors that may confound the causal effect of birth weight. This means that identification stems from differences in nutrition *in utero* (resulting from different placentas for fraternal twins and different positioning on the placenta for monozygotic). Birth order is included as a control. The robustness of the findings is assessed by separate analyses for mothers who have more than one singelton birth, allowing for mother fixed effects rather than pregnancy fixed effects. To assess the role of zygocity the sample is restricted to same-sex twins. Also the sub-sample where there is survey data on zygocity is used. The findings are robust but reveal interesting evidence that those who participate in twins studies are a self-selected sample. Also it should be borne in mind that selection into the sample of registry data for long-run outcomes may be affected by infant mortality. Finally, there are substantial differences between twins and singletons in terms of factors such as gestation and the age of their mothers and twins usually appear in the lower part of the distribution of birth weights.

Communities

Many studies use variation within groups, communities or geographic areas to control for unobservable factors that are common to all those within the community or locality. For example, Wagstaff (2007) controls for village effects in a study of the impact of health shocks, such as the death of a working-age member of the household, on incomes of urban and rural households in Vietnam based on the Vietnam Living Standards Survey (VLSS). Arcidiacono and Nicholson (2005) find that adding fixed effects for individual medical schools eliminates the positive peer effects that appear to exist when selection bias is not taken into account. The inclusion of school effects means that the impact of peer effects on a student's achievements and on their choice of specialty are identified by variations over time within schools in the ability and preferences of students. The aim is to separate correlated effects from exogenous peer effects. The study relies on data on graduates from US Medical Schools over a relatively short period, 1996-98, so identification may be limited by a lack of variation over time. Currie and Neidell (2005) use variation within Californian zip code areas to identify the impact of air pollution on infant mortality. They find a statistically significant effect even at low levels of air pollution. The impact of this effect is quantified: it is estimated that reduction in pollution in California over the 1990s saved around 1000 infant lives. The study takes data from the California birth cohort files and matches it to EPA data on air quality - specifically measures of carbon monoxide, ozone and particulate matter (PM10) – and information on weather patterns from the National Climatic Data Center. A linear model is used to approximate the discrete hazard function for infant deaths and month, year and zip code fixed effects are included in the model, this relies on variation with cells of

observations defined by month, year and locality. The study complements the natural experiment presented by Chay and Greenstone (2003) that is described above.

2.4 'Anti-tests'

One way to assess the robustness of an identification strategy is to find an anti-test (or placebo test). Anti-tests provide counter evidence by applying a model or identification strategy in a context where no effect should be detected. If an apparent "effect" is found then the validity of the identification strategy must be called into question.

For many years the standard empirical strategy to test for the phenomenon of supplier induced demand (SID) in medical care has been to include a measure of the supply of doctors – usually the physician density, measuring the number of doctors in a locality per head of population – in empirical models of health care utilisation or expenditure. This strategy is plagued by omitted variable bias and identification problems. To assess the robustness of the approach, Dranove and Wehner (1994) apply the physician density strategy using the obstetrician/population ratio and the volume of births as the measure of utilisation. The physician density test shows evidence that the number of births (and hence pregnancies) is "supplier-induced": casting obvious doubt on the reliability of the approach. However failure of the methodology does not imply rejection of SID. For example, Gruber and Owings (1996) find evidence of increased C-section rates in response to fall in fertility in the US between 1970 and 1982: a shift by obstetricians to more lucrative procedures in response to economic pressures.

In their "Addiction to milk" paper Auld and Grootendorst (2004) use non-addictive substances, such as milk, eggs and oranges, to construct an anti-test and demonstrate that evidence for the rational addiction hypothesis based on aggregate data may be spurious. Numerous studies have applied the canonical rational addiction equation of Becker et al. (1994) to substances such as alcohol, cigarettes and cocaine, and claim to have found support for rational addiction. But Auld and Grootendorst (2004) show that these findings are mimicked when the model is applied to Canadian aggregate data for non-addictive substances. Monte Carlo simulations show that spurious evidence is likely when the time-series data exhibit high serial correlation, when prices are poor instruments, when over-identified instrumental variable estimators are used, or when theoretical restrictions are imposed by fixing the implied discount rate in the model.

The idea of an anti-test may provide a useful strategy as part of a robustness/sensitivity analysis. A good example of this is Galiani et al.'s (2005) evaluation of the impact of the privatisation of local water services on child mortality in Argentina. They adopt two strategies for assessing the reliability of their difference-in-differences approach that can both be interpreted as anti- or placebo tests. The first, which is a good practice to adopt in any difference-in-differences analysis, is to estimate a placebo regression: the model of interest is estimated using only data from the pre-treatment period, but including an indicator of those cases that will go on to be treated. If this indicator of hypothetical treatments is significant it is a sign that the treated and controls are not comparable and that the "parallel trends" assumption required for difference-in-differences analysis is not valid. The second strategy adopted by Galiani et al. (2005) is that, as well as measuring deaths from infectious and parasitic diseases they include measures of deaths from causes unrelated to water quality. The fact that they detect a reduction for the former but not for the latter creates confidence in their difference-in-differences identification strategy.

3. Data and measurement issues

3.1 Administrative data or sample surveys

Much of the applied work done by health economists uses social surveys. These are often designed to provide representative random samples of the underlying population. Most often the sampling follows a multi-stage design with clustered and/or stratified sampling (see e.g., Jones et al. (2007b)). Data may be collected by face-to-face interviews or postal, telephone or web-based questionnaires and in health surveys this is often supplemented by clinical tests and measurements. Many surveys are one-off cross sections but increasingly researchers have turned to longitudinal, or panel, surveys which give repeated observations on the units of interest, whether they be individuals, households or organisations. Sample surveys are the mainstay of microeconometric research and some of the more popular datasets are summarised in Table 1.

INSERT TABLE 1 AROUND HERE

In health economics administrative datasets often prove more useful and reliable than social surveys. Administrative datasets include sources such as tax records, reimbursement and claims databases and population registers of births, deaths, cancer cases, HIV/AIDS cases, unemployment etc. (see e.g., Aakvik et al. (2003); Aakvik et al. (2005); Atella et al. (2006); Black et al. (2007); Chalkley and Tilley (2006); Dano (2005); Dranove et al. (2003); Dusheiko et al. (2004); Dusheiko et al. (2006); Dusheiko et al. (2007); Farsi and Ridder (2006); Gravelle et al. (2003); Ho (2002); Lee and Jones (2004); Lee and Jones (2006); Martin et al. (2007); Propper et al. (2002); Propper et al. (2004); Propper et al. (2005); Rice et al. (2000); Seshamani and Gray (2004)). These data are collected primarily for administrative purposes and are made available to researchers for secondary analysis. Some countries allow comprehensive linkage of different sources of administrative data based on personal identification numbers (e.g., Black et al. (2007)). Administrative datasets are typically large, often with millions rather than thousands of observations, and are comprehensive, often providing observations on a complete population rather than a random sample. They tend to be less prone to unit and item non-response than survey data and may give better coverage of hard-to-reach groups of the population and the socially disadvantaged. Also they tend to be less affected by reporting bias but are still vulnerable to data input and coding errors. Given their primary purpose, administrative datasets are not designed by and for researchers. This means they may not contain all of the variables that are of interest to researchers, such as socioeconomic characteristics, and that many different sources may have to be combined to produce a usable dataset. In some cases sources of administrative data may be made combined and made available with researchers in mind. For example, the Oxford Record Linkage Study (ORLS), used by Seshamani and Gray (2004), is a longitudinal dataset that links statistical abstracts for hospital inpatient and day cases to birth and death certificates for people living in the Oxford region of England. It provides 10 million records for over 5 million people between 1963 and 1999.

Dusheiko et al.'s (2004) study of the impact of practice budgets for GPs on hospital waiting times in the English NHS provides an example of the complex and painstaking process that is often required to link administrative data. Information on waiting times was obtained from the Hospital Episode Statistics (HES) for 1997/98 to 2000/01. HES

is an annual database of hospital inpatient activity, including day cases, with more than 10 million records per year. Dusheiko et al. (2004) extracted information on the waiting times for over 5 million finished consultant episodes and linked average waiting times to GP practices. Information on practice populations was obtained from the PCT database at the National Primary Care Research and Development Centre (NPCRDC: www.primary-care-db.org.uk). Practice characteristics, such as the GP's age and sex, qualifications, size of the practice and so on, were obtained from the Prescription Pricing Authority, the Department of Health's Organisational Codes Service and their General Medical Statistics, along with the NPCRDC database. Patient characteristics for each practice were obtained from the 1991 Census and components of the Index of Multiple Deprivation, with these small area data mapped to GP practices. Finally, supply side factors, such as distances to hospitals, were obtained from the Department of Health's AREA project.

Some of the key administrative datasets that have been used in health economics are summarised in Table 1.

Non-response and attrition

Non-response and attrition are a common feature of longitudinal survey data. Nicoletti and Peracchi (2005) list possible reasons for non-response: these include demographic events such as death; movement out of scope of the survey such as institutionalization or emigration; refusal to respond at subsequent waves; absence of the person at the address, along with other types of non-contact. Jones et al. (2006) investigate health-related non-response in the first eleven waves of the British Household Panel Survey (BHPS) and the full eight waves of the European Community Household Panel (ECHP). They explore its consequences for dynamic models of the association between socioeconomic status and self-assessed health (SAH). Descriptive evidence shows that there is health-related non-response in the data, with those in very poor initial health more likely to drop out, and variable addition tests provide evidence of non-response bias in the panel data models of SAH. Nevertheless a comparison of estimates - based on the balanced sample, the unbalanced sample and corrected for non-response using inverse probability weights – shows that, on the whole, there are not substantive differences in the average partial effects of the variables of interest.

Inverse probablity weights are used to attempt to control for attrition: this works by estimating separate probit equations for whether an individual responds or does not respond at each of the waves of the panel. Then the inverse of the predicted probabilities of response from these models are used to weight the contributions to the log likelihood function in the pooled probit models for SAH. The rationale for this approach is that a type of individual who has a low probability of responding represents more individuals in the underlying population and therefore should be given a higher weight. The appropriateness of this approach relies on the assumption that non-response is ignorable conditional on the variables that are included in the models for non-response ("selection on observables"). If this assumption holds then inverse probability estimates give consistent estimates. The findings in Jones et al. (2006) and the earlier work by Contoyannis et al. (2004b) suggest that, while health-related non-response clearly exists, on the whole it does not appear to distort the magnitudes of the estimated dynamics of SAH and the relationship between socioeconomic status and self-assessed health. Similar findings have been reported concerning the limited influence of

non-response bias in models of income dynamics and various labour market outcomes and on measures of social exclusion such as poverty rates and income inequality indices.

3.2 Health outcomes

Self-reported data

Self-assessed health (SAH) is often included in general social surveys. For example, in the British Household Panel Survey (BHPS), SAH is an ordered categorical variable based on the question "Please think back over the last 12 months about how your health has been. Compared to people of your own age, would you say that your health has on the whole been excellent/good/fair/poor/very poor?". The validity of self-reported measures of health has caused considerable debate. As a self-reported subjective measure of health, SAH may be prone to measurement error. General evidence of non-random measurement error in self-reported health is reviewed in Currie and Madrian (1999) and Lindeboom (2006).

Self-assessed health is not the only source of concern with self-reported data. Baker et al. (2004) use careful record linkage to check for flaws in self-reported data on specific chronic conditions. Survey data from the Canadian National Population Health Survey for 1996-7 are linked to ICD-9 codes for Ontario residents from administrative data on utilisation of services for the Ontario Health Insurance Plan (OHIP) for the survey year and the five previous years. Linear probability models are used to analyse the probabilities of false negatives and false positives in the self-reported data. This shows that reporting errors are associated with individual characteristics. The data reveal a large number of false negatives, although the probability declines for those with more recorded medical treatments, suggesting that this reflects undiagnosed conditions and lack of information among respondents. The number of false positives is much smaller but there is some evidence of "justification bias": those not in work are more likely to report false positives for conditions such as hypertension, ulcers and bronchitis.

A more favourable view of self-reported measures emerges in the work of Benitez-Silva et al. (2004). They take the, relatively small, sub-sample of respondents to the first three waves of the US Health and Retirement Survey (HRS) who had applied for disability benefit from the Social Security Administration (SSA) and compare their self-reported disability to the outcome of the SSA decision. In this case the SSA decision to award benefits is used as an objective indicator to assess the reliability of the self-reported data on limitations that prevent work. Conditional moment tests for whether self-reported disability is an unbiased indicator of the SSA decision suggest that a large fraction of this population report their health accurately. Unlike Baker et al. (2004) this study relies on information collected within the HRS rather than matching the survey data with administrative records. McGarry (2004) adopts another strategy to get around the problem of 'justification bias'. Rather than using data on actual retirement she uses information from the US Health and Retirement Survey (HRS) on the subjective expected probability of retirement by age 62, which is collected while people are still in work. Using this measure she finds strong effects of health on expected retirement age.

It is sometimes argued that the mapping of health into SAH categories may vary with respondent characteristics. This source of measurement error has been termed "state-dependent reporting bias" (Kerkhofs and Lindeboom (1995), "scale of reference bias"

(Groot (2000)) and "response category cut-point shift" (Sadana et al. (2000); Murray et al. (2001)). Regression analysis of SAH is often done by specifying an ordered probability model, such as the ordered probit or logit. Then the symptoms of measurement error can be captured by making the cut-points dependent on some or all of the exogenous variables used in the model and estimating a generalised ordered model. This requires strong *a priori* restrictions on which variables affect health and which affect reporting in order to separately identify the influence of variables on latent health and on measurement error. Attempts to surmount this fundamental identification problem include modelling the reporting bias based on more "objective" indicators of true health (Kerkhofs and Lindeboom (1995), Lindeboom and Van Doorslaer (2004)) and the use of "vignettes" to fix the scale (Das and Hammer (2005); Murray et al. (2001)). Lindeboom and van Doorslaer (2004) analyse SAH in the Canadian National Population Health Survey and use the McMaster Health Utility Index (HUI-3) as their objective measure of health. They find evidence of reporting bias with respect to age and gender, but not for income, education or linguistic group. A similar identification strategy to Lindeboom and van Doorslaer (2004), that relies on objective measures capturing all of the genuine variation in health, is adopted by Etilé and Milcent (2006) who estimate generalised ordered probit models with a 4-category measure of self-assessed health. They construct synthetic measures of objective health from a latent class analysis of a set of self-reported indicators such as ADLs and BMI. The latent class analysis is used to condense the sample into six classes and indicators for these classes are included in the ordered probit model. This works by constructing classes, such that the underlying health indicators are independent of each other, conditional on class membership. Latent class models are presented as an alternative to the grade of membership approach that has been used in some earlier work (Lindeboom et al. (2002)). Estimates on a sample of 2956 individuals aged under 65 from the French Enquête Permanente sur les Conditions de Vie des Menages (EPCV) survey show evidence of reporting bias and, unlike Lindeboom and Van Doorslaer (2004), there is evidence that this is related to income. Etilé and Milcent (2006) findings suggest a concave relationship between health and income in terms of health production and convexity with respect to reporting, with over-optimism among the rich and over-pessimism among the poor. They conclude that the problems of reporting bias can be minimised by collapsing the 4-pont scale into a binary measure of poor health.

Jurges (2007) focuses on cross-country differences in reporting of self-assessed health as measured for those age over 50 in the ten countries covered by the first wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). Generalised ordered probit models are used to regress SAH on a set of objective measures, such as grip strength, walking speed and body mass index (BMI), to get a set of disability weights. The average thresholds across the SHARE countries are then used to reclassify the reported data, assuming that the disability weights are constant across countries. The variation in SAH is decomposed into the component that is explained by the objective measures and the component attributed to reporting bias. The findings suggest that those in the Danish and Swedish samples over-rate their health, while those in Germany under-rate their health. For Austria and Greece there is little bias.

Anthropometric measures

Anthropometric measures have long played a role in studies of developing countries, especially those focused on child health issues. With the growing problem of adult and

childhood obesity in more affluent nations they are increasingly being used in that context as well. Typical anthropometric measures are height and weight, which may be self-reported or measured by a professional; infant length for children aged under 2; demi-span, which is based on the length of an out-stretched arm and is used among older populations who may have difficulties standing straight; the waist-to-hip ratio; and the body mass index (BMI). BMI is the most commonly used indicator of obesity. It is calculated as weight in kilograms divided by height in meters squared, with a BMI of 30 or greater indicating obesity and 25-30 indicating over-weight. The confounding effect of levels of muscle development means that measures of body fat are sometimes used instead of BMI. The height-for-age z-score standardises a child's measured height using the median (or mean) and standard deviation for children of the same age and sex from a reference population, such as the US National Center for Health Statistics reference population of well-nourished American children (e.g., Duflo (2000)). Height is also compared to a standard distribution to construct measures of stunting (e.g. Gertler (2004)). For example, Chen and Zhou (2007) use height to measure the long-term health consequences of childhood exposure to the 1959-61 famine in China, which is estimated to have caused 15-30 million excess deaths. They adopt a difference-in-differences approach that exploits regional differences in exposure to the famine.

Anthropometric measures may also play a role as biomarkers (discussed in more detail below). For example height and weight can be used as predictors of mortality, stroke and cardio-vascular disease and the waist-to-hip ratio is a predictor for hypertension, late-onset diabetes, cardio-vascular disease, stroke and some forms of cancer.

Biomarkers

Biological markers, or biomarkers, are likely to play an increasing role in future research in health economics as they are incorporated into an increasing range of datasets, including longitudinal datasets such as the US Health and Retirement Survey (HRS), the English Longitudinal Survey of Ageing (ELSA) and the planned UK Longitudinal Household Study. This trend is likely to be enhanced by the availability of DNA information and genetic screening, which provide greater potential to control for individual heterogeneity. Biomarkers are biological or physiological measures that indicate the presence of a disease or the propensity to develop a disease. They can be used to identify risk factors and as objective measures of health that avoid contamination by reporting bias (see e.g., Adda and Cornaglia (2006); Banks et al. (2006); Currie et al. (2007)).

Biomarkers for cardio-vascular disease include elevated blood pressure and variability in the heart rate. Metabolic biomarkers include serum HDL and total cholesterol and triglycerides which are predictors of heart disease; fibrinogen, which is linked to blood clotting and the risk of heart disease; and glycated haemoglobin, which is a proxy indicator for diabetes. Biomarkers linked to the immune system include interleukin-6 (IL-6), which is a predictor of Altzheimer's disease, arthritis, diabetes and osteoperosis; C-reactive protein (CRP), which indicates lupus, pneumonia, rheumatoid arthritis, rheumatic fever and tuberculosis; ferritin and haemoglobin, which indicate iron deficiency; serum retinol, which indicates vitamin A deficiency. Biomarkers linked to hormonal indicators of stress (HPA axis) include cortisol, adrenocotricotrpic hormone (ACTH) and dehydroepiandrosterone-sulphate (DHEA-S). Biomarkers linked to the sympathetic nervous system include norepiphrenine, which is associated with longevity,

and epiphrenine (adrenaline), which is linked to cognitive decline and longevity. Biomarkers may also be used as objective indicators of physical functioning and health limitations; these include lung function tests such as forced expiratory volume (FEV) and forced vital capacity (FVC), as measured by a spirometer, and grip strength, measured by a gripometer.

Datasets which contain biomarkers and that have been used in economic research include the English Longitudinal Survey of Ageing (ELSA), UK Health and Lifestyle Survey (HALS), the US Health and Retirement Survey (HRS), the Health Survey of England (HSE), the UK National Child Development Study (NCDS) and the more recent cohort studies, the US National Health and Nutrition Examination Surveys (NHANES), the Survey of Health, Aging and Retirement in Europe (SHARE), and the Whitehall Study of English civil servants. To take the ELSA as an example: participants in the study are visited by a registered nurse who takes measurements of blood pressure; lung function; height, weight and the waist-to-hip ratio; grip strength, as a measure of upper body strength; a measure of lower body strength, based on standing up from a chair without using the arms; a saliva sample, that is used to measure cortisol which is a marker for stress; and a blood sample, which is used to test total cholesterol, HDL cholesterol, fibrinogen, CRP, ferritin, glycated haemoglobin and haemoglobin. The blood samples from ELSA and from wave 7 of the NCDS will allow DNA to be extracted. DNA can also be collected using mouth or cheek swabs (as in the US AddHealth Survey).

Banks et al. (2006) study the socioeconomic gradient in health in the UK and the US and they compare both self-reported outcomes and objective outcomes based on biomarkers. To do this they use the ELSA data for the UK and the HRS and NHANES for the US. The self-reported measures include the general question on self-assessed health (SAH) as well as self-reported indicators of chronic conditions such as diabetes, hypertension and cancer. The biomarkers are glycosated haemoglobin levels above 6.5%, as a marker for diabetes; systolic blood pressure over 140mm Hg and diastolic blood pressure over 90mm Hg, as a measure of hypertension; CRP greater than 3mg/L, as a marker of high risk of arteriosclerosis; fibrinogen over 400 mg/dl, as a marker for cardiovascular disease and HDL cholesterol over 40 mg/dl, as an indicator of reduced risk of coronary heart disease. Banks et al. (2006) find that, on average, respondents in the US reported better self-assessed health but the opposite holds true for the biomarkers. Their results show a strong socioeconomic gradient in self-assessed health, self-reported diseases and in the biomarkers. The gradient appears strongest for the biomarkers. Comparing the self-reported data with the biomarkers allows a measure of the socioeconomic gradient in undiagnosed cases. A gradient is apparent for diabetes but not for hypertension.

A novel feature of Adda and Cornaglia (2006) is the use of biomarkers, in this case cotinine, within an economic study of smoking. Cotinine is a metabolite of nicotine and can be used as a biomarker for levels of tobacco consumption that is not contaminated by problems of measurement error, such as recall bias and deliberate deception, that may affect self-reported consumption. The study shows that smokers engage in compensatory behaviour, increasing their intensity of smoking and off-setting the impact of tobacco tax increases. Data on cotinine is collected from saliva samples as part of the repeated cross section data in the US National Health and Nutrition Examination Surveys (NHANES) for 1999-2000. Evidence based on the biomarker is contrasted with self-reported consumption. Cotinine has another advantage in studies of smoking and health as it provides a way of measuring passive smoking, especially among children.

3.3 Modelling costs and expenditure

Individual-level data on medical expenditures and costs of treatment are typically distinguished by a spike at zero, if there are non-users in the data, and a strongly skewed distribution with heavy tails. These kind of data are most often used in two areas of application: risk adjustment and cost-effectiveness analysis. In risk adjustment the emphasis is on predicting the treatment costs for particular types of patient, often with very large datasets. Cost-effectiveness analyses tend to work with smaller datasets and the scope for parametric modelling may be more limited (Briggs et al. (2005)). In the context of clinical trials attention has focused on methods to deal with censoring of cost data due to limited follow-up (e.g., Baser et al. (2006); Raikou and McGuire (2004) Raikou and McGuire (2006)).

The presence of a substantial proportion of zeros in the data has typically been handled by using a two-part model: which distinguishes between a binary indicator used to model the probability of any costs and a conditional regression model for the positive costs. OLS applied to the level of costs (y) can perform poorly, due to the high degree of skewness and excess kurtosis, and the positive observations are often transformed prior to estimation. The most common transformation is the logarithm of y although the square root is sometimes used as well. As the policy interest typically focuses on predicting costs on the original scale, the regression results have to be retransformed back to that scale. This weakens the case for working with transformed data and, in particular, problems arise with the retransformation if there is heteroskedasticity in the data on the transformed scale (Manning (1998); Manning and Mullahy (2001); Mullahy (1998)). Ai and Norton (2000) provide standard errors for the retransformed estimates when there is heteroskedasicity.

More recently attention has shifted to other estimators. Basu et al. (2004) compare log-transformed models to the Cox proportional hazard model. Gilleskie and Mroz (2004) propose a flexible approach that divides the data into discrete intervals then applies discrete hazard models, implemented as sequential logit models. Conway and Deb (2005) use a finite mixture model. Cooper et al. (2007) use hierarchical regressions implemented using Bayesian MCMC. But the dominant approach in the recent literature has been the use of generalised linear models (GLM) (e.g., Buntin and Zaslavsky (2004); Manning and Mullahy (2001); Manning et al. (2005); Manning (2006)). The GLM specifies a link function for the relationship between the conditional mean, $\mu=E(y|x)$, and a linear function of the covariates and specifies the form of the conditional variance, $V(y|x)$, usually assuming that it can be specified as a simple function of the mean. The models are estimated using a quasi-likelihood approach derived from the quasi-score or "estimating equations". The most popular specification of the GLM for costs has been the log-link with a gamma error (Blough et al. (1999); Manning and Mullahy (2001); Manning et al. (2005)). Cantoni and Ronchetti (2006) propose a robust variant of GLM that is less sensitive to outliers. In response to the problem of selecting the appropriate link and variance functions Basu and Rathouz (2005) suggest a flexible semiparametric extension of the GLM model. Their model incorporates a Box-Cox transformation into the link function which includes the log-link as a special case along with other power functions of y. The model, which is labelled the extended estimating equations (EEE) approach, also allows for flexible specifications of the variance using the power variance and quadratic variance families to nest common distributions such as the Poisson, gamma, inverse Gaussian and negative binomial. Basu et al. (2006) apply the EEE method to claims data on the incremental costs associated with heart failure.

4. Methods for dealing with unobserved heterogeneity and dependence

4.1 Deviations and conditional estimates

Consider a linear panel data regression model with repeated measurements ($t=1,...., T_i$) for a sample of n individuals ($i=1,.....,n$):

$$y_{it} = x_{it}'\beta + u_i + \varepsilon_{it} \tag{16}$$

Correlation between the unobservable individual effects (u) and the regressors (x) will lead to an omitted variable bias and inconsistent estimates of the βs. The individual effects can be swept from the equation by transforming variables into deviations from their within-group means or by using orthogonal deviations, based on the mean of the future values of the variables. Applying least squares to the mean deviations gives the covariance or within-groups estimator of β. Similarly, the model could be estimated in first differences to eliminate the individual effects. Identification of β rests on there being sufficient variation over time and the estimators may perform poorly when there is insufficient variation.

Many of the outcomes used in health economics are binary or ordered categorical measures, such as self-assessed health. Fixed effects panel data methods, that allow for a correlation between the individual effect and the regressors of the model, are not, however, readily available for categorical data due to the incidental parameter problem. For binary data the problem can be surmounted by using the conditional fixed effects logit, which uses a sufficient statistic to eliminate the individual effect from the log-likelihood function (Chamberlain (1980)). In the case of the logistic regression the within-individual sum of y_{it} is a sufficient statistic and conditional ML estimates are consistent. Although the conditional logit provides consistent parameter estimates, the approach has practical drawbacks for the researcher. First, by only using observations that have within-individual variation in the outcome and in the regressors, the method often leads to a substantial reduction in sample size. Second, it is hard to calculate partial effects of a variable of interest due to the inherent lack of information on the distribution of the individual heterogeneity, which is conditioned out of the model.

The conditional logit can be applied to ordered data by choosing a particular threshold value and collapsing the data into a binary measure. A recent extension of Chamberlain's model, the conditional ordered fixed effects logit, proposed by Ferrer-i-Carbonell and Frijters (2004) and applied to data on self-assessed health by Frijters et al. (2005), suggests a method to reduce the drastic loss in the number of observations by identifying individual-specific threshold values to collapse the ordered dependent variable into a binary format. Das and van Soest (1999) combine adjacent categories so that the dependent variable is summarized as a binary variable, and then use conditional logits. They repeat this for all the possible combinations of adjacent categories to get a set of estimates of the parameters of interest. They then define a linear combination of these estimates, with the optimal weighting matrix used to compute the final estimate obtained from a minimum distance approach. Ferrer-i-Carbonell and Frijters (2004) also propose an estimator that collapses the ordered variable into a binary format, but they use an individual specific threshold value. To find this individual threshold, the authors

maximize a weighted sum of log-likelihood functions, similar to Das and van Soest (1999), subject to the constraint that the sum of squared weights across all possible threshold values across all individuals must be equal to the number of individuals in the sample. The threshold is selected for which the analytical expected Hessian is minimized. However, this formulation of the estimator is highly computation intensive, a fact which makes its wider application less attractive. In a simplification of this estimator, one can simply use the within-individual means as cut-off criterion.

Dynamic models

Even for linear models the within-groups estimator breaks-down in dynamic models such as:

$$y_{it} = \alpha y_{it-1} + u_i + \varepsilon_{it} \tag{17}$$

This is because the group mean is a function of ε_{it} and ε_{it-1}. An alternative is to use the differenced equation:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta \varepsilon_{it} \tag{18}$$

in which case both y_{it-2} and Δy_{it-2} are valid instruments for Δy_{it-1} as long as the error term (ε_{it}) does not exhibit autocorrelation. Arellano and Bond (1991) proposed generalised method of moments (GMM) estimators for dynamic panel data models: linear models that can include leads and lags of the dependent variable as well as a fixed effect. Instruments are created within the model by first taking differences of the equation to sweep out the individual effect and then using lagged levels or differences of the regressors as instruments.

Bover and Arellano (1997) extend the use of GMM to dynamic specifications for categorical and limited dependent variable models, where it is not possible to take first differences or orthogonal deviations as the latent variable y^* is unobserved. One of the advantages of using panel data is the possibility to account for the correlation amongst the effects and the explanatory variables. To allow for this correlation Chamberlain (1984) suggested using a random effects approach and specifying a distribution for the individual effects conditional on the values of the explanatory variables at each wave of the panel. This specification may contain polynomial terms and interactions in the *x's* as well. Combining this with assumptions about the conditional expectation of the initial and final values of the latent variable allows the dynamic model to be solved out to give linear reduced forms for the latent variables at each of the wave of the panel. Estimates of the reduced forms will be sensitive to assumptions about the distribution of the error terms the linearity of the expected value and the conditional mean independence assumption. However, these hypotheses can be checked by specification tests at the level of the reduced form, which is easier to do than testing the dynamic specification. At the second stage, on the basis of the reduced form coefficients, the parameters of the underlying dynamic structural model can be derived using various estimators. The simplest is to apply the within-groups transformation to the dynamic model after replacing the latent variables by their predicted counterparts (Bover and Arellano (1997)). This two-step within-groups procedure is simple to apply but provides inefficient parameter estimates. Chamberlain (1984) proposed a fully efficient minimum distance

(MD) estimator. Instead of using Chamberlain's approach, Bover and Arellano (1997) propose a three-step within-groups GMM which also facilitates tests of the over-identifying restrictions.

4.2 Numerical integration and classical simulation-based inference

In panel data specifications unobserved heterogeneity is often modelled as a random effect and "integrated out" of the log-likelihood function. Monte Carlo simulation techniques can be used to deal with the computational intractability of nonlinear models, such as panel and multinomial probit models. Popular methods of simulation-based inference include classical Maximum Simulated Likelihood (MSL) estimation, and Bayesian Markov Chain Monte Carlo (MCMC) estimation. This section introduces the classical approach (for a review of the methods and applications to health economics see Contoyannis et al. (2004a)).

Numerical integration by quadrature works well with low dimensions but computational problems arise with higher dimensions. Instead Monte Carlo simulation can be used to approximate integrals that are numerically intractable. This includes numerous models derived from the multivariate normal distribution. Simulation approaches use pseudo-random draws of the evaluation points and computational cost rises less rapidly than with quadrature.

The principle behind simulation-based estimation is to replace a population value by a sample analogue. This means that laws of large numbers and central limit theorems can be used to derive the statistical properties of the estimators. The basic problem is to evaluate an integral of the form:

$$\int [h(u)]dF(u) = E_u [h(u)] \tag{19}$$

where $h(u)$ is a nonlinear function of the random vector u which has a multivariate density $f(u)$. This kind of expression arises in panel data models with random effects specifications and with autocorrelated errors and in multiple equation models with correlated unobservables. The integral can be approximated using draws from $f(u)$, u_r, $r=1,\ldots,R$ such that:

$$\int [h(u)]dF(u) \approx \frac{1}{R}\sum_{r=1}^{R} [h(u_r)] \tag{20}$$

Maximum Simulated Likelihood (MSL) is a simple extension of classical maximum likelihood estimation (MLE) and is useful in many cases where the log-likelihood function involves high dimensional integrals. The idea is to replace individual contributions to the sample likelihood function with an average over R random draws:

$$l_i = \frac{1}{R}\sum_{r=1}^{R} [l(u_{ir})] \tag{21}$$

where $l(u_{ir})$ is an unbiased simulator of L_i. The MSL estimates are the parameter values that maximize,

$$Lnl = \sum_{i=1}^{n} \left[Lnl_i \right] \tag{22}$$

For likelihoods derived from the multivariate normal the GHK simulator is often used. In practice, Halton sequences or antithetics can be used to reduce the variance of the simulator (see Contoyannis et al. (2004a) for details).

4.3 Bayesian MCMC

In Bayesian analysis a prior density of the parameters of interest, say $\pi(\theta)$, is updated using information from sample data. Given a specified sample likelihood for the observed data, $l(y \mid \theta)$, the posterior density of θ is given by Bayes' theorem,

$$\pi(\theta \mid y) = \frac{\pi(\theta) l(y \mid \theta)}{\pi(y)} \tag{23}$$

where,

$$\pi(y) = \int \pi(\theta) l(y \mid \theta) d\theta \tag{24}$$

The scaling factor $\pi(y)$ is known as the predictive likelihood and is used to compare models. It determines the probability that the specified model is correct. The posterior density $\pi(\theta \mid y)$ reflects updated beliefs about the parameters. Given the posterior distribution, a 95% credible interval can be constructed that contains the true parameter with probability equal to 95%. Point estimates for the parameters can be computed using the posterior mean,

$$E(\theta \mid y) = \int \theta \pi(\theta \mid y) d\theta \tag{25}$$

Bayesian estimates can be difficult to compute directly. For instance, the posterior mean is an integral with dimension equal to the number of parameters in the model. In order to overcome the difficulties in obtaining the characteristics of the posterior density, Markov Chain Monte Carlo (MCMC) simulation methods are often used. The methods provide a sample from the posterior distribution and posterior moments and credible intervals are obtained from this sample (see Contoyannis et al. (2004a) for details).

Bayesian MCMC simulation is built on the Gibbs sampling algorithm. To implement Gibbs sampling the vector of parameters is subdivided into groups. For example, with two groups let $\theta = (\theta_1, \theta_2)$. Then, a draw from the joint distribution $\pi(\theta_1, \theta_2)$ can be obtained in two steps: first, draw θ_1 from the marginal distribution $\pi(\theta_1)$; then draw θ_2 from the conditional distribution $\pi(\theta_2 \mid \theta_1)$. However, in many situations it is possible to sample from the conditional distribution but it is not obvious how to sample from the marginal. The Gibbs sampling algorithm solves this problem by sampling iteratively from the full set of conditional distributions. Even though the Gibbs sampling algorithm never actually draws from the marginal, after a sufficiently large number of iterations the draws can be regarded as a sample from the joint distribution. There are situations in which it is not possible to sample from a conditional density, and hence the Gibbs sampling cannot be applied directly. In these situations, Gibbs sampling can be combined with a so called

Metropolis step as part of a Metropolis-Hastings algorithm. In the Metropolis step, values for the parameters are drawn from an arbitrary density, and accepted or rejected with some probability. An attraction of MCMC is that latent or missing data can be treated as parameters to be estimated. Although this data augmentation method introduces many more parameters into the model, the conditional densities often belong to well-known families and there are simple methods to sample from them. This makes the use of MCMC especially convenient in nonlinear models, where the latent variables (y*) can be treated as parameters to be estimated. Once the y*s have been simulated the estimation step involves the estimation of normal-linear models for y*.

4.4 Finite mixture models

Latent class models

Recently the latent class framework has been used in models for health care utilisation with individual data. Deb and Trivedi (2002) note that this framework "provides a natural representation of the individuals in a finite number of latent classes, that can be regarded as types or groups"[2]. The segmentation can represent individual unobserved characteristics such as unmeasured health status. The latent class (or finite mixture) framework offers a representation of heterogeneity, where individuals are drawn from a finite number of latent classes. For example, Conway and Deb (2005) show that allowing for heterogeneity between "normal" and "complicated" pregnancies leads to evidence that early prenatal care is effective: on average, bringing the onset of prenatal care forward by one week increases birth weights by 30-60g in normal pregnancies. Using a finite mixture model to capture the bimodality of the distribution of birth weights counteracts evidence from the standard 2SLS approach that the effects of prenatal care are weak or non-existent. Estimates of the mixture model use observational data from the 1988 US National Maternal and Infant Health Survey and the empirical findings are augmented by simulation results that show that the conventional findings could be attributable to the existence of a relatively small proportion (10-15%) of "complicated" pregnancies in the population.

To specify a finite mixture model consider a vector of outcomes y_i that are observed for individual i: these may be repeated observations in a panel data model or related outcomes in a multiple equation model and they are linked by common unobservable heterogeneity. Then assume that each individual belongs to one of a set of latent classes j=1,…,C, and that individuals are heterogeneous across classes. Conditional on the observed covariates, there is homogeneity within a given class j. Given the class that individual i belongs to, the outcomes have a joint density $f_j(y_i|\ x_i;\ \theta_j)$ where the θ_j are vectors of parameters that are specific to each class. The probability of belonging to class j is π_{ij}, where $0<\pi_{ij}<1$ and $\Sigma^{C}_{j=1}\pi_{ij}=1$. Unconditionally on the latent class the individual belongs to, the joint density of y_i is given by:

$$f\left(y_i \mid x_i;\pi_{i1},...,\pi_{iC};\theta_1,...,\theta_C\right)=\sum_{j=1}^{C}\pi_{ij}\, f_j\left(y_i \mid x_i;\theta_j\right) \tag{26}$$

[2] In health economics latent class models (LCM) have typically been applied in the context of nonlinear regression models, to allow for the role of unobserved heterogeneity in the relationship between an observed outcome and a set of regressors. In the statistics literature LCMs are more commonly applied in the context of latent structural variables and a set of observed indicators, such that the indicators are orthogonal conditional on class membership.

The discrete distribution of the heterogeneity has C mass points and the πs need to be estimated along with the θ_js.

In many empirical applications of finite mixture models the class membership probabilities are treated as fixed parameters $\pi_{ij}=\pi_j$, j=1,… ,C (e.g., Deb and Trivedi (1997); Deb and Holmes (2000); Deb (2001); Deb and Trivedi (2002); Jimenez-Martin et al. (2002); Atella et al. (2004); Bago d'Uva (2006)). A more general approach is to parameterise the heterogeneity as a function of individual characteristics. To implement this approach in the case of the latent class model, class membership can be modelled as a multinomial logit (as in, for example, Clark et al. (2005); Etilé (2006)):

$$\pi_{ij} = \frac{\exp\left(z'_i \gamma_j\right)}{\sum_{k=1}^{C} \exp\left(z'_i \gamma_k\right)}, \quad j = 1,...,C, \tag{27}$$

with the normalisation $\gamma_C=0$. This approach uncovers the determinants of class membership. In a panel data context, this parameterisation provides a way to account for the possibility that the observed regressors may be correlated with the individual heterogeneity. Let $z_i = \bar{x}_i$ be the average over the observed panel of the observations on the covariates. This is in line with what has been done in recent studies to allow for the correlation between covariates and random effects, following the suggestion of Mundlak (1978) and Chamberlain (1984). The vectors of parameters θ_1, …, θ_C, γ_1, …, γ_{C-1} are estimated jointly by maximum likelihood.

After estimating the model, it is possible to calculate the posterior probability that each individual belongs to a given class. The posterior probability of membership of class j depends on the relative contribution of that class to the individual's likelihood function. This is given by:

$$P[i \in j] = \frac{\pi_{ij} f_j\left(y_i \mid x_i; \theta_j\right)}{\sum_{k=1}^{C} \pi_{ik} f_k\left(y_i \mid x_i; \theta_k\right)} \tag{28}$$

Each individual can then be assigned to the class that has the highest posterior probability for them.

Finite density estimators and discrete factor models

In latent class models class membership has a discrete distribution with a fixed number of mass points. The models are very flexible in that all of the parameters can be allowed to vary across classes. A special case of this general model assumes the slope coefficients are fixed across classes and that only the intercepts vary. This case is widely used to model unobserved heterogeneity, without imposing parametric assumptions on the distribution of the heterogeneity. The specification has a dual interpretation: the population may truly fall into a discrete set of classes or types or, alternatively, the mass points can be viewed as an approximation of some underlying continuous distribution – the finite density estimator. The finite density estimator was introduced into the econometrics literature by Heckman and Singer (1984) in the context of hazard models (see e.g., Van Ours (2004); Van Ours (2006)). More recently the finite density estimator

has been widely used in multiple equation models where a common factor structure is assumed, as in equation (9) above. This is often called the discrete factor model (DFM).

Since the discrete factor model includes an intercept for each equation, the location of the distribution of the common factor η is arbitrary; also the scale of η is arbitrary and undetermined (Mroz (1999)). Therefore, identification of the DFM requires some normalisations. The existing literature on the DFM offers a range of equivalent strategies to identify the additional parameters of the discrete distribution by fixing the scale and the location of the distribution. If both are fixed, one of the factor loadings is set to 1 and either one of the η_j is set to 0 (see Mroz (1999)) or the mean of the discrete distribution is restricted to be 0, so that one of the η_j can be expressed as a function of the others (Kan et al. (2003)). If only the location is fixed, the first and the last mass points are set to 0 and 1 (this strategy is used by Mroz (1999) when $C > 2$). Other applications also impose that the remaining mass points follow a logistic distribution such that $\eta_k \in (0; 1)$ (see Mello et al. (2002); Picone et al. (2003b)). The π_k can be parameterised using various distributions such as the logistic, normal or the sine function, such that each π_k is between 0 and 1 and they sum to 1.

4.5 Copulas

The presence of common unobservables leads to multiple equation models and the need to specify multivariate distributions. But the menu of parametric forms available for bivariate and, more generally, multivariate distributions is limited. In many applications multivariate normality may be unappealing: for example with heavily skewed and long-tailed data on costs of care or on QALYs (Quinn (2005)) or for rare events. Copulas provide an alternative and are a method of constructing multivariate distributions from univariate marginal distributions (see Trivedi and Zimmer (2005)). A copula is a function that can be interpreted as a joint probability whose arguments are the univariate CDFs of the marginal distributions. The fact that the CDF is used means that the marginal distributions are fixed and invariant to transformations of the random variable. The functional form selected for the copula uniquely determines the form of the dependence, independently of the functional forms of the marginal distributions. The attractions of copulas are that they are flexible – they can mix together marginal distributions of different types, whether they be continuous, integer valued counts or categorical; they allow for richer concepts of dependence than the standard linear measure, including measures of tail dependence; they are computationally tractable and avoid the need for numerical integration or simulation.

A key result in the theory of copulas, Sklar's theorem, shows that all multivariate distributions can be represented by a copula. So in the bivariate case, if two random variables have a joint distribution $F(x_1, x_2)$ and marginal distributions $F_1(x_1)$ and $F_2(x_2)$ then Sklar's theorem establishes that there exits a copula C such that:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)) \tag{29}$$

In practice the unique copula that characterises the true joint distribution is unknown. So particular functional forms have to be selected and compared in terms of their goodness of fit. There is a long list of copulas to choose from. Common choices include the Frank copula and the Farlie-Gumbel-Morgenstern (FGM) copula which is a first-order

approximation of the Frank copula and is tractable to use in applied work (Smith (2003); Prieger (2002); Zimmer and Trivedi (2006)). The bivariate form of the Frank copula is:

$$C_\theta(u,v) = -\theta^{-1}\log\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right) \tag{30}$$

The bivariate form of the FGM copula is:

$$C_\theta(u,v) = uv(1+\theta(1-u)(1-v)) \tag{31}$$

Other common choices include elliptical copulas such as the Gaussian and Student's t and the Clayton copulas. The Gaussian copula is an example of a copula derived by the method of inversion and takes the form:

$$C_\theta(u,v) = \Phi_2(\Phi_1^{-1}(u),\Phi_1^{-1}(v)) \tag{32}$$

where Φ_2 and Φ_1 denote bivariate and univariate standard normal distribution functions.

5. Models for longitudinal data

5.1 Applications of linear models

Models for longitudinal and spatial panels

In Lindeboom et al. (2002) models for cognitive status and emotional well-being are estimated using linear fixed effects specifications, estimated in first differences. The data are taken from three waves of the Dutch Longitudinal Aging Study Amsterdam (LASA) and exhibit considerable attrition. Lindeboom et al. (2002) note that the fixed effects specification is robust to selection associated with time invariant unobservables, but they also include indicators of patterns of response in their model. They find a large impact of life events, such as bereavement, on mental health among the elderly. Carey (2000) estimates the impact of length of stay (LOS) on total hospital costs using a panel of 2792 US hospitals for the period 1987-92. The individual effect in these models captures factors such as the quality of care provided in each hospital, which is likely to be correlated with both costs and LOS. To allow for these correlated effects she uses the Chamberlain minimum distance estimator and finds that the elasticity of total costs with respect to LOS is low.

The rapid growth in dental care expenditure in Taiwan, after the inauguration of national health insurance (NHI) in 1995, lead the government to reform the payment system and introduce global budgeting for outpatient dental care in July 1998. In response to the introduction of global budgets, dentists might alter their supply behaviour, changing the number of visits, the amount of expenditure, and the type of services provided. Lee and Jones (2004) develop two-way fixed effects models to estimate these effects using panel data constructed from outpatient dental care expenditures claims from the Taiwanese National Health Insurance system. The availability of a long panel, with up to 48 monthly observations, allows them to estimate a policy effect for each dentist in the panel, using within-dentist variation and effectively treating each dentist as their own control group.

The individual effects are an important component in the panel data model to investigate dentists' responses to the introduction of global budgeting. The magnitude of the dentist effects measures individual heterogeneity in activities that could not be captured by observable factors in the regression. This allows them to estimate individual-specific responses to the payment reform as well as calculating the average policy effect. They model the individual response to the policy changes by estimating the differences of individual fixed effects between two separate models: pre- and post- global budget. This policy effect can be interpreted as, holding other observable variables constant, the extent to which each individual dentist's activity changed after the introduction of global budgeting. They use OLS to analyze the factors influencing variation in the policy effects across different dentists: which include information on the dentist's demographic characteristics, such as age and gender; the type and ownership of affiliated medical institutions (public or private and hospital or clinic), and exogenous environmental characteristics, such as the dentist-population ratio and annual household income; and whether the dentist practices in a deprived area. The overall effect of global budgets is to constrain costs but there is evidence of a change in the mix of services. Male and younger dentists have higher policy effects than female and older dentists. Global budgets favour dentists in deprived areas and there is some evidence of increases in the expenditure per visit and the volume of composite resin fillings.

Applications of linear models are not confined to longitudinal datasets. Moscone et al. (2007) take a spatial panel approach to data on mental health expenditure by local authorities in England. They compare specifications that allow for a spatial autoregressive process in the error term; a random effects model with spatially lagged dependent variables; and a random effects model with spatial autocorrelation. The random effects model with lagged dependent variables proves to be the preferred specification.

Dynamic panel data models: GMM estimators

Brown et al. (2005) apply dynamic panel data models to assess the impact of HMOs on the supply of doctors at the local level. They construct a panel of Californian counties for the years 1988-98 using data from the American Medical Association and specify reduced-form models for the supply of doctors per 100,000 inhabits for both specialists and primary care. First differencing is used to deal with omitted variables and dynamics are included to allow for inertia in supply responses. The models are estimated by the systems version of the Arellano-Bond estimator (Arellano and Bond (1991); Arellano and Bover (1995); Blundell and Bond (1998)), with both one-step and two-step estimates of the standard errors. The internal instruments, drawn from lagged variables, are augmented by some external instruments. The tests for autocorrelation are consistent with the assumptions of the Arellano-Bond model: there is evidence of first order autocorrelation in the residuals but not of higher order autocorrelation. The results show that the supply of specialists is responsive to changes in the relative market penetration of HMOs, but the supply of primary care doctors is not.

The Arellano-Bond approach is also adopted by Tamm et al. (2007) to estimate price effects on insurers' market shares. They use an unbalanced panel of insurers, who were active in the German social health insurance system between January 2001 and April 2004. This issue is important as price sensitivity among consumers is a precursor for the success of reforms based on the notion of managed competition. They adopt a dynamic

panel data model for the logarithm of each insurers market share. This aggregate model is motivated by an individual-level multinomial logit model for the choice of insurer and dynamics are introduced to capture the fact that only a fraction of consumers will switch companies during a given year. The model is estimated using the Arellano-Bond estimator and the systems-GMM estimator. The standard Arellano-Bond GMM estimator, that relies of lagged levels to instrument lagged differences, may perform poorly. This suggests using the systems approach, but this approach requires stronger restrictions on the initial conditions: such that the first period error term and the first differences of the regressors have to be uncorrelated with the individual effect. In this case the orthogonality conditions for the systems approach are rejected. The hypothesis of a unit root is not rejected and models are also estimated for differences in log-market share as a function of levels of the regressors. The estimated short-run price elasticities are small but, due to the dynamics, the long-run impact of prices is substantial.

Other applications of the GMM and systems-GMM approaches include Baltagi et al. (2005) who estimate dynamic equations for the log of hours worked by Norwegian doctors, using panel data from the personnel register of the Norwegian Association of Local and Regional Authorities. Clark and Etilé (2002) use seven waves of the British Household Panel Survey (BHPS) to estimate dynamic models for cigarette consumption. This creates some problems when the data are first differenced as there is considerable heaping of the self-reported data around focal values such as 20 cigarettes per day. Health shocks are shown to influence levels of smoking. Hauck and Rice (2004) use eleven waves of the BHPS to estimate models for mental health, measured by the GHQ-12 score. They compare static variance components models and dynamic GMM estimators and find greater persistence of mental health problems among those with lower socioeconomic status. Windmeijer et al. (2005) use the systems GMM estimator in panel data models of the demand for outpatient visits by GP practices in England.

5.2 Applications with categorical outcomes

Pooled and random effects specifications

Contoyannis et al. (2003) consider the determinants of a binary indicator for functional limitations using seven waves (1991-1997) of the British Household Panel Survey (BHPS). Their models allow for persistence in the observed outcomes due to state dependence (a direct effect of previous health status), unobservable individual effects (heterogeneity which is due to unobserved factors that are fixed over time) and persistence in the transitory error component. Allowing for persistence is important: a comparison of the observed outcomes with those predicted by a simple binomial model shows that persistence is substantial in the data. They estimate models for the repeatedly observed binary health indicator, with and without state dependence, using panel probit models. These are estimated by Maximum Simulated Likelihood (MSL) using the GHK simulator with antithetic acceleration. They also implemented a test for the existence of asymptotic bias due to simulation which is used to select the number of replications required for use in MSL.

In related work Contoyannis et al. (2004b) explore the dynamics of self-assessed health (SAH) in the British Household Panel Study (BHPS). The variable of interest is an ordered measure of self-assessed health and the BHPS reveals evidence of considerable persistence in individual's health status. As SAH is measured at each wave of the panel

there are repeated measurements for a sample of individuals. SAH is modelled using a latent variable specification, which is estimated using pooled ordered probits (with robust inference) and random effects ordered probit models. The presence of lagged health is designed to capture state dependence, the influence of previous health history on current health. The error term is split into two components, the first captures time invariant individual heterogeneity, the second is a time varying idiosyncratic component.

In this kind of application it is quite likely that the unobserved individual effect will be correlated with the observed regressors, such as household income. To allow for this possibility they parameterize the individual effect (Mundlak (1978); Chamberlain (1984); Wooldridge (2005)). This allows for correlation between the individual effects and the means of the regressors. In addition, because they are estimating dynamic models they need to take account of the problem of initial conditions. It is well known that in dynamic specifications the individual effect will be correlated with the lagged dependent variable, this gives rise to what is known as the *initial conditions problem*, that an individual's health at the start of the panel is not randomly distributed and will reflect the individual's previous experience and be influenced by the unobservable individual heterogeneity. To deal with the initial conditions an attractively simple approach suggested by Wooldridge (2005) is used. This involves parameterising the distribution of the individual effects as a linear function of initial health at the first wave of the panel and of the time means of the regressors, and assuming that it has a conditional normal distribution. As long as the correlation between the individual effect and initial health and the regressors is captured by this equation it will control for the problem of correlated effects. Its ease of implementation stems from the fact that the equation for u_i can be substituted back into the main equation and the model can then be estimated as a pooled ordered probit or a random effects ordered probit using standard software to retrieve the parameters of interest. Contoyannis et al. (2004b) find that self-assessed health is characterized by substantial positive state dependence and unobserved permanent heterogeneity. Including state dependence dramatically reduces the impact of individual heterogeneity. Conditioning on the initial period health outcomes and within-individual averages of the exogenous variables reduces the impact of heterogeneity and state dependence. Unobservable heterogeneity accounts for around 30 per cent of the unexplained variation in health.

Similar dynamic panel probit models are used by Gannon (2005). In her case the outcome of interest is a binary measure of labour force participation which is assumed to be a function of past labour force participation and health limitations. This gives dynamic panel probit models that are estimated in pooled and random effects versions using the Wooldridge (2005) approach to deal with the initial conditions problem. The models are estimated with the Living in Ireland Survey (LIS) which is the Irish component of the European Community Household Panel (ECHP). Nolan (2007) also uses the LIS but uses a dynamic random effects Poisson specification to model the use of GP visits. She adopts the Wooldridge approach to model the initial conditions. In contrast Arulampalam and Bhalotra (2006) use Heckman's approach to specify the initial conditions in a Markov model of infant deaths among Indian families.

GMM estimators

In Jones and Labeaga (2003) a panel of Spanish households is used to test the empirical formulation of the rational addiction model (Becker et al. (1994)). These data raise

problems of measurement errors, censoring, and unobservable heterogeneity. Jones and Labeaga (2003) use sample separation information to exclude those households who never purchase tobacco. To deal with the remaining zeros they compare specifications based on infrequency of purchase and on censoring. GMM and system-GMM are used to deal with errors-in-variables and unobservable heterogeneity (Arellano and Bond (1991); Bover and Arellano (1997)). Within-groups two-step, within-groups three-step GMM and Minimum Distance methods are used to allow for censoring. To reduce the influence of distributional assumptions they adopt a semiparametric approach to estimate each of the *T* cross-section equations using Powell's (1986) Symmetrically Censored Least Squares (SCLS). There is evidence that the rational addiction specification is sensitive to unobservable heterogeneity and censoring and the results suggest that failure to account for heterogeneity may lead to over-estimates of the impact of addiction. The panel data estimators imply that behaviour is more forward looking than suggested by the results that fail to correct for heterogeneity.

Finite mixture models

Deb (2001) applies a random effects probit model in which the distribution of the individual effect is approximated by a discrete density. This is an example of a finite mixture model and it relaxes the normality assumption for the distribution of the random effects. Deb uses Monte Carlo experiments to assess the small sample properties of the estimator. These show that only 3-4 points of support are required for the discrete density to mimic normal and chi-square densities and to provide approximately unbiased estimates of the structural parameters and the variance of the individual effects. Deb (2001) applies the model to a cross-section of individuals clustered in families, where the random effect represents unobserved family effects. It is assumed therefore that all individuals in each family belong to the same latent class. This approach aims to approximate the distribution of the random (family) intercepts, whereas the responses to the explanatory variables are not allowed to vary across latent classes. Clark et al. (2005) develop a latent class ordered probit model for reported well-being, in which individual time invariant heterogeneity is allowed both in the intercept and in the income effect.

5.3 Applications with count data

Poisson-lognormal mixtures

Winkelmann (2004) proposes an alternative two-part, or hurdle, model based on a Poisson-lognormal mixture rather than the usual negbin variants. Hurdle, or two-part, models make a distinction between the decision to seek care, modelled as a binary choice, and the conditional number of visits, modelled as a truncated count data regression. They are the most widely used specification in the recent applied literature (see e.g., Alvarez and Delgado (2002); Chang and Trivedi (2003); Sarma and Simpson (2006); Yen et al. (2001)), although Santos Silva and Windmeijer (2001) have pointed out that they may be problematic in applications where it cannot be assumed that there is a single spell of illness for each period of observation in the data. Winkelmann's (2004) model leads to a probit equation for the first hurdle and a truncated Poisson-lognormal model for the second. Unlike the negbin model, the latter does not have a closed form and is estimated using Gauss-hermite quadrature. An application to the 1997 reform of copayments for prescription drugs in Germany uses data on quarterly doctor visits in the German

Socioeconomic Panel (GSOEP). This confirms Deb and Trivedi's (2002) result that finite mixture models outperform negbin hurdle models. But the results show that the normal hurdle model fits better than both of these specifications.

The innovation in Van Ourti (2004) is to include a Gaussian random effect in the two-part model. The time invariant individual effect appears as a common factor in both parts of the model, with the factor loading in the first equation normalised to one for identification. The individual effect is then integrated out with the resulting integral evaluated by Gauss-Hermite quadrature. In an empirical application to GP and specialist visits and to nights in hospital in the Panel Study of Belgian Households (PSBH), the panel version of the two-part model (2PM-PA) is compared to a one part-model with a Gaussian individual effect (1PM-PA) and to pooled versions of the one-part and two-part models (1PM-PO, 2PM-PO). On the basis of the log-likehoods and the AIC and BIC the 2PM-PA specification is preferred to the simpler specifications.

While Van Ourti (2004) extends the Gaussian random effects model to the two-part specification for count data, Munkin and Trivedi (1999) and Riphahn et al. (2003) do the same for a bivariate count data model, dealing with the case where there are two dependent variables both measured as integer counts. Munkin and Trivedi (1999) propose a model that is designed for cross section data and that is applied to the number of emergency room visits and of hospitalisations in data from the US National Medical Expenditure Survey, 1987-88 (as used by Deb and Trivedi (1997)). They construct the bivariate model by specifying the marginal distributions of the counts and then adding a correlated heterogeneity term to give a Poisson-lognormal mixture. This has conditional mean functions:

$$\begin{aligned} \lambda_{i1} &= \exp(x_{i1}'\beta_1 + \varepsilon_{i1}) \\ \lambda_{i2} &= \exp(x_{i2}'\beta_2 + \varepsilon_{i2}) \\ (\varepsilon_{i1}, \varepsilon_{i2}) &\sim N[(0,0),(\sigma_1^2, \rho\sigma_1\sigma_2, \sigma_2^2)] \end{aligned} \tag{33}$$

The log-likelihood function for the model involves a two-dimensional integral and is estimated by Maximum Simulated Likelihood (MSL) with antithetics and a correction for first order simulation bias. Using a bivariate model does not lead to substantial changes in the estimated effects of the regressors but the overall fit of the model does improve.

The model proposed by Riphahn et al. (2003) is designed for panel data and they apply it to separate measures of doctor visits and hospital inpatient visits from 12 waves of the German Socioeconomic Panel (GSOEP) for 1984-95, focusing on West Germans aged between 25 and 65. Their specification extends the single equation model of Geil et al. (1997) by adding two Gaussian error components to each equation (j), one a time invariant individual effect (u_{ij}) and the other a time-varying idiosyncratic error term (ε_{itj}):

$$\begin{aligned} \lambda_{it1} &= \exp(x_{it1}'\beta_1 + u_{i1} + \varepsilon_{it1}) \\ \lambda_{it2} &= \exp(x_{it2}'\beta_2 + u_{i2} + \varepsilon_{it2}) \end{aligned} \tag{34}$$

Correlation between the two equations is introduced by assuming that the error terms are drawn from a bivariate normal distribution. Although it is not clear from the paper, it appears that the individual effects (u_{ij}) are assumed to be independent of each other which seems rather restrictive. Computation is based on a combination of Gauss-

Hermite quadrature, to integrate over the time invariant individual effect, and Gauss-Legendre quadrature, to integrate over the bivariate distribution within each period.

Finite mixtures

Deb and Trivedi (1997), Deb and Trivedi (2002), Deb and Holmes (2000), Jimenez-Martin et al. (2002) and Sarma and Simpson (2006) estimate finite mixture models for count measures of health care use, in which a negbin distribution is assumed within each latent class. Lourenço and Ferreira (2005) extend the application of the negbin finite mixture model to a truncated sample, from the 2003-04 Europep survey for Portugal, where data is only collected for those who visit health centres. This means that the data are drawn from an endogenous sampling scheme and are truncated at zero and raises the question of whether the distribution of unobserved heterogeneity should be defined over the whole population or only the truncated sample.

Jochmann and Leon-Gonzalez (2004) propose a specification that uses a semiparametric Bayesian approach. This can be seen as an extension of Deb and Trivedi (1997). They start with a parametric 'random coefficients' specification of the Poisson model as a benchmark. In this model the random slopes (b_i) are assumed to be drawn from a multivariate normal distribution, so the conditional mean function takes the form:

$$\lambda_{it} = \exp(x_{it}'\beta + w_{it}'b_i + \varepsilon_{it}) \tag{35}$$

The semiparametric element of the model is introduced by using a Dirichlet process mixture for the prior on the random effects. This gives a mixture model with a random number of components and extends the usual treatment of latent class models that fix the number of components. The Dirichlet process specifies a base distribution, in this case assumed to be normal, and a fixed number of mass points, in this case set equal to 10. Then, new draws of the random effects are a mixture of draws from the base distribution and draws from existing clusters of values. The end product is a discrete distribution where the number of mass points is random. Estimation of the model is done by MCMC, based on Gibbs sampling, with data augmentation (the random effects and latent variables are treated as parameters to be estimated) and incorporating a Metropolis-Hastings step where the Gibbs sampling cannot be used. The MCMC algorithm was run for 30,000 iterations, discarding the first 5,000 for the burn-in period. The application uses data on the number of visits to the doctor in the previous quarter from the 1997-2001 waves of the German Socioeconomic Panel (GSOEP). The aim is to test for horizontal equity in the delivery of care by seeing whether non-need factors play a significant role in explaining variation in utilisation. This is tested using Bayes factors and horizontal equity is not rejected with these data.

The two dominant strands in the recent literature, hurdle models and finite mixture models, are brought together in the latent class hurdle model developed by Bago d'Uva (2006). Her model uses a panel of individuals across time: individuals i are observed T_i times. Let y_{it} represent the number of doctor visits in year t. The joint density of $y_i = (y_{i1}, \ldots, y_{iTi})$ is given by:

$$g(y_i \mid x_i; \pi_{i1}, \ldots, \pi_{iC}; \theta_1, \ldots, \theta_C) = \sum_{j=1}^{C} \pi_{ij} \prod_{t=1}^{T_i} f_j(y_{it} \mid x_{it}; \theta_j) \tag{36}$$

where x_i is a vector of covariates, including a constant and θ_j are vectors of parameters, and $0<\pi_{ij}<1$ and $\Sigma^{C}_{j=1}\pi_{ij}=1$. Conditional on the class that the individual belongs to, the number of visits in period t, y_{it}, is assumed to be determined by a hurdle model. The underlying distribution for the two stages of the hurdle model is the negative binomial (negbin). Formally, for each component $j = 1,\ldots C$, it is assumed that the probability of zero visits and the probability of observing y_{it} visits, given that y_{it} is positive, are given by the following expressions:

$$
\begin{aligned}
f_j\left(0 \mid x_{it};\beta_{j1}\right) &= P\left[y_{it}=0 \mid x_{it},\beta_{j1}\right] = \left(\lambda_{j1,it}^{1-k}+1\right)^{-\lambda_{j1,it}^{k}} \\
f_j\left(y_{it} \mid y_{it}>0, x_{it};\beta_{j2}\right) &= \frac{\Gamma\left(y_{it}+\dfrac{\lambda_{j2,it}^{k}}{\alpha_j}\right)\left(\alpha_j\lambda_{j2,it}^{1-k}+1\right)^{-\frac{\lambda_{j2,it}^{k}}{\alpha_j}}\left(1+\dfrac{\lambda_{j2,it}^{k-1}}{\alpha_j}\right)^{-y_{it}}}{\Gamma\left(\dfrac{\lambda_{j2,it}^{k}}{\alpha_j}\right)\Gamma\left(y_{it}+1\right)\left[1-\left(\alpha_j\lambda_{j2,it}^{1-k}+1\right)^{-\frac{\lambda_{j2,it}^{k}}{\alpha_j}}\right]}
\end{aligned}
\qquad (37)
$$

where $\lambda_{j1,it}=\exp\left(x'_{it}\,\beta_{j1}\right)$, $\lambda_{j2,it}=\exp\left(x'_{it}\,\beta_{j2}\right)$, α_j are overdispersion parameters and k is an arbitrary constant.

As in the standard hurdle model, β_{j1} can be different from β_{j2} reflecting the fact that the determinants of care are allowed to have different effects on the probability of seeking care and on the conditional number of visits. On the other hand, having $\left[\beta_{j1},\beta_{j2}\right]\neq\left[\beta_{l1},\beta_{l2}\right]$ for $j\neq l$, reflects the differences between the latent classes. Various special cases are nested within the general model. It can be assumed that all the slopes are the same, varying only the constant terms, $\beta_{j1,0}$ and $\beta_{j2,0}$, and the overdispersion parameters α_j. This represents a case where there is unobserved individual heterogeneity but not in the responses to the covariates. The most flexible version allows α_j and all elements of β_{j1} and β_{j2} to vary across classes. The finite mixture hurdle model also accommodates a mixture of sub-populations for which health care use is determined by a negbin model (the two decision processes are indistinguishable) and sub-populations for which utilisation is determined by a hurdle model. This is obtained by setting $\beta_{j1}=\beta_{j2}$, for some classes. Setting them equal for all of the classes gives a panel data version of Deb and Trivedi (1997) latent class negbin model. Bago d'Uva (2006) applies the latent class hurdle model to panel data from the RAND Health Insurance Experiment and finds a higher price effect on health care utilisation for the latent class of "low users". This is mostly attributable to the impact of price on the probability of seeing a doctor rather than the conditional number of visits.

5.4 Applications of quantile regression and other semiparametric methods

Quantile regression is of particular value when there is interest in the full conditional distribution of an outcome rather than just the conditional expectation of the variable. The method provides one way to allow for heterogeneity in treatment effects over the range of the conditional distribution. It is a semiparametric approach that avoids distributional assumptions about the error term. Also it has the attractive property of invariance to monotone transformations of y and therefore avoids the retransformation problem.

Kan and Tsai (2004) apply quantile regression to the conditional distribution of body mass index (BMI) using data from the Cardiovascular Disease Risk Factors Two-Township Study (CVDFACTS) for Taiwan. Quantile regression is well suited to an analysis of obesity as interest focuses on the upper tail of the distribution of BMI rather than the area around the mean. Other studies have tended to create an indicator variable for obesity using the published clinical thresholds but the quantile approach makes better use of the available variation in BMI in the tail of the distribution.

Lee and Jones (2006) use the same dataset for Taiwan dentistry as Lee and Jones (2004) to provide evidence on heterogeneity in dentists' activity. The heterogeneity is of particular importance because dentists' responses are likely to differ widely from high to low activity dentists. Quantile regressions provide a useful method to investigate the differential responses of dentists to various observable variables. It is shown that time trends for dentists at higher quantiles have greater fluctuations than those at lower quantiles. The clinic-hospital gap in activity at higher quantiles is greater than at low quantiles. Clinic dentists at higher quantiles have much higher numbers of visits and numbers of treatments than those at lower quantiles, but they provide less intensity of care. Dentists in deprived areas have higher activity than those in non-deprived areas, but only when they are located at higher quantiles.

Winkelmann (2006) applies quantile regression to count data on doctor visits in the German Socioeconomic Panel (GSOEP). This allows an evaluation of the impact of the 1997 reform of copayments for medicines on the full conditional distribution and not just the mean. The quantiles of a count are integer valued and cannot be represented by a continuous function of the covariates, such as $exp(x'\beta)$. So Winkelmann (2006) adopts the method proposed by Machado and Santos Silva (2005). This transforms the data by "jittering": adding a uniform random variable to the counts and then applying quantile regression to the resulting continuous variable. The results show a greater impact on the reform on the lower quantiles, which is consistent with the earlier evidence from a hurdle model in Winkelmann (2004), that showed a larger effect in the first part of the model.

Applications of other semiparametric regression methods are relatively sparse in the health economics literature. Askildsen et al. (2003) use Kyriazidou's (1997) semiparametric estimator for a panel data sample selection model in order to estimate nurses' labour supply in Norway. This allows for individual effects in the selection equation and the hours equation that may be correlated with each other and with the observed covariates. A conditional logit is used to estimate the selection equation. Then the hours equation is estimated in first differences by weighted least squares, with kernel weights applied to the difference in the linear index from the selection equation between different periods. The aim is to difference observations across periods for which the probability of selection is (approximately) the same. Rettenmaier and Wang (2006) use a semiparametric estimator for the Tobit model to model persistence in Medicare reimbursements. The model allows for fixed effects and a lagged dependent variable, but assumes that initial conditions are fixed.

It is widely believed that income has a direct effect on health but it is often argued that indirect income effects due to relative deprivation may be equally important. Wildman and Jones (2007) investigate these relationships using parametric and semiparametric panel data models. By allowing for a flexible functional form for income they seek to ensure that coefficients on relative deprivation variables are not an artefact of a highly

non-linear relationship between health and income. Parametric estimation may lead to biased coefficients if the parameterisation of the explanatory variables is incorrect. Semiparametric partially linear estimation overcomes this problem by allowing an unspecified relationship between health and income (Robinson (1998)). The results provide strong evidence for the impact of income on self-reported measures of health for men and women. These results are robust across a range of techniques and are resilient to the inclusion of measures of relative deprivation. The parametric results for relative deprivation largely reject its influence on health, although there is some evidence of an effect in the semiparametric models.

6. Multiple equation models

6.1 Applications using MSL

Balia and Jones (2007) use the first wave of the British Health and Lifestyle Survey (HALS) from 1984-85 along with the longitudinal follow-up from May 2003 to model the determinants of premature mortality and to assess the relative contribution of lifestyle factors to the gradient in mortality in Britain. A behavioural model, that relates mortality to observable and unobservable factors, is used to motivate the empirical specification. Death, self-assessed health and a range of health-related behaviours are measured as binary outcomes and, to capture the effect of lifestyles on mortality and morbidity in the presence of common unobservable factors, the model is estimated as a recursive multivariate probit. The full system is estimated by maximum simulated likelihood (MSL). Health inequality is explored using a decomposition analysis of the Gini coefficient. The results contradict the view that lifestyles only play a minor part in health inequalities.

Deb and Trivedi (2006) and Deb et al. (2006a) use a latent factor specification to model selection into treatment in nonlinear models and adopt a maximum simulated likelihood estimator (MSL). The aim is to estimate causal effects using a structural model, motivated by a 'selection on unobservables' approach in which the parametric distribution of the unobservables is specified and the full model is estimated by maximum likelihood. In this case the outcomes of interest are binary and count measures of health care utilisation. The treatment variables reflect the individual's choice of insurance plan, which is modelled using a random utility framework. The options are categorised as HMOs, other managed care and non-managed care plans, with the choice of options specified using a multinomial logit specification. In Deb and Trivedi (2006) the data are taken from the 1996 US Medical Expenditure Panel Survey (MEPS) while in Deb et al. (2006a) the MEPS data are augmented by the 1996 Community Tracking Survey (CTS) as a test for the external validity of the findings. Utilisation and insurance plan are modelled simultaneously to take account of the possibility of selection by patients, insurers and providers. Identification of the latent factor model rests on the assumption that each of the multinomial choices depends on a unique latent factor, based on *iid* normal draws, and these are allowed to be freely correlated with the error in the outcome equation. The parametric specification is identified by functional form but exclusion restrictions are also imposed, using employment status and occupational sector as predictors of insurance plans, while excluding them from the outcome equations. Simple linear models are used to provide informal checks for the validity of these instruments. Estimation uses

maximum simulated likelihood, accelerated by using quasi-random draws from Halton sequences. The results do not show evidence of favourable selection into HMO plans but the average treatment effects on the use of care are much larger when selection is taken into account and there is considerable heterogeneity in the effects. Monte Carlo simulation is also used to compute the standard errors of these treatment effects.

Other applications of MSL include Lindeboom et al. (2002) who use the LASA panel to estimate a five equation model for the use of long-term care services among elderly residents of Amsterdam. They take draws from a multivariate normal distribution and use antithetics to accelerate the estimation. The results show strong effects of health status, sex, socioeconomic status and prices on the use of institutional care. Two papers by Pudney and Shields (2000a) and Pudney and Shields (2000b) use MSL to estimate a system of equations made up of a generalised ordered probit model for British nurses's pay grades along with auxiliary equations for training, career breaks, work outside the NHS, and part-time work. A common factor structure leads to a log-likelihood based on the multivariate normal and hence the need for simulation estimation. The MSL estimation only uses 50 replications, without acceleration, which is relatively few for this kind of application. Pudney and Shields (2000b) makes a case for identification based on functional form rather than exclusion restrictions. Vera-Hernandez (2003) makes use of MSL to estimate a structural model of insurance coverage and health care use applied to data from the RAND Health Insurance Experiment.

6.2 Applications using Bayesian MCMC

Patient selection makes it hard to identify the impact of the size and characteristics of hospitals on the quality of their outcomes. The problem arises when there is selective admission that is influenced by unobservables such as unmeasured severity that are also associated with the quality of outcomes. Geweke et al. (2003) find evidence of patient selection among 78,848 Medicare patients treated in 114 hospitals in Los Angeles county between 1989 and 1992. They focus on patients aged over 65 with a diagnosis of pneumonia taken from administrative data on hospital discharges collected by the State of California Office of Statewide Health Planning and Development. Quality of the clinical outcomes is measured by deaths in hospital within 10 days of admission. A structural probit model for deaths is coupled with a reduced form multinomial probit model for the patient's choice of hospital, allowing correlation in the error terms to capture patient selection. The system of equations is estimated by a Bayesian approach using the MCMC simulator of the posterior distribution. Gibbs sampling with data augmentation breaks the estimation into steps, first simulating the latent dependent variables and then estimating the linear simultaneous equations system. The model is identified by using distances from the patients' homes to the hospitals as an instrument. The raw data and simple probits do not show a relationship between hospital size and mortality rates but the MCMC results reveal a U-shaped relationship, with better quality in the smallest and largest hospitals.

Deb et al. (2006b) start with a conventional two-part model for medical expenditure. This is applied to data on ambulatory care, which has 17% of zero observations, and hospital care, which has 94% zero observations and which exhibits positive skewness and excess kurtosis. The data are drawn from the US Medical Expenditure Panel Survey (MEPS) for 1996-2001. Giving six repeated cross sections and 20,460 observations. The standard two-part set up is used with a binary choice equation and a conditional

regression for the log of expenditure. However this is augmented by a multinomial probit model to allow for the endogenous selection of insurance plans, which fall into three categories: HMO, PPO and FFS. To capture the possibility of selection bias the error terms from the insurance equations (u) are assumed to be linearly associated with the error terms in the two parts of the model for expenditure:

$$\begin{aligned} \varepsilon_1 &= u'\delta + \upsilon \\ \varepsilon_2 &= u'\pi + \tau \end{aligned} \tag{38}$$

This assumes that the ε's are only conditionally independent given u and relaxes the usual assumption that the two parts of the model are independent. Like Geweke et al. (2003) the full system of equations is estimated by Bayesian MCMC and Bayes factors are used to construct a test for the exogeneity of the choice of insurance plan. This test shows evidence of substantial selection bias. Having estimated the model the authors show how to define and compute estimated treatment effects for the impact of insurance plan on expenditure, this makes uses of data augmentation to impute the latent variables. It should be noted that the approach used to compute the treatment effects involves a standard retransformation for log-scale data and therefore relies on a strong assumption about the absence of heteroskedasticity in the expenditure data (Manning and Mullahy (2001)).

6.3 Applications using finite mixtures

The common factor model was introduced earlier in this chapter in the context of Aakvik et al. (2005) evaluation of a Norwegian Vocational Rehabiliation programme (see equations 8-10 above). Aakvik et al.'s (2005) application takes a parametric approach and assumes that the common factor is normally distributed. An alternative, semiparametric approach is to use a finite density and estimate a discrete factor model (DFM). The discrete factor model has two main advantages over MSL. First, it is semiparametric and therefore more robust than the parametric approach, which relies on strong distributional assumptions. Also, in general, it is easier to compute, involving standard numerical methods for maximum likelihood estimation or the use of the EM algorithm, rather than computationally intensive Monte Carlo simulation (see Arcidiancono et al. (2007)). However in practice there can be problems with identification, manifested in failure of convergence and problems with multiple optima.

Like Aakvik et al. (2005), Aakvik et al. (2003) use a latent variable framework to specify the impact of multidisciplinary treatment for back pain on the probability of leaving sickness benefits. Using a structural model, based on latent variables, allows them to define the average treatment effect (ATE), the average treatment effect on the treated (ATT) and to allow for heterogeneous marginal treatment effects (MTE). In this application the unobservables are modelled using a discrete factor structure with distance to the nearest hospital used to identify the model. The estimates show a positive effect of around 6 percentage points on the probability of leaving sickness benefits.

Rous and Hotchkiss (2003) use data from 254 Texas counties on all reported births in 1993 to explore the impact of prenatal care on birthweight. They estimate a discrete factor model with three equations: a logit model for whether the pregnancy is carried to full-term and linear regressions for a measure of prenatal care visits and for birthweight. Travel distance to the nearest provider of abortions is used to identify the first equation,

the availability of obstetricians is used for the second and the gender of the child for the third. The factor model shows evidence of adverse selection effects. Picone et al. (2003b) merge panel data from the US National Long-term Care Survey (NLTCS) for 1984-95 with the National Death Index to investigate the impact of treatment intensity on survival rates and other health outcomes. Their model involves a system of three equations for treatment intensity, length of stay and health outcomes. These are assumed to have a common factor structure which is estimated using a one-factor model. The selection of models is based on the LR statistic (Mroz (1999)) and 1000 bootstrap replications are used to avoid the problem of local optima. The model is identified by excluding area data on the cost of capital, the Herfindahl index for hospital concentration and a wage index from the mortality equation. The results suggest that treatment intensity has a beneficial effect. Hamilton et al. (2000) compare US and Canadian data to see whether waiting time for surgery for hip fractures influences the outcomes of the treatment, measured by length of stay and inpatient mortality. They use discharge data for 20,995 patients admitted to acute hospitals in Quebec and Massachusetts between 1990 and 1992 and estimate a competing risks model for delay before surgery and post-surgery length of stay. They use the Heckman-Singer specification for the common heterogeneity with two mass points (Heckman and Singer (1984)). Day of week of admission is used as an instrument for delay until surgery. The raw data shows a strong relationship between delays and outcomes but this disappears once unobserved frailty is taken into account. The higher observed inpatient mortality in Quebec is attributed to the longer length of stay rather than poorer outcomes: the longer the patients remain in hospital the more likely they are to die there rather than at home.

Other applications of the discrete factor model include Bhattacharya et al. (2003) who model the impact of public and private insurance on HIV-related mortality. Mello et al. (2002) use a discrete factor specification for a two equation models of health plan choice, whether to join a Medicare HMO, and various measures of subsequent health care utilisation applied to data from the Medicare Current Beneficiary Survey for 1993-96. A similar model of HMO enrolment and subsequent hospital use by Kan et al. (2003) finds evidence of strong selection effects when a discrete factor specification is used to deal unobservables. Rous and Hotchkiss (2003) use the Nepal Living Standards Survey for 1996 in a model of choice of health care provider, levels of expenditure and health outcomes with a two-factor specification to allow for community and household effects. Holmes (2005) combines a multinomial logit model with a discrete factor specification to evaluate the US National Health Service Corps (NHSC) programme that was designed to encourage doctors to locate in under-serviced areas.

The discrete factor model allows for a common factor in the intercept of each equation. In contrast the latent class model allows all of the parameters to vary across the latent classes. Clark and Etilé (2006) use the latent class framework to approximate the continuous distribution of the individual effects in a dynamic random effects bivariate probit model. They apply the model to data on smoking among couples in the British Household Panel Survey (BHPS) and use the simulated annealing version of the EM algorithm to estimate the model. Atella et al. (2004) develop a latent class model for the joint decisions of consulting 3 types of physician. The authors assume that, within a latent class, each decision can be modelled by an independent probit, so the joint distribution of the three binary outcomes is a product of probits.

6.4 Applications using copulas

The copula approach leads to closed forms and avoids the need for numerical integration. Also it circumvents the problem of the limited menu of parametric specifications of multivariate distributions that are available, especially when normality is an unsuitable assumption, for example when dealing with highly skewed data.

Although copulas are not mentioned explicitly, Prieger (2002) proposes an extension of the sample selection model that is built around the FGM copula. This is applied to data from the 1996 wave of the Medical Expenditure Panel Survey (MEPS). The outcome equation is a duration model for hospital length of stay and the selection equation measures whether the individual had an inpatient stay or not during the survey period. Prieger's approach builds on Lee's (1983) selection model that uses a bivariate normal copula. In both cases the use of copulas to model the joint distribution allows for the marginal distributions to be non-normal. As the outcome is a measure of duration exponential, gamma, log-logistic, lognormal and Weibull specifications are considered and the gamma is selected as the preferred model. This reflects an attractive feature of the copula approach, that model selection can focus on finding an acceptable specification of the marginal distributions before turning to the joint distribution. The specifications of the full selection model are compared using the AIC, CAIC and BIC information criteria as well as the Voung test for non-nested models. These tests favour the FGM over the bivariate normal copula in terms of goodness of fit. Smith (2003) uses the same data as Prieger (2002) but a different copula, the Frank copula from the Archimedean class, that improves the fit of the model and changes the estimates of average length of stay.

Copulas are used by Zimmer and Trivedi (2006) to model dependence in a system of nonlinear reduced form equations. Their application focuses on couples' decisions about insurance coverage and health care use and consists of three equations: one for the husband's utilisation, one for the wife's and one for whether or not the couple take out separate health insurance policies. The marginal distributions for utilisation are assumed to be negbin (NB2) and a probit is used for the insurance equation. The equations are assumed to be linked by common unobservable factors and the joint distribution is modelled using a trivariate Frank copula, which is derived using the mixture of powers approach. The use of copulas avoids having to select a parametric specification for the unobservable heterogeneity and is computationally tractable. To emphasise these points Zimmer and Trivedi (2006) compare their results to those derived from a maximum simulated likelihood approach, based on multivariate normality. The model is applied to data from four waves of the US National Medical Expenditure Panel Survey (MEPS). An apparent limitation of the copula approach is that it works by specifying marginal distributions and then modelling dependence so that the emphasis is on system of reduced form equations rather than on conditional distributions. However Zimmer and Trivedi (2006) show how the estimates can be used to derive the conditional distribution and hence compute the average treatment effect (ATE) of insurance coverage on health care use.

7. Evaluation of treatment effects

7.1 Matching

Matching provides a general approach to deal with selection on observables. It addresses the problem that in the observed data confounding factors may be non-randomly distributed over the treated and controls. Rosenbaum and Rubin (1983) showed that, rather than matching on an entire set of observable characteristics (x), the dimensions of the problem can be reduced by matching on the basis of their probability of receiving treatment, $P(d=1|x)$, known as the *propensity score*. In practice propensity score matching (PSM) estimators do not rely on exact matching and instead weight observations by their proximity, in terms of the propensity score.

An important requirement is that the model for treatment, used to construct the propensity score, should only include variables that are unaffected by participation in the treatment, or the anticipation of participation. The matching variables should be either time invariant characteristics or variables that are measured before participation in the treatment and that are not affected by anticipation of participation. The crucial condition for identification of treatment effects using the matching approach is that the selection into treatment should be ignorable, conditional on the observed covariates[3].

Jones et al. (2007a) use the European Community Household Panel (ECHP) to estimate the impact of private health insurance coverage on the use of specialist visits in four European countries that allow supplementary coverage. The results show that the probability of having private insurance increases with income and with better reported health. Private insurance has a positive effect on the probability of specialist visits in all countries although the magnitude is sensitive to the choice of estimator. They match treated individuals with non-treated individuals inversely weighted for the distance in terms of estimated propensity scores, with weights constructed using kernel smoothed distance weighting. They ensure that all cases are supported by controls. The quality of the matching can be assessed by computing the reduction of the pseudo R^2 of the insurance regression before and after matching. To evaluate the extent to which matching on propensity scores balances the distribution of the x's between the insured and the uninsured group, they compute the bias reduction due to matching for each of the x's.

Dano (2005) uses a 10% sample of the Danish population, drawn from register data, to give a panel for 1981-2000. She estimates the impact of injuries sustained in road traffic accidents on economic outcomes. Although these accidents are unanticipated health shocks their incidence varies with observable and unobservable characteristics that can be associated with the outcomes and the estimates of the treatment effect need to be adjusted for this. Due to the large sample size, one-to-one matching without replacement is used, with matching on the linear index from the propensity score. The matching is combined with a difference-in-differences approach to control for time invariant

[3] Comparisons of treatment effects estimated using randomized experiments with those estimated by matching methods in the labour economics literature cast considerable doubt on the validity of such ignorability conditions and hence on the matching approach (e.g., Lalonde (1986); Smith and Todd (2001); Agodini and Dynarski (2004)).

unobservables. The study finds an impact of injuries on earned income for older and low income individuals but also shows the compensating effects of public transfers in the Danish system.

García-Gómez and López-Nicolás (2006) adopt a matched panel data difference-in-differences estimator. They use panel data from the Spanish component of the European Community Household Panel (ECHP) to explore the impact of health shocks on employment and of employment shocks on health. Their strategy is to match exactly on pre-treatment outcomes, in order to control for time-invariant observables. This means that controls are restricted to those individuals who were identical to the treated in terms of their pre-treatment outcomes. In order to define the treated and the controls a three-year window is adopted. In the case of health shocks, the treated are those who are in good health in the first year and then move to bad health in the next two years, and who are in employment for the first two years. The controls remain in good health throughout the period and are also in employment in the first two years. The outcome is employment status in the third year. In addition to matching on pre-treatments outcomes, propensity score matching is used to make the treated and controls comparable in terms of their observed characteristics. Four methods of matching are compared: nearest neighbour matching on the propensity score; kernel matching on the propensity score; matching of the four nearest neighbours on a set of explanatory variables; and simple matching on the four nearest neighbours according to the propensity score.

Other studies that use matching are summarised in Table 2.

INSERT TABLE 2 AROUND HERE

7.2 Regression discontinuity

The regression discontinuity (RD) design exploits situations where the assignment to treatment changes discontinuously with respect to a threshold value of one or more exogenous variables. The contrast between individuals on either side of the discontinuity is used to identify the treatment effect. In a sharp regression discontinuity design passing the threshold completely determines the allocation of treatment. In a fuzzy design, which is more likely in practice, the allocation of treatment is stochastic and the threshold creates a discontinuity in the probability of treatment. The discontinuity design relies on a comparison of observations "before & after" the threshold and does not have a separate control group. For this reason, applications typically use a narrowly defined neighbourhood around the discontinuity to try and ensure that the treated and untreated observations are comparable in other respects. Studies that use a discontinuity design were described in Section 2 above (Almond (2006); Lleras-Muney (2005); Pop-Eleches (2006)).

7.3 Difference-in-differences

The difference-in-differences, or diff-in-diff, (DD) approach to evaluation with non-experimental data has been applied extensively in the health economics literature. The

method is based on a before & after (pre-post) design with a control group. It can be used with both panel data and repeated cross sections and requires treatment and control groups to be specified.

The basic form of the DD estimator of the average treatment effect, compares mean outcomes for the treated (1) and controls (0) before (B) and after (A) the treatment:

$$ATE_{DD} = (\bar{y}^1_A - \bar{y}^1_B) - (\bar{y}^0_A - \bar{y}^0_B) \tag{39}$$

With individual panel data the DD estimator can be computed using a two-way fixed effects specification:

$$y_{it} = \gamma(T_i P_t) + x'_{it}\beta + \nu_t + u_i + \varepsilon_{it} \tag{40}$$

The treatment effect is identified by the parameter (γ) on the interaction term between the indicator for whether the individual is in the treated group (T) and the indicator for the post-treatment period (P)[4]. The observed regressors (x) control for any observable differences between the treated and controls and the individual effects (u) control for any time-invariant unobservable differences that maybe correlated with the outcome and with the allocation of treatment. In this sense, the DD estimator combines selection on unobservables with selection on observables, so long as the unobservables are time invariant. The time effects (ν) take account of any time trend in the data that is common to both the treatment and control groups. This implies a "parallel trend" assumption When the DD estimator is applied to repeated cross-section data a further assumption is required: that the composition of the treatment and control groups remains stable over time.

The DD estimator is defined above by using the interaction between the post-treatment period (P) and belonging to the treatment group (T). In some applications exposure to treatment may be defined by the interaction between more than two factors. For example, in Schmidt's (2007) evaluation of the impact of infertility insurance mandates on birth rates in the US, an indicator of whether states have mandates is interacted with whether or not women are over 35, as those over 35 are most likely to suffer infertility. Multiple interactions can be used to define exactly who is exposed to treatment and also to allow for heterogeneity in the size of the treatment effect. This approach is often labelled difference-in-difference-in-differences (DDD).

Chalkley and Tilley (2006) show how economic incentives can influence dental practice. This study exploits the comparison between self-employed and salaried dentists working for the NHS in Scotland to show that the financial incentives of fee-for-service (FFS) increase the intensity of treatment by around 21%. Using a difference-in-differences approach the paper finds that self-employed dentists treat exempt patients, who are assumed to be more likely to be influenced by supplier inducement, more intensively than non-exempt patients, relative to salaried dentists who do not face the financial incentive of FFS. These findings are based on an administrative database, the Management Information and Dental Accounting System (MIDAS), that records claims for self-employed and salaried dentists. The database provides a panel of dentists and

[4] Bertrand et al (2004) highlight the risk of making misleading inferences using the standard DD estimator if there is serial correlation in the outcomes and the standard errors are not adjusted to take account of it.

patients and can be used to control for the practice style of individual dentists as well as measures of patient need.

In January 1994 the health authorities in Belgium increased copayment rates for home and office visits to GPs and for visits to specialists. The increases were substantial: 35% for GP home visits, 45% for GP office visits and 60% for specialist visits. Cockx and Brasseur (2003) use this as a natural experiment. To create a control group they use those who were exempt from charges due to low income among widows, orphans, the disabled and retired. This means that identification relies on the treatment and control groups being comparable and the authors note that identification can only be achieved for low income groups. A difference-in-differences estimator is applied to the logarithm of utilisation and the model is extended to a Rotterdam demand system to accommodate substitution effects induced by the change in relative prices. Interaction terms are used to allow for heterogeneity in the treatment effects. A similar set of reforms in Germany is used as a natural experiment by Winkelmann (2004a). In this case copayments for prescription drugs increased by 6 Deutsche Marks on 1st July 1997, leading to relative price increases of up to 200% depending on the pack size. The policy is evaluated using data from the German Socioeconomic Panel (GSOEP) for the years around the reform, 1995-96 and 1998-99. The control group here is those exempt from these charges, made up of those with private insurance, coinsured children and low income households. A difference-in-differences strategy is adopted with the effect of the copayments on doctor visits identified by the interaction between the timing of the reform and exemption from charges. Winkelmann (2004a) argues that the assumption of a common trend, which is essential for identification, is plausible in this context.

The use of a control group in difference-in-differences methods help avoids the spurious inferences that can arise in a simple before & after comparison. For example, in Wagstaff and Yu's (2007) evaluation of the impact of the World Bank's Health VIII project in Gansu province in China they find only a small reduction in out-of-pocket spending on health care in counties exposed to the programme. A before & after comparison would suggest that there was no significant improvement in this outcome. However the trend in the control group of counties shows a rise in out-of-pocket payments, so the DD estimate reveals better outcomes among the treated relative to the controls. An important caveat is that the validity of the DD estimates relies on the comparability of the treated and controls in terms of the underlying trend in the outcomes. The comparability of treatment and control groups can be enhanced by combining the DD approach with matching estimators (as in Galiani et al. (2005); Dawson et al. (2007); Wagstaff and Yu (2007)). In Wagstaff and Yu (2007) the unmatched DD estimate for the impact of Health VIII on the availability of medical equipment in township health centres (THCs) does not show a significant effect. But when the controls are matched with treated counties an effect is revealed. This is because the counties selected for the project tended to be poorer than the average among the controls so that, on average, the funds available to invest in equipment in the control counties lead to higher rate of increase. Matching with control counties that face the same sort of financial constraints allows a reliable comparison to be made.

The careful use of matching estimators, which should include tests of whether the treated cases and selected controls are balanced in terms of their observed characteristics, provides a link to strategies for testing the robustness of the identification assumptions that are built into the differences-in-differences approach. The comparability of the treatment and control groups can be assessed by comparing their observed characteristics

prior to treatment and, in particular, by testing the parallel trend assumption prior to treatment. A good example of this is Galiani et al. (2005). They use a difference-in-differences strategy applied to panel data on municipalities in Argentina. The treatment of interest is the privatisation of local water services that took place in the 1990s and the outcomes are measures of general and cause-specific child mortality. There is sufficient data for the pre-treatment period to do graphical analysis of the trends in the treatment and control groups. More importantly it is possible to estimate the two-way fixed effects specification for mortality rates only using the data from the pre-treatment period but including an indicator of which municipalities would go on to be treated. Evidence that this indicator is significant would undermine the parallel trends assumption and mean that areas that privatised their water supply were systematically different in terms of (trends in) mortality. In fact, Galiani et al. (2005) find that common trend assumption is not rejected. Their difference-in-differences estimators are refined using a propensity score matching approach. The results show a significant reduction in deaths from infectious and parasitic diseases and suggest that privatisation helped to reduce health inequalities. Other studies that combine difference-in-differences with matching are Dano (2005), Garcia-Gomez and Lopez-Nicolas (2006), and Girma and Paton (2006).

Numerous studies in health economics use the difference-in-differences strategy. Some of these are summarised in Table 3.

INSERT TABLE 3 AROUND HERE

7.4 Instrumental variables

The difference-in-differences design is often applied in the context of natural experiments. Natural experiments and natural controls also form the basis for the instrumental variable (IV) approach, which is intended to capture "selection on unobservables" (see e.g., Auld (2006b)). This approach relies on the variation in treatment that can be attributed to variation in an exogenous variable, or instrument; assigning individuals to treatments on the basis of the instrument mimics the random assignment of an experimental design. This approach is often hard to apply in practice as instruments should be both powerful predictors of treatment and have no direct effect on outcomes. The search for convincing instruments is therefore fraught with difficulty.

There are two broad estimation strategies. The FIML approach specifies a complete system of equations for the outcomes and treatments and estimates them jointly, allowing for common unobservable factors and identifying the model through exclusion restrictions. Estimation can be by MLE, MSL, MCMC, DFM or copulas, as described in Section 6 above. The more commonly used approach is the limited-information or single equation approach, using IV estimators such as 2SLS, GMM and 2SCML. Some studies that use instrumental variables were described in Section 2 above (Arendt (2005); Auld and Sidhu (2005); Evans and Lien (2005); Gardner and Oswald (2007); Lakdawalla et al. (2006); Lindahl (2005)). Other applications are too numerous to describe in detail here (some examples are Cawley et al. (2006); Contoyannis et al. (2005); Dor et al. (2006); Dubay et al. (2001); Dusheiko et al. (2004); Dusheiko et al. (2007); Elliott et al. (2007); Guariglia and Rossi (2004); Hadley et al. (2003); Jewell and Triunfo (2006); Kessler and

McClellan (2002); Lindrooth and Weisbrod (2007); Sasso and Buchmueller (2004) Meer et al. (2003); Schellhorn (2001); Sloan et al. (2001); Houtven and Norton (2004); Yelowitz (2000)).

In Section 1 above it was emphasised that, when treatment effects are heterogeneous, the IV estimator identifies a local average treatment effect (LATE) and that this estimate is conditional on the set of instruments that are used. In a recent paper Basu et al. (2007) apply Heckman and Vytlacil's (1999) local-IV (LIV) estimator which identifies marginal treatment effects (MTE) over the support of the propensity score $p(d=1|x,z)$. Computation of the LIV estimator involves regressing the outcome y on the observed regressors x and on a flexible function of the propensity score, which is estimated using x and the instruments z. The model could be estimated semiparametrically, for example by using a partially linear model, or as in Basu et al. (2007), by adding polynomial and interaction terms between x and $p(d=1|x,z)$. Then the LIV estimator of the $MTE(x,u_d)$ is:

$$LIV = \left[\frac{\partial E(y \mid x, p(x,z))}{\partial p(x,z)} \right]_{1-p(x,z)=u_d} \tag{41}$$

This can be used to test for heterogeneity in the treatment effect and to construct estimates of the other treatment effects of interest such as the ATE and ATT.

Table 1: Key datasets cited in the review

Acronym	Full Title, origin	Format	Homepage
AddHealth	National Longitudinal Study of Adolescent Health, USA	Panel survey	www.cpc.unc.edu/projects/addhealth/
AHEAD	Assets and Health Dynamics Among the Oldest-Old, USA	Panel survey	hrsonline.isr.umich.edu/
BHPS	British Household Panel Survey, UK	Panel survey	www.data-archive.ac.uk/
BRFSS	Behavioral Risk Factor Surveillance System,US	Telephone survey	www.cdc.gov/brfss/
CHNS	China Health and Nutrition Surveys	Panel survey	www.cpc.unc.edu/projects/china
ECHP	European Community Household Panel, EC-15	Panel survey	forum.europa.eu.int/Public/irc/dsis/echpanel/home
ELSA	English Longitudinal Survey of Ageing	Panel survey	www.ifs.org.uk/elsa/
GSCF	Gansu Survey of Children and Families, China	Longitudinal multilevel survey	china.pop.upenn.edu/Gansu/intro.htm
GSOEP	German Socioeconomic Panel	Panel survey	www.diw.de/english/sop/
HALS	Health and Lifestyle Survey, GB	Panel survey	www.data-archive.ac.uk/
HCSUS	HIV Cost and Services Utilization Study, USA	Panel survey	www.rand.org/health/projects/hcsus/
HCUP	Healthcare Cost and Utilization Project, USA	Administrative	www.hcup-us.ahrq.gov/overview.jsp
HES	Hospital Episode Statistics, England & Wales	Administrative	www.dh.gov.uk/en/Publicationsandstatistics/Statistics/HospitalEpisodeStatistics/index.htm
HRS	Health and Retirement Survey, USA	Panel survey	hrsonline.isr.umich.edu/
HSE	Health Survey for England, Welsh Health Survey, Scottish Health Survey	Repeated cross sections	www.data-archive.ac.uk/
HSN	Historical Sample of the Netherlands	Longitudinal sample from census and registers	www.iisg.nl/~hsn/
LASA	Longitudinal Aging Study Amsterdam	Panel survey	www.lasa-vu.nl
LSMS	Living Standards Measurement Study, World Bank	Repeated cross sections	www.worldbank.org/html/prdph/lsms/lsmshome.html
MTR	Minnesota Twin Registry, USA	Longitudinal Register	www.psych.umn.edu/psylabs/mtfs/default.htm
NDF	Natality Detail Files, USA	Census of births	www.cdc.gov/nchs/products/elec_prods/subject/natality.htm
MEPS	National Medical Expenditure Panel Survey, USA	Panel Survey	www.ahrq.gov/data/mepsweb.htm
NCDS	National Child Development Survey, UK	Cohort study	www.data-archive.ac.uk/
NHANES	National Health and Nutrition Examination Surveys, USA	Repeated cross sections	www.cdc.gov/nchs/nhanes.htm
NLSCY	National Longitudinal Survey of Children and Youth, Canada	Panel survey	www.statcan.ca/
NLSY	National Longitudinal Survey of Youth, USA	Panel survey	www.bls.gov/nls/
NLTCS	National Long Term Care Survey, USA	Panel survey	www.nltcs.aas.duke.edu/index.htm
NPHS	National Population Health Survey, Canada	Panel survey	www.statcan.ca/english/survey/household/health/health.htm
PSBH	Panel Study of Belgian Households	Panel survey	www.psbh.be/
PSID	Panel Study of Income Dynamics, USA	Panel Survey	psidonline.isr.umich.edu/

RAND HIE	RAND Health Insurance Experiment, USA	Panel, randomised experiment	www.icpsr.umich.edu/ICPSR/access/index.html
RLMS	Russian Longitudinal Monitoring Study	Panel survey	www.cpc.unc.edu/rlms/
SHARE	Survey of Health, Ageing and Retirement in Europe	Panel Survey	www.share-project.org/
SLID	Survey of Labour and Income Dynamics, Canada	Panel survey	www.statcan.ca/
US Census	United States Census	Population Census	www.census.gov/
WHS	World Health Survey, WHO	Repeated cross sections	www.who.int/healthinfo/survey/en/index.html

Table 2: Studies that use matching estimators

Study	Outcome	Treatment	Method	Comment
Dano (2005)	Earnings, annual employment rate, disposable income, public transfer income	Road traffic accidents	Uses difference-in-differences matching estimator with panel constructed from register data. One-to-one matching without replacement based on linear index from propensity score. Checks for balancing.	Finds an impact on earned income for older and low income individuals. Also shows compensating effects of public transfers in the Danish system.
Frolich et al. (2004)	Labour market outcomes: reintegration into the labour force	Vocational rehabilitation programme in Western Sweden	Multiple treatments with matching based on multivariate balancing scores computed from multinomial probit models. Uses nearest neighbours with replacement. Checks for balancing of covariates after matching	Finds a negative effect of rehabilitation. Many of the effects are insignificant: many controls are used repeatedly in the matching
García-Gómez and López-Nicolás (2006)	Employment, income, self-assessed health	Health shocks and employment shocks	Use matched difference-in-differences estimator. Combine exact matching on pre-treatment outcomes with propensity score matching. Use both nearest neighbour and kernel-smoothed matching.	Find effects of health shocks on employment and activity. Also find an effect of transition to unemployment on self-assessed health.
Girma and Paton (2006)	Teenage pregnancy rates	Free OTC access to emergency birth control for teenagers at pharmacies in	Use difference-in-differences matching estimator with panel data.	No evidence that policy leads to lower teenage pregnancy rates

		England	Matching based on nearest neighbours. Impose common support and test for balancing	
Jalan and Ravallion (2003)	Prevalence and duration of diarrhea in children aged under 5 in rural India	Piped water supply	PSM using logit, matching to 5 nearest neighbours	Finds that matching on household as well as village level controls makes a difference. Average effects are misleading as there is a lot of heterogeneity in the treatment effects, with little impact on poorer and less educated mothers.

Table 3: Studies that use Difference-in-Differences (DD)

Study	Outcome	Treatment	Treated/Controls	Comments
Adams (2007)	Relative wages for older workers	Introduction of pure community rating among small group health insurers in New York in 1993	i) Data for small and large firms in New York only: time interacted with older workers and employment in small firm. ii) Data for small firms in all states: time interacted with older workers and New York	Uses two sets of DDD estimates. Finds an increase in relative wages of older workers in small firms
Carpenter (2004)	Self-reported alcohol use and drunk driving	Zero tolerance drunk driving laws	In DDD estimators zero tolerance laws interacted with age 18-20, the group affected by the laws, and age 22-24	Finds an effect on episodic drinking among males, but for other outcomes.
Chalkley and Tilley (2006)	Intensity of dental treatment	Remuneration of dentists, FFS or salary	Interacts salaried v self-employed dentists with exempt v non-exempt patients.	Find that self-employed dentists treat exempt patients more intensively
Chen et al. (2007)	Use of outpatient and inpatient care and health status among the elderly	Introduction of National Health Insurance (NHI) in Taiwan in 1995	Those who became insured under NHI/those already insured	Find an impact on use of care, which is larger for those on low incomes, but not on health.
Chen and Zhou (2007)	Height, labour supply, earnings	1959-61 famine in China	Birth cohorts interacted with regional differences in severity of famine (proxied by excess mortality)	Find worse outcomes for those exposed to the famine
Cockx and Brasseur (2003)	Demand for physician	Increase in copayments	Non-exempt/exempt	DD applied to a Rotterdam

	services			demand system
Currie and Hotz (2004)	Childhood accident requiring medical attention	Child care regulations	Aged 0-4/aged 5-9	
Davidoff et al. (2005)	Employer sponsored insurance coverage	Reforms of state regulations of small group insurance market in US	Interacts indicators for reform with high/low risk and small/large firms	Uses DD and DDD estimates with repeated cross sections
Dawson et al. (2007)	Waiting times in opthamology	London patient choice experiment (LPCP)	LCPC hospitals/three comparator groups: all hospitals in rest of England; a matched control group; all hospitals from 4 large metropolitan areas	DD combined with propensity score matching The results show a small reduction in average waiting times and reduced dispersion within the LPCP hospitals
Dranove et al. (2003)	Health care expenditures and health outcomes for patients receiving coronary artery bypass graft (CABG)	Mandatory CABG report card laws adopted in New York State in 1991 and Pennsylvania in 1993	Hospitals affected by reform/hospitals in other states	Use DD estimators in both hospital level and patient level analyses. Find an adverse effect of report cards on outcomes, due to the patient selection they induce. The validity of the DD strategy is assessed by using a cohort of AMI patients who, as emergency admissions, should not be subject to patient selection.
Dusheiko et al. (2006)	Admissions to hospital (chargeable elective, non-chargeable elective & emergency)	Abolition of GP fundholding in England	(Ex-)fundholders/non-fundholders	Find that abolition of fundholding increased admissions
Finkelstein (2004)	Private insurance coverage for prescription drug expenditure	Medicare coverage	Those aged 63-64 at start of HRS panel survey, who became eligible for Medicare/(younger controls) those aged 60-62	Finds no evidence that Medicare is associated with levels of private coverage for drug expenditures

			who were never eligible, (older controls) those aged 65-67 who were eligible throughout	
Galiani et al. (2005)	Child mortality (general and cause specific)	Privatisation of local water services in Argentina	Privately provided water services/publically provided	Includes careful analysis of robustness of results to 'parallel' trends assumption, matching methods and use of cause-specific mortality to assess validity of findings
Jensen and Richter (2004)	Poverty, nutrition, use of medical care	1996 pensions crisis in Russia	In arrears/those who received pensions	
Liu et al. (2004)	Hospital length of stay and charges	Postpartum discharge laws in US states	States that enacted the laws/states that did not	Find heterogeneity in the effects depending on how the laws were implemented
Propper et al. (2002)	Waiting times for hospital admissions in England (North West Anglia region)	GP fundholding	Fundholders' patients/non-fundholders' patients	Effects on waiting only applied to a limited set of patients and of procedures
Schmidt (2007)	Infertility (first birth rates)	Infertility insurance mandates in US states	Uses interaction between time, whether states have a mandate and age of women (over 35).	Uses DDD estimates. Finds that mandates significantly increase birth rates for women over 35
Wagstaff and Yu (2007)	Use of services, catastrophic expenses, health outcomes	World Bank's Health VIII project, Gansu province, China	Project/non-project counties	DD combined with propensity score matching
Winkelmann (2004a)	Doctor visits	Increase in copayments for prescription drugs	Non-exempt/exempt	
Wolfe et al. (2006)	Public health care coverage of "welfare leavers"	Wisconsin BadgerCare Program (expanded public health insurance eligibility)	i) Between cohort strategy: 1997 cohort/1995 cohort of leavers ii) Within cohorts: newly eligible/continuously eligible	DD estimates suggest that BadgerCare increased public coverage by 17-25% points.

References

Aakvik, A., Heckman, J. J. and Vytlacil, E. J. (2005) Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *Journal of Econometrics* **125**, 15-51.

Aakvik, A. and Holmas, T. H. (2006) Access to primary health care and health outcomes: The relationship between GP characteristics and mortality rates. *Journal of Health Economics* **25**, 1139-53.

Aakvik, A., Holmas, T. H. and Kjerstad, E. (2003) A low-key social insurance reform-effects of multidisciplinary outpatient treatment for back pain patients in Norway. *Journal of Health Economics* **22**, 747-62.

Abadie, A. and Gay, S. (2006) The impact of presumed consent legislation on cadaveric organ donation: a cross-country study. *Journal of Health Economics* **25**, 599-620.

Adams, S. (2007) Health Insurance market reform and employee compensation: The case of pure community rating in New York *Journal of Public Economics* **91**, 1119 - 33.

Adda, J. and Cornaglia, F. (2006) Taxes, cigarette consumption and smoking intensity. *American Economic Review* **96**, 1013-28.

Agodini, R. and Dynarski, M. (2004) Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics* **86**, 180-94.

Ai, C. and Norton, E. C. (2000) Standard errors for the retransformation problem with heteroscedasticity. *Journal of Health Economics* **19**, 697-718.

Almond, D. (2006) Is the 1918 influenza pandemic over? Long term effects of in utero influence influenza exposure in the post 1940 US. *Journal of Political Economy* **114**, 672-712.

Almond, D., Chay, K. Y. and Lee, D. S. (2005) The costs of low birth weight. *Quarterly Journal of Economics* **120**, 1031-83.

Alvarez, B. and Delgado, M. A. (2002) Goodness-of-fit techniques for count data models: an application to the demand for dental care in Spain. *Empirical Economics* **27**, 543-67.

Arcidiacono, P. and Nicholson, S. (2005) Peer effects in medical school. *Journal of Public Economics* **89**, 327-50.

Arcidiancono, P., Sieg, H. and Sloan, F. (2007) Living Rationally under the Volcano? An empirical analysis of heavy drinking and smoking. *International Economic Review* **48**, 37-65.

Arellano, M. and Bond, S. (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations *Review of Economic Studies* **58**, 277 - 97.

Arellano, M. and Bover, O. (1995) Another look at the instrumental variables estimation of error component models. *Journal of Econometrics* **68**, 29 - 52.

Arendt, J. N. (2005) Does education cause better health? A panel data analysis using school reforms for identification. *Economics of Education Review* **24**, 149-60.

Arulampalam, W. and Bhalotra, S. (2006) Sibling death clustering in India: state dependence versus unobserved heterogeneity. *Journal of the Royal Statistical Society Series A* **169**, 829-48.

Arulampalam, W., Naylor, R. A. and Smith, J. P. (2004) A hazard model of the probability of medical school drop-out in the UK. *Journal of the Royal Statistical Society Series A* **167**, 157-78.

Askildsen, J. E., Baltagi, B. H. and Holmas, T. H. (2003) Wage policy in the health care sector: a panel data analysis of nurses' labour supply. *Health Economics* **12**, 705-19.

Askildsen, J. E., Bratberg, E. and Nilsen, O. A. (2005) Unemployment, labour force composition and sickness absence: a panel data study. *Health Economics* **14**, 1087-101.

Atella, V., Brindisi, F., Deb, P. and Rosati, F. C. (2004) Determinants of access to physician services in Italy: a latent class seemingly unrelated probit approach. *Health Economics* **13**, 657-68.

Atella, V., Peracchi, F., Depalo, D. and Rossetti, C. (2006) Drug compliance, co-payment and health outcomes: evidence from a panel of Italian patients. *Health Economics* **15**, 875-92.

Au, D. W. H., Crossley, T. F. and Schellhorn, M. (2005) The effect of health changes and long term health on the work activity of older Canadians. *Health Economics* **14**, 999-1018.

Auld, M. C. (2002) Disentangling the effects of morbidity and life expectancy on labor market outcomes. *Health Economics* **11**, 471-83.

Auld, M. C. (2006a) Estimating behavioral response to the AIDS epidemic. *The Berkeley Electronic Press* **5**.

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

Auld, M. C. and Grootendorst, P. (2004) An empirical analysis of milk addiction. *Journal of Health Economics* **23**, 1117-33.

Auld, M. C. and Sidhu, N. (2005) Schooling, cognitive ability and health. *Health Economics* **14**, 1019-34.

Bago d'Uva, T. (2005) Latent class models for use of primary care: evidence from a British panel. *Health Economics* **14**, 873-92.

Bago d'Uva, T. (2006) Latent class models for utilisation of health care. *Health Economics* **15**, 329-43.

Baker, M., Stabile, M. and Deri, C. (2004) What do self-reported, objective, measures of health measure? *Journal of Human Resources* **39**, 1067-93.

Balia, S. and Jones, A. M. (2007) Mortality, lifestyle and socioeconomics status. *Journal of Health Economics* **26,** in press.

Baltagi, B. H., Bratberg, E. and Holmas, T. H. (2005) A panel data study of physicians' labor supply: the case of Norway. *Health Economics* **14**, 1035-45.

Banks, J., Marmot, M., Oldfield, Z. and Smith, J. P. (2006) Disease and disadvantage in the United States and in England. *Journal of the American Medical Association* **295**, 2037-45.

Baser, O., Gardiner, J. C., Bradley, C. J., Yuce, H. and Given, C. (2006) Longitudinal analysis of censored medical cost data. *Health Economics* **15**, 513-25.

Basu, A., Arondekar, B. V. and Rathouz, P. J. (2006) Scale of interest versus scale of estimation: comparing alternative estimators for the incremental costs of a comorbidity. *Health Economics* **15**, 1091-107.

Basu, A., Heckman, J., Navarro, S. and Urzua, S. (2007) Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. *Health Economics* **16**, in press.

Basu, A., Manning, W. G. and Mullahy, J. (2004) Comparing alternative models: log vs Cox proportional hazard? *Health Economics* **13**, 749-65.

Basu, A. and Rathouz, P. J. (2005) Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* **6**, 93-109.

Becker, B., Grossman, M. and Murphy, K. (1994) An empirical analysis of cigarette addiction *American Economic Review* **84**, 396-418.

Behrman, J. R. and Rosenzweig, M. R. (2004) Returns to birthweight. *Review of Economics and Statistics* **86**, 586-601.

Benitez-Silva, H., Buchinsky, M., Chan, H. M., Cheidvasser, S. and Rust, J. (2004) How large is the bias in self reported disability. *Journal of Applied Econometrics* **19**, 649-70.

Bertrand, M., Duflo, E. and Mullainathan, S. (2004) How much should we trust differences-in-differences Estimates? *The Quarterly Journal of Economics* **119**, 249-75.

Bhattacharya, J., Goldman, D. and Sood, N. (2003) The link between public and private insurance and HIV-related mortality. *Journal of Health Economics* **22**, 1105-22.

Black, S., Devereux, P. and Salvanes, K. (2007) From the cradle to the labour market ? The effect of birth weight on adult outcomes. *The Quarterly Journal of Economics* **122**, 409 - 39.

Bleakley, H. (2007) Disease and development evidence from hookwork eradication in the Americal South *The Quarterly Journal of Economics* **122**, 73-117.

Blough, D. K., Madden, C. W. and Hornbrook, M. C. (1999) Modeling risk using generalized linear models. *Journal of Health Economics* **18**, 153-71.

Blundell, R. and Bond, S. (1998) Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* **87**, 115-43.

Bover, O. and Arellano, M. (1997) Estimating dynamic limited-dependent variable models from panel data *Investigaciones Economicas* **21**, 141 - 65.

Bradford, D. W., Kleit, A. N., Krousel-Wood, M. A. and Re, R. N. (2001) Stochastic frontier estimation of cost models within the hospital. *Review of Economics and Statistics* **83**, 300-08.

Bradley, C. J., Neumark, D., Bednarek, H. L. and Schenk, M. (2005) Short-term effects of breast cancer on labor market attachment: results from a longitudinal study. *Journal of Health Economics* **24**, 137-60.

Briggs, A. (2006) Statistical methods for cost-effectiveness analysis alongside clinical trials. In Jones, A.M. (ed) *Elgar Companion to Health Economics* Cheltenham: Edward Elgar.

Briggs, A., Nixon, R., Dixon, S. and Thompson, S. (2005) Parametric modelling of cost data: some simulation evidence. *Health Economics* **14**, 421-28.

Brown, T., Coffman, J., Quinn, B., Scheffler, R. and Schwalm, D. (2005) Do physicians always flee from HMO's? New results using dynamic panel estimation methods. *Health Services Research* **40**, 357-73.

Buntin, M. B. and Zaslavsky, A. M. (2004) Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics* **23**, 525-42.

Burgess, J. (2006) Productivity analysis in health care. In Jones, A.M. (ed) *Elgar Companion to Health Economics* Cheltenham: Edward Elgar.

Cantoni, E. and Ronchetti, E. (2006) A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures. *Journal of Health Economics* **25**, 198-213.

Carey, K. (2000) Hospital cost containment and length of stay: an econometric analysis. *Southern Economic Journal* **67**, 363 - 80.

Carpenter, C. (2004) How do zero tolerance drunk driving laws work? *Journal of Health Economics* **23**, 61-83.

Cawley, J., Grabowski, D. C. and Hirth, R. A. (2006) Factor substitution in nursing homes. *Journal of Health Economics* **25**, 234-47.

Chalkley, M. and Tilley, C. (2006) Treatment intensity and provider remuneration: dentists in the British national health service. *Health Economics* **15**, 933-46.

Chamberlain, G. (1980) Analysis of covariance with qualitative data. *Review of Economic Studies* **47**, 225 - 38.

Chamberlain, G. (1984) Panel data. In Griliches, Z. and Intriligator, M. (eds), *Handbook of Econometrics*. Amsterdam: North-Holland.

Chandra, A. and Staiger, D. (2007) Productivity spillovers in health care: evidence from the treatment of heart attacks. *Journal of Political Economy* **115**.

Chang, F.-R. and Trivedi, P. K. (2003) Economics of self medication: theory and evidence. *Health Economics* **12**, 721-39.

Chay, K. Y. and Greenstone, M. (2003) The impact of air polution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. *Quarterly Journal of Economics* **118**, 1121-67.

Chen, L., Yip, W., Chang, M., Lin, H., Lee, S., Chiu, Y. and Lin, Y. (2007) The Effects of Taiwan's National Health Insurance of access and health status of the elderly. *Health Economics* **26**, 223 - 42.

Chen, Y. and Zhou, L.-A. (2007) The long term health and economic consiquences of the 1959 - 1961 famine in China. *Journal of Health Economics* **26**, 659 - 81.

Chou, S.-Y. (2002) Asymmetric information, ownership and quality of care: an empirical analysis of nursing homes. *Journal of Health Economics* **21**, 293-311.

Chou, W. L. (2007) Explaining China's reginal health expenditures using LM-type unit root tests *Journal of Health Economics* **26**, 682 - 98.

Clark, A. and Etile, F. (2002) Do health changes affect smoking? Evidence from British panel data. *Journal of Health Economics* **21**, 533-62.

Clark, A., Etile, F., Postel-Vinay, F., Senik, C. and Van Der Straeten, K. (2005) Heterogeneity in Reported Well-being: Evidence from Twelve European Countries. *Economic Journal* **115**, C118-C32.

Clark, A. E. and Etile, F. (2006) Don't give up on me baby: spousal correlation in smoking behaviour. *Journal of Health Economics* **25**, 958-78.

Claxton, K., Fenwick, E. and Sculpher, M. J. (2006) Decision making with uncertainty: The value of information. In Jones, A.M. (ed) *Elgar Companion to Health Economics* Cheltenham: Edward Elgar.

Cockx, B. and Brasseur, C. (2003) The demand for physician services: evidence from a natural experiment. *Journal of Health Economics* **22**, 881-913.

Contoyannis, P., Hurley, J., Grootendorst, P., Jeon, S.-H. and Tamblyn, R. (2005) Estimating the price elasticity of expenditure for prescription drugs in the presence of non-linea price schedules: an illustration from Quebec, Canada. *Health Economics* **14**, 909-23.

Contoyannis, P., Jones, A. M. and Leon-Gonzalez, R. (2004a) Using simulation-based inference with panel data in health economics. *Health Economics* **13**, 101-22.

Contoyannis, P., Jones, A. M. and Rice, N. (2003) Simulation-based inference in dynamic panel probit models: An application to health. *Empirical Economics* **28**, 1-29.

Contoyannis, P., Jones, A. M. and Rice, N. (2004b) The dynamics of health in the british household panel survey. *Journal of Applied Econometrics* **19**, 473-503.

Contoyannis, P. and Rice, N. (2001) The impact of health on wages: evidence from the British Household Panel Survey. *Empirical Economics* **26**, 599-622.

Conway, K. S. and Deb, P. (2005) Is prenatal care really ineffective? Or, is the 'devil' in the distribution? *Journal of Health Economics* **24**, 489-513.

Cooper, N. J., Lambert, P. C., Abrams, K. R. and Sutton, A. J. (2007) Predicting costs over time using Bayesian Markov chain Monte Carlo methods: an application to early inflammatory polyarthritis. *Health Economics* **16**, 37-56.

Cowell, A. J. (2006) The relationship between education and health behaviour: some empirical evidence. *Health Economics* **15**, 125-46.
Currie, A., Shields, M., Price, S. and Wheatley (2007) The child health / family income gradient: evidence from England. *Journal of Health Economics* **26**, 213 - 32.
Currie, J. and Hotz, V. J. (2004) Accidents will happen? Unintentional childhood injuries and the effects of child care regulations. *Journal of Health Economics* **23**, 1-215.
Currie, J. and Madrian, B. (1999) Health, health insurance and the labour market. In Ashenfelter, O. and Card, D. (eds.), *Handbook of Labor Economics*. Vol 3C. Amsterdam: Elsevier
Currie, J. and Neidell, M. (2005) Air pollution and infant health: what can we learn from California's recent experience. *Quarterly Journal of Economics* **120**, 1003-30.
Currie, J. and Stabile, M. (2006) Child mental health and human capital accumulation: the case of ADHD. *Journal of Health Economics* **25**, 1094-118.
Dano, A. M. (2005) Road injuries and long run effects on income and employment. *Health Economics* **14**, 955-670.
Das, J. and Hammer, J. (2005) Which doctor? Combining vignettes and item response to measure clinical competence. *Journal of Development Economics* **78**, 348 - 83
Das, M., and van Soest, A. (1999) A panel data model for subjective information on household income growth. *Journal of Economic Behaviour and Organization* **40**, 409-26.
Davidoff, A., Blumberg, L. and Nichols, L. (2005) State health insurance market reforms and access to insurance for high-risk employees. *Journal of Health Economics* **24**, 725-50.
Dawson, D., Gravelle, H., Jacobs, R., Martin, S. and Smith, P. (2007) The effects of expanding patient choice of provider on waiting times: evidence from a policy experiment *Health Economics* **16**, 113 - 28
Deb, P. (2001) A discrete random effects probit model with application to the demand for preventive care. *Health Economics* **10**, 371-83.
Deb, P. and Holmes, A. M. (2000) Estimates of use and costs of behavioural health care: a comparison of standard and finite mixture models. *Health Economics* **9**, 475-89.
Deb, P., Li, C., Trivedi, P. K. and Zimmer, D. M. (2006a) The effect of managed care on use of health care services: results from two contemporaneous household surveys. *Health Economics* **15**, 743-60.
Deb, P., Munkin, M. K. and Trivedi, P. K. (2006b) Bayesian analysis of the two-part model with endogeneity: application to health care expenditure. *Journal of Applied Econometrics* **21**, 1081-99.
Deb, P. and Trivedi, P. K. (1997) Demand for medical care by the elderly: a finite mixture approach. *Journal of Applied Econometrics* **12**, 313-36.
Deb, P. and Trivedi, P. K. (2002) The structure of demand for health care: latent class versus two-part models. *Journal of Health Economics* **21**, 601-25.
Deb, P. and Trivedi, P. K. (2006) Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: application to health care utilization. *Econometrics Journal* **9**, 307-31.
Deb, P. and Trivedi, P. K. (2006a) Restrictions on provider access in health plans and socioeconomic status. *Health Services Research* **41**, 759-87.
Dee, T. S., Grabowski, D. C. and Morrisey, M. A. (2005) Graduated driver licensing and teen traffic fatalities. *Journal of Health Economics* **24**, 571-89.
Disney, R., Emmerson, C. and Wakefield, M. (2006) Ill health and retirement in Britain: a panel data-based analysis. *Journal of Health Economics* **25**, 621-49.

Dor, A., Sudano, J. and Baker, D. (2006) The effect of private insurance on the health of older, working age adults: evidence from the health and retirement study *Health Services Research* **41**, 759-787.

Doyle, J. J. (2005) Health insurance, treatment and outcomes: using auto accidents as health shocks. *Review of Economics and Statistics* **87**, 256-70.

Dranove, D., Kessler, D., McClellan, M. and Satterthwaite, M. (2003) Is more information better? The effects of 'report cards' on health care providers. *Journal of Political Economy* **111**, 555-88.

Dranove, D. and Lindrooth, R. (2003) Hospital consolidation and costs: another look at the evidence. *Journal of Health Economics* **22**, 983-97.

Dranove, D. and Wehner, P. (1994) Physician-induced demand for childbirths. *Journal of Health Economics* **13**, 61-73.

Dubay, L., Kaestner, R. and Waidmann, T. (2001) Medical malpractice liability and its effects on prenatal care utilization and infant health. *Journal of Health Economics* **20**, 591-611.

Duflo, E. (2000) Child health and household resources in South Africa: evidence from the old age pension program. *American Economic Review* **90**, 393 - 98.

Dusheiko, M., Gravelle, H. and Jacobs, R. (2004) The effect of practice budgets on patient waiting times: allowing for selection bias. *Health Economics* **13**, 941-58.

Dusheiko, M., Gravelle, H., Jacobs, R. and Smith, P. (2006) The effect of financial incentives on gatekeeping doctors: evidence from a natural experiment. *Journal of Health Economics* **25**, 449-78.

Dusheiko, M., Gravelle, H., Yu, N. and Campbell, S. (2007) The impact of budgets for gatekeeping physicians on patient satisfaction: evidence from fundholding. *Journal of Health Economics* **26**, 742 - 62.

Elliott, R. F., Ma, A. H. Y., Scott, A., Bell, D. and Roberts, E. (2007) Geographically differentiated pay in the labour market for nurses. *Journal of Health Economics* **26**, 190-212.

Etilé, F. (2006) Who does the hat fit? Teenager heterogeneity and the effectiveness of information policies in preventing cannabis use and heavy drinking. *Health Economics* **15**, 697-718.

Etilé, F. and Milcent, C. (2006) Income-related reporting heterogeneity in self assessed health: evidence from France. *Health Economics* **15**, 965-81.

Evans, W. N. and Lien, D. S. (2005) The benefits of prenatal care: evidence from the PAT bus strike. *Journal of Econometrics* **125**, 207-39.

Farsi, M. and Ridder, G. (2006) Estimating the out-of-hospital mortality rate using patient discharge data. *Health Economics* **15**, 983-95.

Ferrer-i-Carbonell, A. and Frijters, P. (2004) How important is methodology for the estimates of the determinants of happiness. *The Economic Journal* **114**, 641-59.

Finkelstein, A. (2004) The interaction of partial public insurance programs and residual private insurance markets: evidence from the US Medicare program. *Journal of Health Economics* **23**, 1-24.

Forster, M. and Jones, A. M. (2001) The role of tobacco taxes in starting and quitting smoking: duration analysis of British data. *Journal of the Royal Statistical Society Series A* **3**, 517-47.

French, E. (2005) The effects of health, wealth and wages on labour supply and retirement behavior *Review of Economic Studies* **72**, 395 - 427.

Frijters, P., Haisken-Denew, J. P. and Shields, M. A. (2005) The causal effect of income on health: evidence from German reunification. *Journal of Health Economics* **24**, 997-1017.

Frijters, P., Shields, M. A. and Price, S. W. (2006) Investigating the quitting decision of nurses: panel data evidence from the british national health service. *Health Economics* **16**, 57-73.

Frolich, M., Heshmati, A. and Lechner, M. (2004) A microeconometric evaluation of rehabilitation of long-term sickness in Sweden. *Journal of Applied Econometrics* **19**, 375-96.

Galiani, S., Gertler , P. and E, S. (2005) Water for life: the impact of the privatization of water services on child morality *Journal of Political Economy* **113**, 83-120.

Gallet, C. A. and List, J. A. (2003) Cigarette demand: a meta-analysis of elasticities. *Health Economics* **12**, 821-53.

Gannon, B. (2005) A dynamic analysis of disability and labour force participation in Ireland 1995-2000. *Health Economics* **14**, 925-38.

García-Ferrer, A., Juan, A. D. and Poncela, P. (2007) The relationship between road traffic accidents and real economic activity in Spain: common cycles and health issues *Health Economics* **16**, 603 - 26.

García-Gómez, P. and López-Nicolás, A. (2006) Health shocks, employment and income in the Spanish labour market. *Health Economics* **15**, 997-1009.

Gardner, J. and Oswald, A. J. (2007) Money and mental wellbeing: a longitudinal study of medium-sized lottery wins. *Journal of Health Economics* **26**, 49-60.

Geil, P., Million, A., Rotte, R. and Zimmerman, K. F. (1997) Economic incentives and hospitalization in Germany. *Applied Economics* **12**, 295-311.

Gemmill, M., Costa-Font, J. and Mcguire, A. (2007) In search of a corrected prescription drug elasticity estimate: a meta-regression approach. *Health Economics* **16**, 627 - 43.

Gertler, P. (2004) Do conditional cash transfers improve child health? Evidence from PROGRESA's control randomized experiment. *American Economic Review* **94**, 336-41.

Geweke, J., Gowrisankaran, G. and Town, R. J. (2003) Bayesian inference for hospital quality in a selection model. *Econometrica* **71**, 1215-38.

Gilleskie, D. B. and Mroz, T. A. (2004) A flexible approach for estimating the effects of covariates on health expenditures. *Journal of Health Economics* **23**, 391-418.

Girma, S. and Paton, D. (2006) Matching estimates of the impact of over the counter emergency birth control on teenage pregnancy. *Health Economics* **15**, 1021-32.

Gravelle, H., Sutton, M., Morris, S., Windmeijer, F., Leyland, A., Dibben, C. and Muirhead, M. (2003) Modelling supply and demand influences on the use of health care: implications for deriving a needs based capitation formula. *Health Economics* **12**, 985-1004.

Groot, W. (2000) Adaptation and scale of refrence bias in self-assessments of quality of life *Journal of Health Economics* **19**, 403 - 20.

Gruber, J. and Owings, M. (1996) Physician financial incentives and caesarean section delivery. *RAND Journal of Economics* **57**, 99-123.

Guariglia, A. and Rossi, M. (2004) Private medical insurance and saving: evidence from the British Household Panel Survey. *Journal of Health Economics* **23**, 761-83.

Hadley, J., Polsky, D., Mandelblatt, J. S., Mitchell, J. M., Weeks, J. C., Wang, Q. and Hwang, Y.-T. (2003) An exploratory instumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. *Health Economics* **12**, 171-86.

Hamilton, B. H., Vivian, H. and Goldman, D. P. (2000) Queuing for surgery: is the US or Canada worse off? *Review of Economics and Statistics* **82**, 297-308.
Harris, M. N., Ramful, P. and Zhao, X. (2006) An ordered generalised extreme value model with application to alchohol consumption in Australia. *Journal of Health Economics* **25**, 782-801.
Harrison, T. (2007) Consolidations and closures: an empirical analysis of exits from the hospital industry *Health Economics* **16**, 457 - 74.
Hauck, K. and Rice, N. (2004) A longitudinal analysis of mental health mobility in Britain. *Health Economics* **13**, 981-1001.
Heckman, J. and Singer, B. (1984) A method of minimizing the distributional impact in econometric models for duration data. *Econometrica* **52**, 271-30.
Heckman, J. and Vytlacil, E. (1999) Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences* **96**, 4730-34.
Heckman, J. and Vytlacil, E. (2007) Econometric evaluation of social programs. In Engle, R. and McFadden, D. (eds), *Handbook of Econometrics Vol.6*. Amsterdam: Elsevier:
Ho, K. (2006) The welfare effects of restricted hospital choice in the US medical care market. *Journal of Applied Econometrics* **21**, 1039-79.
Ho, V. (2002) Learning and the evolution of medical technologies: the diffusion of coronary angioplasty. *Journal of Health Economics* **21**, 873-85.
Hoch, J. S., Briggs, A. H. and Willan, A. R. (2002) Something old, something new, something borrowed, something blue: a framework for the marriage of health econometrics and cost-effectiveness analysis. *Health Economics* **11**, 415-30.
Hogelund, J. and Holm, A. (2006) Case management interviews and the return to work of disabled employees. *Journal of Health Economics* **25**, 500-19.
Holmas, T. H. (2002) Keeping nurses at work: a duration analysis. *Health Economics* **11**, 493-503.
Holmes, G. (2005) Increasing physician supply in medically underserved areas. *Labour Economics* **12**, 697 - 725.
Holmlund, H. (2005) Estimating long-term consequences of teenage childbearing. *Journal of Human Resources* **40**, 716-43.
Houtven, C. H. V. and Norton, E. C. (2004) Informal care and health care use of older adults. *Journal of Health Economics* **23**, 1159-80.
Imbens, G. W. and Angrist, J. (1994) Identification and estimation of local average treatment effects. . *Econometrica* **62**, 467-75.
Jalan, J. and Ravallion, M. (2003) Does piped water reduce diarrhea for children in rural India? *Journal of Econometrics* **112**, 153-73.
Jensen, R. T. and Richter, K. (2004) The health implications of social security failure: evidence from the Russian pension crisis. *Journal of Public Economics* **88**, 209-36.
Jewell, R. T. and Triunfo, P. (2006) The impact of prenatal care on birthweight: the case of Uruguay. *Health Economics* **15**, 1245-50.
Jiménez-Martin, S., Labeaga, J. M. and Martínez-Granado, M. (2002) Latent class versus two-part models in the demand for physician services across the European Union. *Health Economics* **11**, 301-21.
Jochmann, M. and Leon-Gonzalez, R. (2004) Estimating the demand for health care with panel data: a semiparametric Bayesian approach. *Health Economics* **13**, 1003-14.
Jones, A. M. (2000) Health Econometrics. In Culyer, A. J. and Newhouse, J. P. (eds), *Handbook of Health Economics*. Amsterdam: Elsevier.

Jones, A. M., Koolman, X. and Doorslaer, E. V. (2007a) The impact of supplementary private health insurance on the use of specialists in European countries. *Annales d'Economie et de Statistiques* **83/84**, 251-75.

Jones, A. M., Koolman, X. and Rice, N. (2006) Health-related non-response in the British Household Panel Survey and European Community Household Panel: using inverse-probability-weighted estimators in non-linear models. *Journal of the Royal Statistical Society Series A* **169**, 543-69.

Jones, A. M. and Labeaga, J. M. (2003) Individual heterogeneity and censoring in panel data estimates of tobacco expenditure. *Journal of Applied Econometrics* **18**, 157-77.

Jones, A. M., Rice, N., D'uva, T. B. and Balia, S. (2007b) *Applied Health Economics*. London: Routledge.

Jurges, H. (2007) True health vs response styles: exploring cross-country diffences in self-reported health *Health Economics* **16**, 163 - 78.

Kan, H., Goldman, D., Keeler, E., Dhanani, N. and Melnick, G. (2003) An analysis of unobserved selcetion in an inpatient diagnostic cost group model. *Health Services & Outcomes Research Methodology* **4**, 71 - 91.

Kan, K. and Tsai, W.-D. (2004) Obesity and risk knowledge. *Journal of Health Economics* **23**, 907-34.

Kerkhofs, M. and Lindeboom, M. (1995) Subjective health measures and state dependednt reporting errors. *Health Economics* **4**, 221 - 35.

Kessler, D. P. and Mcclellan, M. B. (2002) How liability law affects medical productivity. *Journal of Health Economics* **21**, 931-55.

Kremer, M. (2003) Randomized evaluations of educational programs in developing countries: some lessons. *American Economic Review* **93**, 102-06.

Kyle, M. (2007) Pharmaceutical Price Controls and Entry Strategies. *The Review of Economics and Statistics* **89**, 88 - 99.

Kyriazidou, E. (1997) Estimation of a panel data sample selection model. *Econometrica* **65**, 1335-64.

Lakdawalla, D., Sood, N. and Goldman, D. (2006) HIV breakthroughs and risky sexual behavior. *Quarterly Journal of Economics* **121**, 1063-102.

Lalonde, R. (1986) Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* **76**, 604-20.

Lee, L. F. (1983) Generalised econometric models with selectivity. *Econometrica* **51**, 507-12.

Lee, M.-C. and Jones, A. M. (2004) How did dentists respond to the introduction of global budgets in Taiwan? An evaluation using individual panel data. *International Journal of Health Care Finance and Economics* **4**, 307-26.

Lee, M.-C. and Jones, A. M. (2006) Heterogeneity in dentists' activity in Taiwan: an application of quantile regression. *Empirical Economics* **31**, 151-64.

Leigh, A. and Jencks, C. (2007) Inequality and mortality: long-run evidence from a panel of countries. *Journal of Health Economics* **26**, 1-24.

Levitt, S. D. and Porter, J. (2001) Sample selection in the estimation of air bag and seat belt effectiveness. *The Review of Economics and Statistics* **83**, 603-15.

Lindahl, M. (2005) Estimating the effect of income on health and mortality using lottery prizes as exogenous source of variation in income. *Journal of Human Resources* **40**, 144-68.

Lindeboom, M. (2006) Health and work of older workers In Jones, A. M. (ed), *Elgar Companion to Health Economics*. Cheltenham: Edward Elgar.

Lindeboom, M., Portrait, F. and Berg, G. J. V. D. (2002) An econometric analysis of the mental health effects of major events in the life of older individuals. *Health Economics* **11**, 505-20.

Lindeboom, M. and Van Doorslaer, E. (2004) Cut-point shift and index shift in self-reported health. *Journal of Health Economics* **23**, 1083-99.

Lindrooth, R. and Weisbrod, B. (2007) Do religious non-profit and for-profit organisations respond diffrently to finacial incentives? The hospice industry *Journal of Health Economics* **26**, 342 - 57.

Liu, Z., Dow, W. H. and Norton, E. C. (2004) Effect of drive-through delivery laws on postpartum length of stay and hospital charges. *Journal of Health Economics* **23**, 129-55.

Lleras-Muney, A. (2005) The relationship between education and adult mortality in the United States *Review of Economic Studies* **72**, 189-221.

Lourenço, O. D. and Ferreira, P. L. (2005) Utilization of public health centres in Portugal: effect of time costs and other determinants. Finite mixture models applied to truncated samples. *Health Economics* **14**, 939-53.

Machado, J. A. F. and Santos Silva, J. M. C. (2005) Quantiles for counts. *Journal of the American Statistical Association* **100**, 1226-37.

Manning, W. (1998) The logged dependent variable, heteroscedasticity, and the retransformation problem. *Journal of Health Economics* **17**, 283-95.

Manning, W. (2006) Dealing with skewed data on costs and expenditure. In Jones, A.M. (ed) *Elgar Companion to Health Economics* Cheltenham: Edward Elgar.

Manning, W., Newhouse, J. P., Duan, N., Keeler, E., Leibowitz, A. and Marquis, M. S. (1987) Health insurance and the demand for medical care: evidence from a randomized experiment. *American Economic Review* **77**, 251-77.

Manning, W. G., Basu, A. and Mullahy, J. (2005) Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* **24**, 465-88.

Manning, W. G. and Mullahy, J. (2001) Estimating log models: to transform or not to transform? *Journal of Health Economics* **20**, 461-94.

Martin, S., Rice, N., Jacobs, R. and Smith, P. (2007) The market for elective surgery: joint estimation of supply and demand *Journal of Health Economics* **26**, 263 - 85

McClellan, M., Newhouse, J. P. and Mcneil, B. (1994) Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? *Journal of the American Medical Association* **272**, 859-66.

McGarry, K. (2004) Health and retirement. *Journal of Human Resources* **39**, 624-48.

Meer, J., Miller, D. L. and Rosen, H. S. (2003) Exploring the health-wealth nexus. *Journal of Health Economics* **22**, 713-30.

Mello, M. M., Stearns, S. C. and Norton, E. C. (2002) Do medicare HMOs still reduce health services use after controlling for selection bias? *Health Economics* **11**, 323-40.

Miguel, E. and Kremer, M. (2004) Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica* **72**, 159-217.

Morris, S. (2006) Body mass index and occupational attainment. *Journal of Health Economics* **25**, 347-64.

Morris, S. (2007) The impact of obesity on employment *Labour Economics* **14**, 413-33.

Moscone, F., Knapp, M. and Tosetti, E. (2007) Mental health expenditure in England: A spatial panel approach. *Journal of Health Economics* **26**, 842 - 64.

Mroz, T. A. (1999) Discrete factor approximations in simultaneous equation models: Estimating the impact of a dummy endogenous variable on a continuous outcome. *Journal of Econometrics* **92**, 233-74.

Mullahy, J. (1998) Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics* **17**, 247-81.

Mundlak, Y. (1978) On the pooling of time series and cross section data. *Econometrica* **46**, 69-85.

Munkin, M. K. and Trivedi, P. K. (1999) Simulated maximum likelihood estimation of multivariate mixed-poisson regression models, with application. *Econometrics Journal* **2**, 29-48.

Murray, C. J. L., Tandon, A., Salomon, J. and Mathers, C. D. (2001) Enhancing cross-population comparability of survey results. GPE Discussion Paper Nr 35, WHO/EIP, World Health Organisation, Geneva.

Nicoletti, C. and Peracchi, F. (2005) Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society Series A* **168**, 763 - 81.

Nolan, A. (2007) A dynamic analysis of GP visiting in Ireland 1995 - 2001. *Health Economics* **16**, 129 - 43.

Or, Z., Wang, J. and Jamison, D. (2005) International differences in the impact of doctors on health: a multilevel analysis of OECD countries. *Journal of Health Economics* **24**, 531-60.

Paton, D. (2002) The economics of family planning and underage conceptions. *Journal of Health Economics* **21**, 207-25.

Picone, G., Wilson, R. M. and Chou, S.-Y. (2003a) Analysis of hospital length of stay and discharge destination using hazard functions with unmeasured heterogeneity. *Health Economics* **12**, 1021-34.

Picone, G. A., Sloan, F. A., Chou, S.-Y. and Jr, D. H. T. (2003b) Does higher hospital cost imply higher quality of care? *Review of Economics and Statistics* **85**, 51-62.

Pop-Eleches, C. (2006) The impact of an abortion ban on socioeconomic outcomes of children: evidence from Romania. *Journal of Political Economy* **114**, 744-73.

Powell, J. (1986) Symmetrically trimmed least squares estimators for Tobit models. *Econometrica* **54**, 1435 - 60.

Prieger, J. E. (2002) A flexible parametric selection model for non-normal data with application to health care usage. *Journal of Applied Econometrics* **17**, 367-92.

Propper, C., Burgess, S. and Green, K. (2004) Does competition between hospitals improve the quality of care: hospital death rates and the NHS internal market. *Journal of Public Economics* **88**, 1247-72.

Propper, C., Croxson, B. and Shearer, A. (2002) Waiting times for hospital admissions: the impact of GP fundholding. *Journal of Health Economics* **21**, 227-52.

Propper, C., Eachus, J., Chan, P., Pearson, N. and Smith, G. D. (2005) Access to health care resources in the UK: the case of care for arthritis. *Health Economics* **14**, 391-406.

Pudney, S. and Shields, M. (2000a) Gender, race, pay and promotion in the British nursing profession: estimation of a generalized ordered probit model. *Journal of Applied Econometrics* **15**, 367-99.

Pudney, S. and Shields, M. A. (2000b) Gender and racial discrimination in pay and promotion for NHS nurses. *Oxford Bulletin of Economics and Statistics* **62**, 801-35.

Quinn, C. (2005) Generalisable regression methods for cost-effectiveness using copulas. *Health, Econometrics and Data Group Working Paper WP 05/13, University of York* York

Raikou, M. and Mcguire, A. (2004) Estimating medical care costs under conditions of censoring. *Journal of Health Economics* **23**, 443-70.

Raikou, M. and Mcguire, A. (2006) Estimating costs for economic evaluation. In Jones, A.M. (ed) *Elgar Companion to Health Economics* Cheltenham: Edward Elgar.

Rettenmaier, A. J. and Wang, Z. (2006) Persistence in Medicare reimbursements and personal medical accounts. *Journal of Health Economics* **25**, 39-57.

Rice, N., Dixon, P., Lloyd, D. and Roberts, D. (2000) Derivation of a needs based capitation formula of allocating prescribing budgets to health authorities and primary care groups in England: regression analysis. *BMJ - British Medical Journal* **320**, 284 -88.

Riphahn, R. T., Wambach, A. and Million, A. (2003) Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics* **18**, 387-405.

Robinson, P. (1998) Root-N-consistent semiparametric regression *Econometrica* **56**, 931-54.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41-55.

Rous, J. J. and Hotchkiss, D. R. (2003) Estimation of the determinants of household health care expenditures in Nepal with controls for endogenous illness and provider choice. *Health Economics* **12**, 431-51.

Royalty, A. B. and Abraham, J. M. (2006) Health insurance and labor market outcomes: joint decision making within households. *Journal of Public Economics* **90**, 1561-77.

Ruhm, C. J. (2003) Good times make you sick. *Journal of Health Economics* **22**, 637-58.

Ryan, M., Gerard, K. and Currie, G. (2006) Using discrete choice experiments in health economics. In Jones, A.M. (ed) *Elgar Companion to Health Economics* Cheltenham: Edward Elgar.

Sadana, R., Mathers, C. D., Lopez, A. D., Murray, C. J. L. and Iburg, K. (2000) Comparative analysis of more than 50 household surveys on health status. GPE Discussion Paper No 15, EIP/GPE/EBD, World Health Organisation, Geneva.

Sahn, D. E., Younger, S. D. and Genicot, G. (2003) The demand for health care services in rural Tanzania. *Oxford Bulletin of Economics and Statistics* **65**, 241-60.

Santos Silva, J. M. C. and Windmeijer, F. (2001) Two-part multiple spell models for health care demand. *Journal of Econometrics* **104**, 67-89.

Sarma, S. and Simpson, W. (2006) A microeconometric analysis of Canadian health care utilization. *Health Economics* **15**, 219-39.

Sasso, A. T. L. and Buchmueller, T. C. (2004) The effect of the state children's health insurance program on health insurance coverage. *Journal of Health Economics* **23**, 1059-82.

Schellhorn, M. (2001) The effect of variable health insurance deductibles on the demand for physician visits. *Health Economics* **10**, 441-56.

Schmidt, L. (2007) Effects of infertility insurance mandates on fertility *Journal of Health Economics* **26**, 413 - 46.

Seshamani, M. and Gray, A. (2004) Ageing and health care expenditure: the red herring argument revisited. *Health Economics* **13**, 303-14.

Sloan, F. A., Picone, G. A., Taylor, D. H. and Chou, S.-Y. (2001) Hospital ownership and cost and quality of care: is there a dime's worth of difference? *Journal of Health Economics* **20**, 1-21.

Smith, J. A. and Todd, P. E. (2001) Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review* **91**, 112-18.

Smith, M. D. (2003) Modelling sample selection using Archimedean copulas. *Econometrics Journal* **6**, 99-123.

Smith, P. C. and Street, A. (2005) Measuring the efficiency of public services: the limits of analysis. *Journal of the Royal Statistical Society Series A* **168**, 401-17.

Stewart, J. M. (2001) The impact of health status on the duration of unemployment spells and the implications for studies of the impact of unemployment on health status. *Journal of Health Economics* **20**, 781-96.
Tamm, M., Tauchmann, H., Wasem, J. and Gress, S. (2007) Elasticities of market shares and social health insurance choice in german: a dynamic panel data approach. *Health Economics* **16**, 243 - 56.
Terza, J. V. (2002) Alcohol abuse and employment: a second look. *Journal of Applied Econometrics* **17**, 393-404.
Trivedi, P. K. and Zimmer, D. M. (2005) Copula modelling: an introduction for practitioners. *Foundations and Trends in Econometrics* **1**, 1-111.
Van den Berg, G. J., Lindeboom, M. and Portrait, F. (2006) Economic conditions early in life and individual mortality. *American Economic Review* **96**, 290-302.
Van Ours, J. C. (2004) A pint a day raises a man's pay; but smoking blows that gain away. *Journal of Health Economics* **23**, 863-86.
Van Ours, J. C. (2006) Dynamics in the use of drugs. *Health Economics* **15**, 1283-94.
Van Ourti, T. (2004) Measuring horizontal inequity in Belgian health care using a Gaussian random effects two part count data model. *Health Economics* **13**, 705-24.
Vera-Hernandez, M. (2003) Structural estimation of a principal agent model: moral hazard in medical insurance. *The RAND Journal of Economics* **34**, 670-93.
Wagstaff, A. (2007) The economic consequences of health shocks: evidence from Vietnam. *Journal of Health Economics* **26**, 82-100.
Wagstaff, A. and Yu, S. (2007) Do health sector reforms have their intended impacts? The World Bank's health VIII project in Gansu province, China *Journal of Health Economics* **26**, 505 - 35.
Wang, Z. and Rettenmaier, A. (2007) A note on cointegration of health expenditures and income. *Health Economics* **16**, 599 - 78.
Wildman, J. and Jones, A. M. (2007) Disentangling the relationship between health and income. *Journal of Health Economics*, in press.
Willan, A. R., Briggs, A. H. and Hoch, J. S. (2004) Regression methods for covariate adjustment and subgroup analysis for non-censored cost effectiveness data. *Health Economics* **13**, 461-75.
Wilson, P. W. and Carey, K. (2004) Nonparametric analysis of returns to scale in the US hospital industry. *Journal of Applied Econometrics* **19**, 505-24.
Windmeijer, F., Gravelle, H. and Hoonhout, P. (2005) Waiting lists, waiting times and admissions: an empirical analysis at hospital and general practice levels. *Health Economics* **14**, 971-85.
Winkelmann, R. (2004) Health care reform and the number of doctor visits - an economic analysis. *Journal of Applied Econometrics* **19**, 455-72.
Winkelmann, R. (2004a) Co-payments for prescription drugs and the demand for doctor visits - evidence from a natural experiment. *Health Economics* **13**, 1081-89.
Winkelmann, R. (2006) Reforming health care: evidence from quantile regressions for counts. *Journal of Health Economics* **25**, 131-45.
Wolfe, B., Kaplan, T., Haveman, R. and Cho, Y. (2006) SCHIP expansion and parental coverage: an evaluation of Wisconsin's BadgerCare. *Journal of Health Economics* **25**, 1170-92.
Wooldridge, J. (2005) Simple solutions to the initial conditions problem in dynamic nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics* **20**, 39-54.

Yelowitz, A. S. (2000) Public policy and health insurance choices of the elderly: evidence from the medicare buy-in program. *Journal of Public Economics* **78**, 301-24.

Yen, S. T., Tang, C.-H. and Su, S.-J. B. (2001) Demand for traditional medicine in Taiwan: a mixed Gaussian-Poisson model approach. *Health Economics* **10**, 221-32.

Zimmer, D. M. and Trivedi, P. K. (2006) Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business & Economic Statistics* **24**, 63-76.