DISCUSSION PAPER SERIES

IZA DP No. 2251

Sibling Death Clustering in India: State Dependence vs. Unobserved Heterogeneity

Wiji Arulampalam
Sonia Bhalotra

August 2006

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Sibling Death Clustering in India: State Dependence vs. Unobserved Heterogeneity

Wiji Arulampalam
University of Warwick
and IZA Bonn

Sonia Bhalotra
University of Bristol,
CSAE, QEH and CHILD

Discussion Paper No. 2251
August 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 2251
August 2006

ABSTRACT

Sibling Death Clustering in India: State Dependence vs. Unobserved Heterogeneity*

Data from a range of different environments indicate that the incidence of death is not randomly distributed across families but, rather, that there is a clustering of death amongst siblings. A natural explanation of this would be that there are (observed or unobserved) differences across families, for example in genetic frailty, education or living standards. Another hypothesis of considerable interest for both theory and policy is that there is a causal process whereby the death of a child influences the risk of death of the succeeding child in the family. Drawing language from the literature on the economics of unemployment, the causal effect is referred to here as state dependence (or scarring). This paper investigates the extent of state dependence in India, distinguishing this from family-level risk factors common to siblings. It offers a number of methodological innovations upon previous research. Estimates are obtained for each of three Indian states, which exhibit dramatic differences in socio-economic and demographic variables. The results suggest a significant degree of state dependence in each of the three regions. Eliminating scarring, it is estimated, would reduce the incidence of infant mortality (among children born after the first child) by 9.8% in the state of Uttar Pradesh, 6.0% in West Bengal and 5.9% in Kerala.

JEL Classification: J1, C1, I1, O1

Keywords: death clustering, infant mortality, state dependence, scarring, unobserved heterogeneity, dynamic random effects logit, multi-level model, India

Corresponding author:

Wiji Arulampalam
Department of Economics
University of Warwick
Coventry, CV4 7AL
United Kingdom
E-mail: wiji.arulampalam@warwick.ac.uk

* We are grateful to Marcel Fafchamps, Norman Ireland, Ian Preston, Arthur van Soest, Richard Smith, Helene Turon, Mike Veall and to two anonymous referees and an editor of the journal for helpful comments, and to Alfonso Miranda for help with the Stata programming. We have benefited from presentations at the Royal Economic Society Conference (Warwick, 2003), the ESRC Econometric Study Group Conference (Bristol, 2004) and at the Universities of Bristol, Oxford, McMaster, Toronto, Essex, Tilburg and Southampton. We are grateful to ORC Macro International for providing us with the data. This organisation bears no responsibility for the analysis or interpretations presented in this paper. An earlier version of the paper was circulated under a slightly different title: Sibling death clustering in India: genuine scarring vs unobserved heterogeneity.

1. INTRODUCTION

Data from a range of different environments indicate that the incidence of childhood death is not randomly distributed across families but, rather, that there is a positive association of sibling deaths- or death clustering (see, for example, Zenger (1993), Guo (1993), Curtis et al (1993), Miller et al (1992), DasGupta (1990), Bean *et al* (1988) and Hobcraft *et al* (1985)). A natural explanation of this would be that families in which child deaths are concentrated are poorer or share genetic or environmental risk factors that predispose all of their children to higher death risks. In other words, families are different or there is inter-family heterogeneity in risk. To the extent that these differences are observed (e.g. maternal education), they can be captured by including these variables as regressors in a model of child mortality. Recent research has gone further in incorporating unobservable heterogeneity (like genetic traits or maternal ability) by allowing for a family-level random effect. In this paper we investigate whether, in *addition* to the positive correlation of sibling deaths arising from shared traits, there is a *causal* process set off by the event of death of a child that results in an elevation of the risk of death of the next child in the family. Borrowing language from the literature on the economics of unemployment, this causal process is called state dependence or scarring. This is formally defined in section 3. Heuristically, the idea is that the death of one child "scars" the family, making the next child in that family more vulnerable. In this paper, we analyse infant mortality, that is, mortality in the first year of life. This definition applies to both the index child and its preceding sibling.

So as to clarify the notion of causality, consider what mechanisms might drive scarring

effects. One causal mechanism is that which operates by the death of an infant shortening the time to the next birth. As it can take up to 24 months for the mother to recuperate physiologically from a birth, a short preceding birth interval for the index child elevates this child's mortality risk. For previous evidence of such effects, see, for instance, Hobcraft *et al* (1983), Cleland and Sathar (1984), Koenig *et al* (1990), Gribble (1993), Zenger (1993), Nath *et al* (1994). The reason that it can take a mother time to recuperate from a previous pregnancy, especially if she is under-nourished, is that a new pregnancy requires replenishment of vital nutrients like calcium and iron that are needed to support foetal development (e.g. DaVanzo and Pebley 1993). The process by which a child death leads to a shorter birth interval may operate in either of two ways. One possibility is that the death of an infant results in the mother ceasing to breastfeed and, thereby, being able to conceive sooner than otherwise (see Bongaarts and Potter (1983), Cantrelle *et al* (1978), Chen *et al* (1974).) Henceforth, this is referred to as *the fecundity hypothesis*. An alternative possibility is that the death of a child leads parents to (intentionally) conceive sooner in a desire to "replace" their loss (e.g. Preston 1985). This is the *replacement hypothesis*. A further possibility, hitherto unrecognised in this literature, is that a child death leaves the mother depressed, as a result of which her subsequent child's health is compromised, both in the womb and in early infancy (see, for example, Steer et al. 1992, Rahman et al. 2004). This is referred to here as the *depression hypothesis.*

It is plausible that there are learning effects, which result in the mortality risk of the index child *falling* on account of the death of the preceding sibling. For instance, if the older

sibling died of diarrhoea, the mother may rush to learn how to prevent diarrhoea-related infant death. Any positive degree of scarring that is identified is then net of learning effects. Although it is of policy significance to establish which mechanism or mechanisms underlie scarring and there is little definitive research in this area, this paper does not attempt to offer any conclusive results in this direction. The examples provided are only illustrative. This paper is concerned primarily with the prior task of identifying whether there are any scarring effects after controlling for observed and unobserved heterogeneity.

This paper contributes to previous research in this area in two main ways. First, it introduces the notion of intra-family scarring which is conceptually distinct from inter-family heterogeneity in death risk. Second, in suggesting how robust estimates of scarring may be obtained, it offers methodological improvements on previous research. While Zenger (1993) describes causal mechanisms that stem from the death of a child and impact on the death risk of the next sibling, the models that she estimates include either the previous child's survival status or unobserved heterogeneity but, in no case, both. A few earlier studies do include the survival status of the preceding sibling in the model, while also allowing for unobserved heterogeneity (e.g. Curtis *et al* 1993, Guo 1993, Sastry 1997a & 1997b, Bolstad and Manda 2001). However, they do not interpret these effects in terms of causality and correlation respectively. Further, for reasons elaborated in section 4.1, the estimated coefficient on the survival status of the previous sibling is likely to be biased in all of these studies.

The analysis is conducted for infant mortality in three Indian states, and we find evidence of significant scarring. Results are presented to show the percentage of observed

persistence in child death that can be explained by scarring (i.e. by the survival status of the preceding sibling), and the reduction in mortality that would be achieved if scarring were eliminated. By virtue of generating inertia in the mortality process, scarring will tend to exercise a drag on the rate of mortality decline. This makes it important to recognise scarring and estimate its significance. Evidence of scarring immediately raises the payoff to policy interventions that reduce mortality because it implies that preventing the death of a child also contributes to preventing the death of siblings of that child.

The next Section describes the data used, the overall incidence of infant death and the extent of sibling death clustering. The statistical model is set out in Section 3, where scarring is formally defined and distinguished from unobserved heterogeneity. Issues that arise in the estimation of the model given the nature of the available data are discussed in Section 4, which further delineates the relation of this paper to previous research. Section 5 describes the empirical model and defines the variables. The results are set out in Section 6. The sensitivity of the estimated scarring effect to alternative specifications and procedures used in the existing literature is investigated in Section 7. This section demonstrates the potential for bias in previous research and, at the same time, suggests how it might be addressed. Section 8 concludes with a discussion of the findings and limitations of this study.

2. THE DATA AND DEATH CLUSTERING IN INDIA

This paper uses the second round of the National Family Health Survey of India (NFHS-II), which interviewed 92300 ever-married women aged 15-49 in 1998-99 and recorded complete

fertility histories for the 73,775 mothers amongst them, including the time and incidence of child deaths. There are 248,785 children in the sample, the mean number per mother being 3.4 and the median number 3. NFHS-II was conducted in 26 Indian states and covered more than 99 percent of India's population. For details on sampling strategy and context, see IIPS and ORC Macro (2000). The data are in the public domain and can be downloaded from www.macrodhs.com.

In a companion paper, we investigate scarring for each of the 15 major states of India and find evidence of scarring effects in 14 of the 15 states (see Arulampalam and Bhalotra 2004). As the current paper has a methodological emphasis, with alternative specifications being explored in section 7, the analysis here is restricted to the three states of Uttar Pradesh , West Bengal and Kerala. These states describe the spectrum of conditions within India. They exhibit remarkable differences in social, demographic, economic and political development (see Dreze and Sen 1997). Uttar Pradesh is the largest Indian state with social and demographic indicators that put it below the Indian average. Kerala is an exceptional state that leads India in almost every index of human development. West Bengal lies between the two in social-demographic development while exhibiting better economic indicators (level of per capita income, poverty incidence) than the other two states. A profile of the three states is presented in Table 1. Of every 1000 births in India, 82 die before the age of 12 months. There is remarkable inter-state variation. The corresponding numbers are 116 in Uttar Pradesh, 76 in West Bengal and 35 in Kerala (see Table 1). These figures are averages over the data sample. As this contains complete retrospective fertility histories, it includes children born across three

decades, 1968-1999. The average number of infant deaths per 1000 live births in India in 2001 is estimated to have been 67 (UNDP 2003).

The top panel of Table 2 shows the raw data probabilities of infant death conditional on the survival status of the preceding sibling. This is a useful description since, in the formal analysis conducted in this study and also in some previous studies, a first-order Markov model is specified in which, conditional on the survival status of the preceding child, the survival status of earlier children does not influence the survival status of the index child (see Section 3). Consider, for illustration, the state of Uttar Pradesh. The probability of infant death is higher by 0.16 (i.e. it is 0.25 rather than 0.09) if the preceding sibling died as an infant. An alternative expression of the relative risk is that an infant in Uttar Pradesh is 2.7 times as likely to die if the preceding sibling died rather than survived. Since the model presented in this paper is a logit, we also present the relative odds associated with the death of the preceding sibling (row [7] of Table 2). For example, in Uttar Pradesh, the relative odds of an infant dying in a family where the preceding sibling has died are 3.1 times higher than for a family in which the preceding sibling has survived infancy. The figures for West Bengal and Kerala are 4.2 and 6.3. Overall, the Indian data exhibit a remarkable degree of death clustering. Without further analysis, however, it is impossible to say whether this reflects scarring or whether it merely reflects risks common to siblings on account of shared family characteristics (heterogeneity).

3. THE STATISTICAL MODEL

This Section sets out a statistical model that permits identification of state dependence (scarring), taking account of the potentially confounding effects of unobserved inter-family heterogeneity.

Let there be n_i children in family *i*. For child *j* (*j=2,…,*n_i) in family *i* (*i=1,2,…, N*), the unobservable propensity to experience an infant death, y_{ij}^*, is specified as

$$y_{ij}^* = \boldsymbol{x}_{ij}'\beta + \gamma y_{ij-1} + \alpha_i + u_{ij} \qquad (1)$$

where $\boldsymbol{x}$ is a vector of strictly exogenous observable child and family specific characteristics that influence y_{ij}* and β is the vector of coefficients associated with $\boldsymbol{x}$. An infant is observed to die when his or her propensity for death crosses a threshold; in this case, when y_{ij}* > 0, and this binary outcome is denoted as y_{ij}=1. The term α_i captures unobserved heterogeneity. It accounts for all time-invariant unobserved and, possibly, unobservable family characteristics that influence the index child's propensity to die. This will include genetic characteristics and variables such as innate maternal ability. The null of no state dependence implies γ=0. The estimated parameter γ should be interpreted as the 'average' effect over the time period considered. In work in progress we investigate whether scarring has declined over time.

The model is dynamic in that it allows the unobservable propensity of death of the index child to be a function of whether the previous child died in infancy or not (i.e. y_{ij-1}=1 or 0). Defining a state as a realisation of a stochastic process, one may think of *state dependence* in terms of the mortality risk facing a child being dependent upon the state (died in infancy or

not) revealed for the previous child in the family. Since time is implicit in the sequencing of children, models that include the previous child's survival status are analogous to dynamic models. Note that, in principle, the preceding sibling may die after the index child. This can happen if, for example, the birth interval between them is 9 months and the first child dies at 11 months while the second child dies in the first month of life. There were no such cases in the data.

In this model, conditional on $y_{ij-1,}$ $\boldsymbol{x_{ij}}$ and α_i, the history of infant deaths amongst older siblings other than the immediately preceding child is assumed to have no impact on y_{ij}^*. If child (*j-2*) died in infancy then, in our model, this will affect the risk of death of child (*j-1*) and, *thereby*, affect the risk of death of child *j*. This is the first-order Markov assumption common in models of this sort (see Zenger (1993), for example). Moreover, a model restricted to first-order effects is consistent with the mechanisms that we suggest might drive scarring (Section 1). Of course, risk factors common to all siblings are captured by α_i.

For an account of dynamic (causal) models with unobserved heterogeneity in the econometrics literature, see Hsiao (1986), Wooldridge (2002). The distinction made in this paper between scarring and unobserved heterogeneity has been made in other contexts in both statistics and economics (see Heckman 1981a, 1981b, 1981c) although its relevance to death clustering has not formerly been recognised. For example, in the literature on the economics of unemployment, scarring refers to the effect of a past episode of unemployment on the future probability of experiencing unemployment, after controlling for all observable (e.g. education)

and unobservable (e.g. ability) individual characteristics.

Given the above assumptions, and dropping $\boldsymbol{x}$ and the index i for convenience, the joint probability of the observed sequence of binary outcomes is

$$P(y_n,\ldots,y_2,y_1|\alpha)=P(y_n|y_{n-1},\alpha)\,P(y_{n-1}|y_{n-2},\alpha)\ldots P(y_2|y_1,\alpha)\,P(y_1|\alpha) \qquad (2)$$

and therefore requires a specification for $P(y_1|\alpha)$. If there were no unobserved heterogeneity α_i, then the initial condition y_1 could be treated as exogenous, and the model given by equation (1) could be estimated using the sample of children j ($j=2,\ldots,n$). Alternatively, even in a dynamic model that incorporates unobserved heterogeneity, the initial conditions problem is avoided if the time dimension of the panel (n_i) is large (Hsiao (1986), pp170.). However, n_i in our model is given by the number of births of mother i, and this cannot be assumed to tend to infinity. As a result, consistent estimation requires that we endogenise (and model) the initial condition. This is done by specifying an equation for the mortality risk of the *first-born* child of each mother as

$$y_{i1}^{*} = \boldsymbol{z_i}'\boldsymbol{\lambda} + \theta\,\alpha_i + u_{i1} \qquad i=1,\ldots,N \qquad (3)$$

where $\boldsymbol{z_i}$ is a vector of exogenous covariates. In principle, the vector of covariates in $\boldsymbol{x}$ and $\boldsymbol{z}$ need not be the same, and θ need not equal one. Equations (1) and (3) together specify a complete model for the infant survival process.

Assuming that u_{ij} for $j=2,..,n_i$ as well as u_{i1}, are independently distributed as logistic, the joint conditional probability for the observed sequence of binary indicators for family i is

$$\mathrm{P}\left(y_{i1},...,y_{in_i} \mid \boldsymbol{x}_i, \boldsymbol{z}_i, \alpha_i\right) = \Lambda\left[\left(\boldsymbol{z}_{i1}'\lambda + \theta\alpha_i\right)\left(2y_{ij}-1\right)\right]\prod_{j=2}^{n_i}\Lambda\left[\left(\boldsymbol{x}_{ij}'\beta + \gamma y_{ij-1} + \alpha_i\right)\left(2y_{ij}-1\right)\right] \quad (4)$$

where Λ denotes the CDF of the logistic. Marginalising the likelihood with respect to α_i gives the likelihood function for family *i*

$$L_i = \int\left(\Lambda\left[\left(\mathrm{z}_{i1}'\lambda + \theta\alpha_i\right)\left(2y_{ij}-1\right)\right]\prod_{j=2}^{n_i}\Lambda\left[\left(\mathrm{x}_{ij}'\beta + \gamma y_{ij-1} + \alpha_i\right)\left(2y_{ij}-1\right)\right]\right) f(\alpha_i)d\alpha_i \quad (5)$$

where $f(\alpha)$ is the probability density function of the unobservable family-specific heterogeneity. Following the literature, we assume that α_i is independently and identically distributed as normal with zero mean and variance σ_α^2, but subject to the following restriction. Referring back to equations (1) and (3), a very large positive (negative) value for α_i will give a very large (small) value for y_{ij}^* and hence a very large (small) probability of observing death of the index child. Infant deaths are a rare occurrence, and some families never experience any infant deaths. As seen in section 2, there is also a tendency towards sibling death clustering, as a result of which some families may lose all of their children in infancy. Probability masses implied by the normal heterogeneity distributional assumption may not be sufficient to accommodate this phenomenon. This is the well-known mover-stayer problem (Blumen, *et. al.* 1955). In order to account for this, we follow the literature and allow for defectiveness at the two extremes by allowing empirically-determined masses at plus and minus infinity of the Normal mixing distribution. See Narendranathan and Elias (1993) for an application of this distributional assumption in the context of modelling individual unemployment and Barry *et.al*

(1989) for relevant software (though we program the estimation in STATA because of the presence of equation (3)).

The likelihood for family *i* is thus given by

$$L_i^* = \frac{\psi_0}{1+\psi_0+\psi_1}\left[\prod_{j=1}^{n_i}(1-y_{ij})\right] + \frac{L_i}{1+\psi_0+\psi_1} + \frac{\psi_1}{1+\psi_0+\psi_1}\left[\prod_{j=1}^{n_i} y_{ij}\right] \qquad (6)$$

where L_i is given by equation (5) and ψ_0 and ψ_1 are the unknown end-point parameters. The first term in the above sum allows for a mass point at minus infinity (y_{ij}=0, $\forall j$, for these families since they are assumed to experience no infant deaths) and the third term allows for a mass point at plus infinity (y_{ij}=1, $\forall j$, for these families since all children are assumed to die in infancy). Thus, the estimated proportion of families predicted to be located at -∞ and +∞ are given by p_0 and p_1 respectively, where,

$$p_0 = \frac{\psi_0}{1+\psi_0+\psi_1} \quad \textbf{\textit{and}} \quad p_1 = \frac{\psi_1}{1+\psi_0+\psi_1}. \qquad (7)$$

In order to ensure the non-negativity of ψ, it was parameterised as $exp(\kappa)$ and κ was estimated. In practice, the data may not contain enough variation in order to allow us to estimate ψ_1 and this is, indeed, what was found in this study (see Table 1 where the proportion of families that lose all or none of their children in infancy is reported).

Given the binary nature of the observed data, a normalisation is required in order to identify the parameters. A conventional practice is to fix the scale of the distribution of the error term, *u*. Even if we set θ=1 initially, if the variance of u_1 was different to that of the other *u* terms, the scale normalisation can imply a non-unit value for θ. Although we report

results for the logistic case, we also estimated the model under the assumption of normality and the results were very similar after allowing for scale differences in the coefficient estimates (Amemiya, 1981).

In addition to mother-specific unobserved heterogeneity, community-level random effects were included in the model to account for the sampling design, which involved clustering at the community level. Failure to allow for community-level unobserved heterogeneity in the likelihood maximisation would provide consistent parameter estimators but inconsistent standard errors (e.g. Deaton 1997, chapter 2). Although the model is multi-level, we have chosen to treat the community-level effect as a nuisance parameter. This is because we cannot interpret a time-invariant community-level effect in any meaningful manner. To the extent that families migrate or the infrastructure of different communities develops at different rates, the assumption of a time-invariant community effect is restrictive: we expect that children of the same mother, born at different dates, may experience different community-level effects. In any case, in this paper, the focus is not on estimation of the variance associated with mothers *vs* communities but rather, on robust estimation of the scarring effect, captured in the parameter, γ.

A test of H_0: $\sigma_\alpha^2=0$ is a test that there are no unobservable family characteristics in the model. This can be tested using a likelihood ratio (LR) (or a standard normal test statistic) but the test statistic will not have a standard χ^2 (or a standard normal) distribution since the parameter under the null is on the boundary of the parameter space. The standard LR (normal)

test statistic has a probability mass of 0.5 at zero and $0.5\chi^2(1)$ (0.5 N(0,1)) for positive values. Thus a one-sided 5% significance level test requires the use of the 10% critical value (Lawless (1987), Andrews (2001)).

Although equation (1) is a standard dynamic random effects logit model (with three levels), the inclusion of equation (3) in the estimation (to account for initial conditions) makes it non-standard. Hence a routine (available on request from the authors) was written in Stata (2000) using Stata's maximisation procedures, to obtain parameter estimates. Gaussian-quadrature was used to approximate the integral in (5).

4. ISSUES OF MODEL SPECIFICATION AND TESTING

This Section describes potential problems that arise in an empirical specification of the model, indicating how common biases in parameter estimators may be avoided. Problems discussed include those of left-truncation, endogeneity, measurement error and time-inconsistency.

4.1 The Initial Conditions Problem

In our survey, women aged 15-49 in 1998/99 were interviewed and retrospective data on their birth histories were collected. A well-recognised problem with retrospective data, when an age cut-off is used to select the interviewees, is a selectivity issue. The interviewees may be a representative sample as of the survey date, but will not be so for earlier years (Rindfuss, *et. al.*, 1982). For this reason and for reasons of recall bias, a common practise in previous research has been to discard information on children born before an arbitrarily selected date, such as five, ten or fifteen years before the date of the survey (e.g., Guo 1993, Curtis, *et al*

1993, Madise and Diamond 1995, Sastry 1997a & 1997b, Bolstad and Manda 2001, Bhargava 2003). This left truncation of the data by calendar time occurs at different points in the birth history of different households, creating additional complications. Many studies also discard the first-born child in every family. This can also result in a severe loss of information (see the number of observations recorded in rows 1-4 of Table 3).

Moreover, left truncation of the data, whether by calendar time or by birth-order of child, results in the problem that the start of the sample does not coincide with the start of the stochastic process under study. On account of the presence of family unobservables, α_i, in equation (1), the survival status of the previous child, y_{ij-1}, is endogenous. Thus, discarding observations at the beginning of the sample produces an endogenously truncated sample. In other words, since α_i is a family-specific term, it will appear in the equation for every child in the family. In particular, it will appear in the equation for y_{ij}* and also in the equation for y_{ij-1}*. Therefore, in the equation for y_{ij}*, the regressor, y_{ij-1}, is necessarily correlated with the error-component, α_i. This is what is meant by endogeneity of y_{ij-1} and, left unaddressed, it will tend to produce a (positive) bias on the coefficient estimate of y_{ij-1}, which provides an estimate of scarring effects.

This is an instance of the 'initial conditions problem' in dynamic models with unobserved heterogeneity (e.g. Heckman (1981c)). Intuitively, the problem is that the model describes a dynamic process, and we need to allow for how it starts. The death risk of the third child in a family depends upon whether the second child died and the death risk of the second

child depends upon whether the first child died, but information on whether the first child died is missing if the data are left-truncated or if information on first-borns is discarded. This information is especially relevant because the first child shares unobservable traits (α_i) with her younger siblings.

In order to address the initial conditions problem, we use the complete birth histories of the women in our sample, and specify equation (3) to describe mortality risk for the first-born child of each mother. In this way, we model the start of the dynamic process that operates within families, with the death of one sibling impacting upon the death risk of the next sibling. Availability of information on the start of the process is an unusual feature of the series of Demographic and Health Surveys of which the Indian survey we use is one. In many other applications of dynamic models with unobserved heterogeneity, data on the start of the process are unavailable. For example, in studying unemployment spells of individuals, researchers would ideally like to have data on school-leavers but must often make do with left-truncated data, that is, data that do not include the first spell of unemployment for each individual.

There are two suggested techniques in the literature that enable one to account for endogeneity of the initial conditions in the estimation. Heckman (1981c) suggests using an equation such as (3) as an approximation for the initial condition $y_{i1.}$ An alternative strategy proposed by Wooldridge (2005) is to approximate the distribution of α_i conditional on y_{i1} rather than the distribution of y_{i1} conditional on α_i, as suggested by Heckman (see equation (2)). In the absence of guidance from economic theory, there are no *a priori* reasons to expect

one method to be superior to the other. Although the main specification in this paper involves avoiding left truncation altogether, in Section 7 we discuss how the initial conditions problem can be mitigated in situations in which the nature of the data or the nature of the research favour using left-truncated data. We have chosen to use Heckman's suggested method to illustrate this point as this procedure is the more natural one to compare with the case in which full birth-histories are used in the estimation.

The fact that left truncation can create a bias if not properly accounted for in the estimation has not been previously recognised. This is true of all of the relevant demographic research that we are aware of. The only study that reflects awareness of the endogeneity problem arising *via* the correlation of the survival status of previous children and family unobservables is Bhargava (2003). That study left-truncates the data and attempts to address the endogeneity problem by imposing the restriction that household possessions and the number of boys and girls born before the index child influence the *number of surviving older children* but, conditional on this variable, have no influence on the mortality risk of the index child. Also, Bhargava (2003) is not concerned with scarring effects. Our analysis confirms the importance of addressing the initial conditions problem. Section 6 provides statistical tests on the parameter θ (see equations 3 and 4) that are useful in assessing the empirical relevance of the initial conditions problem. Furthermore, estimates of alternative models presented in Section 7 indicate the direction and size of the bias induced by left-truncation.

The available data are *right*-censored in the sense that, at the time of the survey in

1998/9, some of the women who were interviewed had not completed their fertility. This creates a different problem, which is that we have in our sample a disproportionately large representation of children of older mothers. Although mother's age is included as an intercept effect, it is not interacted with y_{ij-1}. Therefore our estimates of scarring may not be "representative"- that is, including all children of younger mothers in the sample may reduce the estimated scarring effect. We have not selected mothers with completed fertility histories because completion of fertility involves a *choice* – selecting on a choice variable can produce endogenous selection bias, addressing which would demand the joint modeling of fertility and mortality.

4.2 Measurement Error

A reason that previous studies have left-truncated the data is that this minimises recall error in the recorded date of child death, which is assumed to be larger the further away the mother is from the event (e.g. Sastry 1997a). It may seem implausible, *a priori*, that mothers ever forget the date of death of a child but the data do exhibit some age-heaping. In particular, the Indian data that are used in this study show heaping at six-month intervals (also see IIPS and ORC Macro 2000: Section 6.2). What effect is this expected to have on estimates of our model? Since the model has infant death on both sides of the equation, with the index child's risk a function of the preceding child's survival status, positively correlated measurement error in these variables will tend to create an upward bias in the scarring coefficient. This potential problem is addressed as follows. The dependent variable and the survival status of the preceding child are both coded as binary variables that are unity if the child dies before the age

of 12 months and zero otherwise. To investigate sensitivity of the estimates to age-heaping at 12 months the models were re-estimated with these variables defined to include deaths occurring at 12 months. The results were very similar (and so are not shown but available on request).

4.3. Time Inconsistency

Survey data used to study childhood mortality typically contain retrospective histories of births and child deaths experienced by ever-married women aged 15-49 at the time of the survey. They also typically gather information on variables such as household assets, toilet facility, electricity or access to piped water at the date of the survey. The data we use for India are similar. A woman aged 49 in 1999 may have experienced a birth and an infant death as long ago as 1968. The time-inconsistency problem is that, in such cases, data that pertain to the date of the survey are less informative, the further from the date of survey is the event of interest (e.g. childhood death). The practice of left-truncation limits this problem, even if it remains somewhat questionable given that growth, migration and structural change can occur quite rapidly in developing countries. In the current analysis, where the entire birth history of each mother is used, the problem is more severe. We therefore do not include any current-dated variables as regressors.

Another, less-recognised problem with some of these variables is that they are endogenous. For example, families will tend to simultaneously decide what resources to allocate to the purchase of assets (a bicycle, a TV), and what resources to spend on inputs into

child health that will reduce child mortality risk (see Becker 1991, for example). Alternatively, access to facilities like piped water will be endogenous if families migrate to regions with piped water.

4.4. Specification of Scarring and Birth Interval Effects

As discussed in Section 1, earlier studies of death clustering were not specifically looking to identify scarring effects as distinct from unobserved heterogeneity across families. This is reflected in the specifications that they employ. Most previous analyses of death clustering model unobserved heterogeneity alone, although a few include variables related to the scarring process. For example, Bhargava (2003) and Muhuri and Preston (1991) include the number of surviving older siblings instead of the survival status of the previous child. This is a compound indicator of fertility and mortality in the family. Moreover, it is insensitive to sequencing- and so it does not reflect 'scarring' as defined in Section 3- as that involves a causal relation between the death of one child and the risk of death of the next child in the family.

A handful of previous studies follow a specification similar to that in this paper in that they include the survival status of the previous sibling in the model (Curtis *et al* 1993, Guo 1993, Sastry 1997a & 1997b, Bolstad and Manda 2001). However, an important difference in specification is that all of these studies also include the preceding birth interval as a regressor. In Section 1 it was argued that the fecundity and replacement mechanisms are two plausible causal processes by which scarring effects may appear, and both involve the birth interval as a proximate variable. To the extent that the previous child's survival status, y_{ij-1}, impacts on the index child's death risk, y_{ij}^{*}, by altering the length of the birth interval, conditioning on the

birth interval will tend to weaken the coefficient on y_{ij-1}. As a result, the degree of scarring will tend to be under-estimated. In other words, to include the birth interval along with an indicator for the death of the preceding sibling (y_{ij-1}) in our model would amount to including both the ultimate and the proximate cause. The reason we prefer to include y_{ij-1} and not the birth interval is that scarring may occur for reasons that do not create a short birth interval. One example of this possibility is the depression mechanism referred to in section 1. So our strategy implies that the total scarring effect is captured in the coefficient on y_{ij-1}. If the birth interval is included as an additional regressor then the coefficient on y_{ij-1} will denote only a partial scarring effect (ie that part of scarring that is not attributable to short birth intervals).

Another problem with the specification used in previous studies is that the birth interval is an endogenous variable (that is, potentially chosen by the family) and one for which valid instruments may be difficult to find. Although *uptake* of contraception is a choice variable (endogenous), the *availability* of contraception is a potential instrument for birth interval, not considered in the previous literature. Since information on contraception in the NFHS data is limited to recent births and using it would involve endogenous left truncation of the data (see Section 4.1), exploration of this is left to future work. Previous research does not appear to have acknowledged the potential for endogeneity bias in estimates of the effects of birth intervals on mortality (although see Bhalotra and van Soest (2004) for a recent attempt at this in the context of neonatal mortality).

There are also measurement problems with birth intervals as they may be shorter on account of premature birth (e.g., Gribble 1993) or longer on account of miscarriage (e.g.

Madise and Diamond 1995). If these events are sufficiently common in the data, the coefficient on birth interval will reflect a compound of these effects.

In this paper, the scarring effect is captured entirely by the coefficient on previous sibling's survival status (γ). To allow comparison with previous studies and, for the Indian data, to assess the impact on γ, results are presented, in Section 7, for a variant of the model in which preceding birth interval is included as an additional regressor. Of course, in the absence of controls for the endogeneity of birth spacing, these results are only indicative.

5. THE EMPIRICAL MODEL

The dependent variable, y_{ij}, is defined as unity if the child is observed to die before the age of 12 months and zero otherwise (infant death). The regressor of interest, y_{ij-1}, is similarly defined as the infant survival status of the preceding sibling. As discussed in Section 4.2, sensitivity of the results to "heaping" in the reported age of death was also investigated. Children who have not had 12 months exposure (i.e. who are younger than 12 months) at the time of the survey are dropped from the sample. When the index child is not a singleton but, instead, a twin (or triplet) then care is taken to ensure that the preceding sibling is correctly identified and is the same for each twin. When the previous child is one of a multiple birth, then y_{ij-1} is defined as unity if all children of that multiple birth died in infancy and as zero otherwise. This is the relevant assumption if the mechanism underlying scarring is the fecundity mechanism since the mother is only likely to stop breastfeeding if both twins (or all three triplets) die. We have confirmed that altering this definition so that y_{ij-1} is defined, as

unity when *at least one* of the multiple births dies does not change the results. This is unsurprising since multiple births are uncommon (see Table 1).

The rest of this Section describes the variables in the vector $\boldsymbol{x_{ij}}$, which are assumed to be identical to the variables $\boldsymbol{z_{ij}}$ in the first-child equation. Means and standard deviations of all variables in the model are in Appendix Table 1. Covariates often used in previous research that are time-inconsistent or endogenous are avoided. The only potentially endogenous variable in the model is y_{ij-1} and, as discussed in Sections 3 and 4.1, addressing this potential problem is an important part of the statistical approach taken here. Since this involves using retrospective histories that go back several years in time, cohort effects are introduced into the model.

Child-specific regressors in the equation are child birth-order, gender and an indicator for whether the child is one of a multiple birth (twin, triplet, etc). The age of the mother at birth of the index child is included to reflect the physiological condition of the mother at a relevant time. Since several studies show child mortality risk to be U-shaped in mother's age, this is specified as a quadratic. Education of the mother is denoted by a set of dummy variables for level of education attained. This is relatively flexible, allowing for non-linear effects (Cleland and van Ginneken (1989) and Hobcraft (1993) provide reviews of the effects of maternal education on childhood mortality; also see Rosenzweig and Schultz (1982)). A similar set of indicators for educational level of the father is included. This is likely to be an important control for socio-economic status to the extent that fathers are the main earners (available data on household assets are not used because of the time inconsistency problem).

Other family-level observable variables included in the model are religion and caste. These allow for "sociological" influences on child death risks.

Cohort effects are modelled by including dummy variables for year of birth of the mother. Mothers in the sample are born between 1948 and 1984. Roughly equal frequency groups are created by defining dummy variables for births during 1948-1959, 1960-1969 and 1970-1984. The cohort effects are expected to pick up any secular decline in death risks over time, other things equal. Note that the child's date of birth is effectively in the model since it also includes the age of the mother at the birth of the index child. To see this, consider a woman who was born in 1940 and gave birth to the index child in 1960 so that age of the mother at birth of the child is 20. The model includes "20" and "1940" and so it implicitly includes "1960".

There were missing values for religion, caste and parental education. In most cases, less than half a percent of observations had missing values, but caste information was missing for 4.8% of cases in Uttar Pradesh and father's education was missing for 0.08% in West Bengal (see Appendix Table 1) How one treats these depends on the assumptions one is prepared to make regarding the structure of the missing data; whether it is: missing at random, missing completely at random, ignorable or nonignorable, for instance (see Cameron and Trivedi 2005, Chapter 27). Alternative methods for dealing with missing data, including single and multiple imputation techniques, are discussed by these authors and also in Rubin (2004). In this paper, we assume that the missing data mechanism is ignorable, that is, the missingness in the variable does not depend upon its value and the parameters of the missing data-

generation process are unrelated to the model parameters of interest. Under this assumption, one can proceed with the estimation by only including observations without any missing values. In this paper, we proceed with this assumption. However, we report the results from an alternative, more ad hoc approach, and check the sensitivity of our results to the approach taken. The alternative approach involves creating binary variables to pick up the cases with missing values, which are then included as additional regressors in the model. In cases where the number of observations with missing data are too small for a coefficient on the missing-value dummy to be precisely estimated (e.g. father's education is missing for 8 families in Kerala), we combine the missing cases with the omitted category (which, for the case of father's education, for example, is fathers with no education). The results were not sensitive to the imposition of this restriction. For the cases (caste in UP and father's education in WB) where we are able to estimate coefficients on these dummies, we find that they are insignificant. The reported results are based on this second approach. The following estimates obtain when, instead, missing values in these cases are dropped: $\hat{\gamma}$ (standard error): for Uttar Pradesh, 0.686 (0.070) and for West Bengal, 0.428 (0.171). We will see that these are very similar to the results reported in the following section.

6. RESULTS

The main result is that we find evidence of scarring in each of the three Indian states, after controlling for a number of exogenous child and family-specific characteristics and for all unobserved differences between families (row [12] Table 2). A full set of results for other

covariates is available on request from the authors.

In the logit model, the log-odds ratio is a linear function of the explanatory variables. Thus, for instance, $exp(\gamma)$ is the effect of the previous infant's death on the relative odds of the index child's death. We also present the marginal effect associated with γ. This is computed as the difference between the sample averages of the probability of death predicted by the estimated model when $y_{ij-1}=0$ and when $y_{ij-1}=1$, which is approximately equivalent to the first partial derivative of the conditional probability of death of the index child with respect to y_{ij-1}. This is what we call state-dependence or scarring in this paper.

Comparing the estimated scarring effect with the difference and ratio of the "raw data probabilities" discussed in Section 2 (and reported in the top panel of Table 2) affords an estimate of the percentage of raw persistence (or clustering) that is explained by scarring, using the model specified in Section 3. Scarring explains about 40% of the clustering observed in the data in Uttar Pradesh; the corresponding proportions being 14% for West Bengal and 21.5% for Kerala (row [14], Table 2). As discussed, previous research has identified clustering with unobserved heterogeneity- these estimates show that in fact, almost half of observed clustering in Uttar Pradesh is attributable to scarring, after holding constant unobserved heterogeneity.

Comparing the averaged model predicted probability of death (excluding first borns) with that of the averaged predicted probability of death setting $\gamma=0$ offers an estimate of the reduction in mortality that would be achievable if scarring were eliminated- a useful

alternative expression of its significance. The estimates suggest that, in the absence of scarring, mortality rates among children born after the first, would fall by 9.8%, 6.0% and 5.9% in Uttar Pradesh, West Bengal and Kerala respectively.

The Table (row [17]) reports a test of the null hypothesis that $\theta=0$ in equation (3). This is a test of the hypothesis that the initial sample observation (child) within a family can be treated as exogenous (as in previous research). Clearly, if $\theta=0$ then unobservables in the equation for the first observation are uncorrelated with unobservables in the [dynamic] equations for subsequent observations. In this case, the model described by (1) and (3) reduces to a simple random effects model and a separate specification of the equation for the initial sample observation is unnecessary. The null that $\theta=0$ is decisively rejected in the case of Uttar Pradesh and West Bengal, which confirms the importance of specifying a distinct reduced form equation for the first child that is estimated jointly with the dynamic equations for other children.

The proportion of the variance attributable to family-level unobservables (α_i) is estimated to be 11% in Uttar Pradesh, 21% in West Bengal and 7% in Kerala (row [19], Table 2). The estimates decisively reject the null of no family-level unobservables for the states of Uttar Pradesh and West Bengal. This and the finding that exclusion of α_i from the model results in over-estimation of scarring (results not shown but available) underline the importance of controlling for α_i.

Many of the covariates in the vector $\mathbf{x}_{ij}$ are estimated to be significant determinants of

mortality (results available upon request). The end-point of the α distribution at $-\infty$, p_0, is insignificant in Uttar Pradesh and West Bengal, but significant in Kerla. This end point mass is included in the model to pickup families that never experience an infant death. As expected, in Kerala, which had the lowest incidence of infant mortality and small family sizes, the model predicts that about 54% of the families are in this category. There was insufficient variation in the data for p_1 to be determined (these terms are defined in Section 4.5). Of course, the additional flexibility allowed by introducing mass points at the two extremes of the distribution may turn out to be important in other data sets.

7. SENSITIVITY OF ESTIMATED SCARRING EFFECT

7.1 Estimates obtained on a left-truncated sample

As discussed in Section 4.1, previous studies left-truncate the sample without seeming to recognise that, if the survival status of the preceding child is amongst the regressors, then this will result in a (positive) bias in its estimated coefficient. To confirm this prediction and to establish the extent of the bias, estimates of the model are obtained under these conditions (Table 3) and compared with the estimates reported in Table 2. Three specifications are investigated.

First, the first-*born* child in each family is discarded from the sample. This is relevant because previous studies do not model the survival status of first-borns. As expected, the resulting 'initial conditions' problem creates a positive bias. The scarring effect increases in all three states, the percentage increase in West Bengal and Kerala being quite dramatic (row

2, Table 3).

The next experiment follows the previous literature in discarding all children born before a certain calendar year. Most previous studies discard observations 5 or 10 years before the date of the survey. This again introduces the initial conditions problem and, therefore, a positive bias (Section 4.1). However, if scarring has been decreasing over time, then a *smaller* scarring effect may be observed in the sub-sample of children born 5 or 10 years before the survey date as compared with the full sample of children in the data, who were born over a span of 36 years. So as to focus on the initial conditions problem and minimise time effects, the left-truncation performed in this second experiment is pushed further back in time. Data are discarded for children born before 1971 so that information for 28 years is retained, with only the initial eight years of data, corresponding to 3.0% of children, being discarded. In this now truncated sample, y_{ij-1} is, of course, undefined for the first-*observed* child in each family. In line with previous research, these children are also excluded from the estimated model (results in row 3). The scarring parameter shows the expected upward bias, and it is of roughly similar magnitude to that obtained in row 2. Rows 2 and 3 of Table 3 establish that in the few existing papers that implicitly contain estimates of scarring, these are likely to be over-estimates (e.g. Curtis *et al* 1993, Guo 1993, Sastry 1997a & 1997b, Bolstad and Manda 2001).

What can be done to mitigate these biases in situations in which left-truncation is necessary? Thus, for example, information on breastfeeding or antenatal care may be essential to the purpose of a study and these data are only available (in the Indian NFHS and also in several other DHS surveys) for the 3-5 years preceding the survey. Consistent estimators may

be obtainable from an endogenously truncated sample if a reduced form equation for the first-*observed* child in the truncated sample is specified and estimated (Heckman 1981c). Results are in Row-4. The scarring estimates are similar to the preferred estimate in row-1, indicating that this strategy goes a fair way towards redressing the initial conditions problem. Also, $\theta=0$ is again rejected for Uttar Pradesh and West Bengal, which confirms the relevance of modelling the first-observed child. This result is likely to be of considerable practical importance.

7.2 Introducing preceding birth interval as a regressor

Refer Section 4.4 where it was argued that the preferred model is one without the birth interval but that a model including this variable both indicates the bias in the scarring parameter in previous research. It also offers insight into the mechanism underlying scarring. In this Section we present, for comparison, results obtained when birth interval is included as an additional explanatory variable in the model.

The preceding birth interval for the index child is defined as a set of dummy variables for 8-17, 18-23, 24-29 and more than 29 months; unsurprisingly, there are no observations with a value of less than 8 months (the average birth interval is reported in Table 1). A set of four intervals is preferred to a quadratic in the birth interval because the distribution exhibits a long tail, which the quadratic form would exaggerate. The choice of intervals is guided by examination of the distribution of the variable and by the demographic literature. It is set to zero for first-born children. The data were coded to ensure that all children in a multiple birth

have the same preceding birth interval. The birth interval dummies are positive and significant and their inclusion is seen to reduce the scarring effect in each of the three states (see row 5 Table 3). In Uttar Pradesh, the scarring coefficient (γ) remains significant but, in West Bengal and Kerala, it is rendered insignificant. The results suggest that a mechanism generating short birth intervals is one part of the scarring story but that, at least in Uttar Pradesh, there is also some other scarring mechanism at work. As discussed in Section 4.4, these results are only tentative since the endogeneity of the birth interval has not been addressed in this experiment.

8. CONCLUSIONS

This paper has investigated the clustering of sibling infant deaths in India. In a departure from previous research in this area, the main aim of the paper was to introduce the idea of scarring as a causal process that might contribute, together with inter-family heterogeneity, to the phenomenon of death clustering. Scarring is of considerable theoretical interest, contributing to understanding the inter-relations of family behaviour, fertility and mortality. It is also clearly of interest to policy-making. As indicated in Section 1, evidence of scarring raises the payoff to interventions that reduce mortality. It can also be useful in targeting interventions at the most vulnerable households. Previous analyses of death clustering have equated death clustering with unobserved heterogeneity. This has been thought to represent genetic traits (e.g. Sastry 1997a, 1997b) or maternal ability (e.g. DasGupta 1990). There is not a lot that policy can do in these cases: it is difficult to engineer genetic change or to influence maternal

ability beyond education. As scarring is a causal process, there is immediate scope for intervention. For example, if the causal process works through the fecundity mechanism (see Section 1) then policies that improve uptake of contraception are likely to reduce death clustering. More specific policy insight depends upon identifying the mechanism underlying scarring. A further reason that scarring is interesting is that it generates inertia or short-term persistence in the mortality process, as a result of which it will tend to exert a drag on the rate of mortality decline.

The statistical issues raised in this paper are expected to be widely applicable in further demographic research. The Indian National Family Health Survey analysed here is one of about 69 Demographic and Health Surveys (DHS) available for low and middle income countries. The DHS data typically contain information on all children of a mother including the first-born. Data on first-born children have quite consistently been thrown away and it is argued here that this not only constitutes a considerable loss of information but is also a source of bias in dynamic models with unobserved heterogeneity. A set of testable restrictions on the model confirms the importance of some of the statistical innovations that are made. Estimation of some variants of the preferred model shows the extent of bias in the scarring parameter that would arise if some of the specification issues highlighted here were ignored.

The main result is that there is a significant degree of scarring in all of the three Indian states for which data were analysed. In order to assess the size of this effect, it is useful to consider the reduction in mortality (excluding first-born children from the sample) that could be achieved if scarring were set to zero by a hypothetical policy intervention. Among second

and higher birth-order births to mothers born between 1949 and 1984, this is estimated to be 9.8% in Uttar Pradesh, 6.0% in West Bengal and 5.9% in Kerala. The fact that Kerala and West Bengal have smaller families (and a higher proportion of first-born children) probably limits the overall impact of scarring in these states: the raw data also clearly indicate a greater degree of clustering in families with a larger number of children. It would be interesting to investigate, in future work, whether the degree of scarring is increasing in birth order and whether it varies with the gender of the preceding sibling. Also, as indicated earlier, these estimates reflect average behaviour over the period under consideration. Further work investigating whether scarring has declined over time and comparing the rate of decline across states is merited.

Preliminary investigation of alternative mechanisms driving scarring suggests that shorter birth intervals following the death of a child in the family constitute an important part of the story, although birth spacing alone does not entirely account for scarring, particularly in the state of Uttar Pradesh. Further research into the processes underlying scarring is merited.

REFERENCES

Amemiya, T. (1981) Qualitative response models: a survey. *Journal of Economic Literature*, 19, 1483-1536.

Andrews, D. W. K. (2001) Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69(3), 683-734.

Arulampalam, W. and Bhalotra, S. (2004) Infant survival in India: frailty and state-dependence, Working Paper 04/558, Department of Economics, University of Bristol, May.

Barry, J., Francis, B., and Davies, R., (1989) SABRE, Statistical analysis of binary recurrent events. Centre for Applied Statistics, Lancaster Universitty.

Bean, L. L., Geraldine, P. M. and Douglas, L. A. (1988) Reproductive behaviour and child survival among Nineteenth Century Mormons. In **t**he 13th annual meeting of the Social Science History Association, November 1988 Chicago.

Becker, G. (1991) *A Treatise on the Family*, 2nd ed. Cambridge, Mass: Harvard University Press.

Bhalotra, S. and Van Soest, A. (2004) Birthspacing and neonatal mortality in India: Dynamics, frailty and fecundity. Working Paper *04/567*, University of Bristol, December.

Bhargava, A. (2003) Family planning, gender differences and infant mortality: evidence from Uttar Pradesh, India. *Journal of Econometrics*, 112, 225-240.

Blumen, I., Kogan, M., and McCarthy, P. J. (1955) *The industrial mobility of labor as a probability process.* Cornell Studies in Industrial and Labor Relations. Ithaca, NY: Cornell University Press.

Bolstad, W. M. and Manda, S. O. (2001) Investigating child mortality in Malawi using family and community random effects: a Bayesian analysis. *Journal of the American Statistical Association*, 96, 12-19.

Bongaarts, J. and Potter, R. G. (1983) *Fertility, Biology and Behaviour: An Analysis of the Proximate Determinants*, New York: Academic Press.

Cameron, A. C. and Trivedi, P. K. (2005) Microeconometrics: Methods and Applications, Cambridge University Press.

Cantrelle, P., Ferry, B. and Mondot, J. (1978) Relationships between fertility and mortality in Tropical Africa. In: *The Effects of Infant and Child Mortality on Fertility* (ed. S. Preston), pp 181-205. New York: Academic Press.

Chamberlain, G. (1984) Panel data. In: *Handbook of Econometrics* (ed. S. Griliches and M. Intriligator), pp 1247-1318. Amsterdam: North-Holland.

Chay, K. Y. and Hyslop, D. R. (2000) Identification and estimation of dynamic binary response panel data models: empirical evidence using alternative approaches. Mimeograph, August.

Chen, L., Ahmed, S., Gesche, M. and Mosley, W. (1974) A prospective study of birth interval dynamics in rural Bangladesh. *Population Studies*, 28, 277-297.

Cleland, J. G. and Sathar, Z. A. (1984) The effect of birth spacing on childhood mortality in Pakistan. *Population Studies*, 38, 401-418.

Cleland, J. G. and Van Ginneken, J. K. (1989) Maternal education and child survival in developing countries: the search for pathways of influence. *Social Science and Medicine*, 27(12): 1357-1368.

Curtis, S. L., Diamond, I., and McDonald, J. W. (1993) Birth interval and family effects on postneonatal mortality in Brazil. *Demography*, 33(1), 33-43.

DaVanzo, J. and Pebley A. R. (1993) Maternal depletion and child survival in Guatemala and Malaysia. *Labor and Population Program Working Paper* 93-18, RAND.

DasGupta, M. (1990) Death clustering, mothers' education and the determinants of child mortality in Rural Punjab, India. *Population Studies*, 44, 489-505.

Deaton, A. (1997) The Analysis of Household Surveys: A Microeconometric Approach to Development Policy. The Johns Hopkins University Press for the World Bank.

Dreze, J. and Amartya, S. (1997) *Indian Development: Selected Regional Perspectives.* Oxford: Clarendon Press.

Gribble, J. N. (1993) Birth intervals, gestational age and low birth weight: are the relationships confounded? *Population Studies*, 47, 133-146.

Guo, G. (1993) Use of sibling data to estimate family mortality effects in Guatemala. *Demography*, 30(1), 15-32.

Heckman, J. J. (1981a) Statistical models for discrete panel data. In: *Structural Analysis of Discrete Data with Econometric Applications* (ed. C. F. Manski and D. McFadden), pp 114-178. Cambridge: MIT Press.

Heckman, J. J. (1981b) Heterogeneity and state dependence. In: *Studies in Labor Markets* (ed. S. Rosen). Chicago: Chicago Press.

Heckman, J. J. (1981c) The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In: *Structural Analysis of Discrete Data with Econometric Applications* (ed. C. F. Manski and D. McFadden), pp 114-178. Cambridge: MIT Press.

Hobcraft, J. N. (1993) Women's education, child welfare and child survival: A review of the evidence. *Health Transition Review*, 3(2): 159-175.

Hobcraft J., McDonald J. W., and Ruttstein S. (1983) Child spacing effects on infant and early child mortality. *Population Index*, 49, 585-618.

Hobcraft J. N., McDonald J. W., and Ruttstein S.O. (1985) Demographic determinants of infant and early child mortality: A comparative analysis. *Population Studies*, 39, 363-385.

Hsiao, C. (1986) *Analysis of Panel Data*. Cambridge University Press.

Hyslop, D. R. (1999) State dependence, serial correlation and heterogeneity in intertemporal labour force participation of married women. *Econometrica*, 67(6), 1255-94.

IIPS and ORC Macro (2000) *National Family Health Survey (NFHS-2) 1998-9*: India. Mumbai: International Institute for Population Sciences (IIPS).

Koenig, M. A., Phillips, J. F., Campbell, O. M. and D'Souza, S. (1990), Birth interval and childhood mortality in rural Bangladesh. *Demography*, 27, 251-265.

Lawless, J. F. (1987) Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15, 209-25.

Madise, N. J. and Diamond, I. (1995) Determinants of infant mortality in Malawi: an analysis to control for death clustering within families. *Journal of Biosocial Science*, 27(1), 95-106.

Miller, J. E., Trussell, J., Pebley, AR and Vaughan, B. (1992) Birth spacing and child mortality in Bangladesh and the Philippines. *Demography*, 29, 305-318.

Muhuri, P. and Preston, S. (1991) Effects of family composition on mortality differentials by sex among children in Matlab Bangladesh. *Population and Development Review*, 17, 415-434.

Narendranathan, W. and Elias, P. (1993) Influences of past history on the incidence of youth unemployment: Empirical findings for the UK. *Oxford Bulletin of Economics and Statistics*, 55 (2), 161-185.

Nath, D. C., Land, K. C., Singh, K. K. (1994) Birth spacing, breastfeeding and early child mortality in a traditional Indian society: a hazards model analysis. *Social Biology*, 41, 168-180.

Preston, S. H. (1985) Mortality in childhood: lessons from WFS. In: *Reproductive Change in Developing Countries* (ed. J. G. Cleland and John Hobcraft), pp 46-59. Oxford: Oxford University Press.

Rahman, A., Iqbal, Z., Bunn, J., Lovel, H. and Harrington, R. (2004) Impact of maternal depression on infant nutritional status and illness: A cohort study. *Archives of General Psychiatry*, 61: 946-952.

Rindfuss, R. R., Palmore, J. A., and Bumpass, L. L. (1982) Selectivity and the analysis of birth intervals from survey data. *Asia and Pacific Census Forum*, 8(3), 5-16.

Rosenzweig, M. and Schultz, T. P. (1982) Child mortality in Colombia: individual and community effects. *Health Policy and Education*, 2(2), 305-348.

Rubin, D. B. (2004) Multiple Imputation for Nonresponse in Surveys, Wiley-IEEE.

Sastry, N. (1997a) Family-level clustering of childhood mortality risk in Northeast Brazil. *Population Studies*, 51(3), 245-261.

Sastry, N. (1997b) A nested frailty model for survival data, with an application to the study of child survival in Northeast Brazil. *Journal of the American Statistical Association*, 92(438), 426-435.

Stata 8 (2004), *Stata Statistical Software,* Stata Corportation.

Steer R. A., Scholl, T. O., Hediger, M. L. and Fischer, R. L. (1992) Self-reported depression and negative pregnancy outcomes. *Journal of Clinical Epidemiology*, 45(10), 1093-9.

UNDP (2003) Human Development Report: *Millenium Development Goals: A Compact Among Nations To End Human Poverty*.

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

Wooldridge, J. M. (2005) Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics*, 20(1), 39-54.

World Bank (2000) *India- Reducing Poverty, Accelerating Development: A World Bank Country Study*. Oxford: Oxford University Press.

Zenger, E. (1993) Siblings' neonatal mortality risks and birth spacing in Bangladesh. *Demography*, 30(3), 477-488.

Table 1: Descriptive Statistics

	Uttar Pradesh	West Bengal	Kerala
Demographic variables			
Probability of infant death [all live births]	0.116	0.076	0.035
Probability of infant death excluding first borns	0.111	0.073	0.033
Age of mother in 1998/9	34.9	35.0	37.0
Age of mother at first marriage	15.7	16.2	18.9
Age of mother at first birth	18.1	18.1	20.3
% women that have never used any method of contraception	55.0	16.8	15.9
% women who can read and write	10.5	25.7	52.5
Total children ever born per mother	5.5	4.2	3.3
% women with 1-2 children	27.8	51.0	59.1
% women with 3-4 children	33.7	32.3	33.4
% women with 5 or more children	38.5	16.7	7.5
Mean (median) birth interval in months[(v)]	30.8 (26)	33.3 (28)	35.4 (29)
% families with no infant deaths	69.6	84.3	92.6
% families in which all births die in infancy	1.22	0.60	0.42
% multiple births	1.37	1.53	1.54
% first-born children	24.2	34.1	39.3
Probability of infant death amongst first-borns	0.131	0.080	0.038
Economic & infrastructure variables			
Rank in per capita income	12	6	8
Growth rate	2.2	3.2	3
Poverty incidence	40.2	26	29.2
Toilet facility	26.7	45.1	85.2
Electricity	36.6	36.7	71.8
Population and sample size			
Population share	17.1	7.91	3.2
Population in millions	171.5	79.3	32.4
Number of mothers in sample	*6901*	*3547*	*2332*
Number of live births in sample	*29426*	*10627*	*5950*

Notes:

(i) The demographic variables and the sample sizes are authors' calculations from the Indian NFHS-II and refer to the period spanned by the entire fertility history of women aged 15-49 in 1998-9. Unless otherwise indicated, figures are sample averages.

(ii) The economic variables are from World Bank (2000). Poverty incidence is for 1994, the growth rate of economy is for the period 1991-2 to 1996-7 and the ranking of states by per capita income is for 1996-97. The growth rate and rankings use the 1980/81-based GDP series.

(iii) The toilet and electricity data are from the NFHS-II Fact Sheets in the NFHS-II final report (2000).

(iv) Population is as recorded by the Registrar-General's Office of the 2000 Census on 1 July 2000.

(v) This is the average preceding birth interval and so it is calculated on a sample excluding first-born children.

Table 2: Clustering and Scarring in Sibling Infant Deaths

		Uttar Pradesh	West Bengal	Kerala
	Panel 1: Raw Data			
[1]	Incidence of infant death	0.116	0.076	0.036
[2]	Incidence of infant death excluding first-borns	0.111	0.073	0.033
[3]	*Prob*(y_{ij}=1|y_{ij-1}=1)	0.249	0.221	0.159
[4]	*Prob*(y_{ij}=1|y_{ij-1}=0)	0.093	0.064	0.029
[5]	Persistence due to y_{ij-1} (difference measure)([3]-[4])	0.156	0.157	0.130
[6]	Persistence due to y_{ij-1} (ratio measure) ([3]/[4])	2.68	3.45	5.48
[7]	Relative odds ratio	3.10	4.18	6.30
	Panel 2: Estimates			
[8]	$\hat{\gamma}$ (standard error)	0.662 (0.068)	0.429 (0.171)	0.687 (0.401)
[9]	*Exp*($\hat{\gamma}$)	1.94	1.54	1.99
[10]	*Prob*(y_{ij}=1|y_{ij-1}=1, .)	0.146	0.069	0.060
[11]	*Prob*(y_{ij}=1|y_{ij-1}=0, .)	0.083	0.047	0.032
[12]	Persistence due to y_{ij-1} (diff measure) ([10]-[11])	0.063	0.022	0.028
[13]	Persistence due to y_{ij-1} (ratio measure) ([10]/[11])	1.759	1.468	1.880
[14]	% Raw persistence explained ([12]/[5])	40.4	14.0	21.5
[15]	Predicted probability of infant death excluding first borns	0.092	0.050	0. 034
[16]	% reduction in mortality if γ=0 (with respect to [15]) ([11]*100/[15])	9.78	6.00	5.88
[17]	θ [z: θ=0] [z: θ=1]	0.695 [3.68] [-1.61]	0.944 [2.51] [-0.21]	1.795 [0.61] [0.27]
[18]	Variance of family level heterogeneity [standard error]	0.387 (0.07)	0.885 (0.24)	0.253 (0.50)
[19]	% variance explained by family level heterogeneity	10.36	21.17	7.14
[20]	Probability mass at $-\infty = p_0$ (standard error)	0.000 (0.000)	0.000 (0.000)	0.538 (0.081)
[21]	Maximised value of log likelihood	-10072.5	-2600.01	-818.7
[22]	Number of women in sample	*7297*	*3606*	*2340*
[23]	Number of children	*29937*	*10627*	*5950*

Notes:

(i) The relative odds ratio in [7] is calculated as the ratio of the odds of an infant death when the previous sibling dies to not dying. This is equivalent to the interpretation of the coefficient on y_{j-1} in the logit model.

(ii) In addition to the previous child's survival status the equations also include child gender, mother's education, father's education, an indicator for whether the child is one of a multiple birth, dummy variables denoting the birth order of the index child, indicators of ethnicity and religion, a quadratic in the age of the mother at the birth of the index child and cohort dummies. A full set of results is available from the authors. The dependent variable y_{ij} is 1 if child j in family i died before the age of 12 months and zero otherwise.

(iii)[10] is obtained by using the estimated parameters to predict y_{ij} for each observation under the condition that y_{ij-1}=1, and then averaging over all observations excluding the first-borns. [11] is similarly obtained by setting y_{ij-1}=0.

(iv) For [17], see equation (3) and Section 4.1. For [18] and [19] see Section 3. For [20] see equations (6) and (7).

Table 3

Scarring Estimates Under Alternative Sample Selections and Specifications

Specification	Uttar Pradesh	West Bengal	Kerala
1. Preferred model (Table 2, [12] and [9])	0.063 [1.94]** *(29937)*	0.022 [1.54] ** *(10627)*	0.028 [1.99]* *(5950)*
2. Drop first-borns	0.074 [2.14]** *(22640)*	0.049 [2.09]** *(7021)*	0.046 [2.97]** *(3610)*
3. Left truncate & drop first observation	0.073 [2.11]** *(22026)*	0.055 [2.26]** *(6709)*	0.038 [2.70]** *(3466)*
4. Left truncate but model first observation	0.067 [2.02]** *(29316)*	0.025 [1.61]** *(10302)*	0.025 [2.26]** *(5801)*
5. Add birth interval	0.048 [1.70] ** *(29937)*	0.015 [1.34] *(10627)*	0.021 [1.75] *(5950)*

Notes: Refer discussion in Section 7 of the text. Reported figures are marginal effects of scarring computed by the difference measure (see Notes to Table 2), and the corresponding relative odds ratios are in brackets [ ]. **, * indicate significance of the estimated coefficient, γ, at the 5% and 10% levels respectively. Figures in parentheses are the number of observations used in the estimation.

Appendix: Table 1

Means (Standard Deviations) of Variables Used in the Analysis

	INDIA	*Uttar Pradesh*	*West Bengal*	*Kerala*
Infant mortality	0.08 (0.27)	0.12 (0.32)	0.08 (0.26)	0.04 (0.19)
Infant mortality (sibling)	0.07 (0.25)	0.10 (0.30)	0.07 (0.25)	0.03 (0.16)
Female children	0.48	0.47	0.49	0.48
Multiple birth	0.01	0.01	0.02	0.02
Birth order 1	0.30	0.24	0.34	0.39
Birth order 2	0.25	0.21	0.26	0.32
Birth order 3	0.18	0.17	0.17	0.16
Birth order 4	0.12	0.13	0.10	0.07
Birth order 5	0.07	0.09	0.06	0.03
Birth order >5	0.08	0.13	0.07	0.03
Hindu	0.77	0.84	0.79	0.52
Muslim	0.12	0.15	0.19	0.31
Other religion	0.11	0.01	0.02	0.17
Scheduled caste	0.17	0.19	0.22	0.09
Scheduled tribe	0.12	0.02	0.05	0.01
Caste data missing	0.01	0.05	0.00	0.00
Ma education missing	0.00	0.00	0.00	0.00
Ma no education	0.52	0.69	0.40	0.08
Ma incomplete primary ed	0.10	0.05	0.18	0.16
Ma complete primary education	0.07	0.08	0.06	0.08
Ma incomplete secondary education	0.16	0.08	0.21	0.33
Ma secondary, higher	0.15	0.10	0.15	0.35
Pa education missing	0.00	0.00	0.01	0.00
Pa no education	0.27	0.29	0.25	0.06
Pa incomplete primary education	0.11	0.06	0.19	0.16
Pa complete primary education	0.08	0.10	0.05	0.11
Pa incomplete secondary education	0.23	0.21	0.24	0.33
Pa secondary education	0.13	0.14	0.08	0.20
Pa higher education	0.18	0.21	0.18	0.14
Age ma at birth of index child	22.8 (5.1)	23.2 (5.5)	22.0 (5.0)	23.3 (4.5)
Number of mothers	73775	7297	3606	2340
Number of children	248785	29937	10627	5950

Source: Authors' calculations based on NFHS-2 (1998-99).
Notes: ma=mother, pa=mother. Caste, Religion and Education variable means are calculated over the sample of motheres and the rest over the children.