

THE CENTRE FOR MARKET AND PUBLIC ORGANISATION

The Centre for Market and Public Organisation (CMPO) is a leading research centre, combining expertise in economics, geography and law. Our objective is to study the intersection between the public and private sectors of the economy, and in particular to understand the right way to organise and deliver public services. The Centre aims to develop research, contribute to the public debate and inform policy-making.

CMPO, now an ESRC Research Centre was established in 1998 with two large grants from The Leverhulme Trust. In 2004 we were awarded ESRC Research Centre status, and CMPO now combines core funding from both the ESRC and the Trust.



Centre for Market and Public Organisation
Bristol Institute of Public Affairs
University of Bristol
2 Priory Road
Bristol BS8 1TX
<http://www.bristol.ac.uk/cmpo/>

Tel: (0117) 33 10799

Fax: (0117) 33 10705

E-mail: cmpo-office@bristol.ac.uk

Sibling-Linked Data in the Demographic and Health Surveys

Sonia Bhalotra

October 2008

Working Paper No. 08/203

Published in Economic and Political Weekly 43 (48) pp.39-43 (2008)

ISSN 1473-625X

Sibling-Linked Data in the Demographic and Health Surveys

Sonia Bhalotra

Department of Economics & CMPO, University of Bristol

Forthcoming in a Special Issue of the Economic and Political Weekly on the NFHS

October 2008

Abstract

This paper highlights an aspect of the enormous and little-exploited potential of the Demographic and Health Surveys, namely the use of data on siblings. Such data can be used to control for family-level unobserved heterogeneity that might confound the relationship of interest and to study correlations in sibling outcomes. These uses are illustrated with examples. The paper ends with a discussion of potential problems associated with the sibling data being derived from retrospective fertility histories of mothers.

Keywords: siblings, unobserved heterogeneity, retrospective fertility histories, state dependence, DHS, India.

JEL Classification: I12, O12, J10, C23, H31.

Electronic version: www.bristol.ac.uk/cmipo/publications/papers/2008/wp203.pdf

Address for Correspondence

CMPO, Bristol Institute of Public Affairs
University of Bristol
2 Priory Road
Bristol
BS8 1TX
s.bhalotra@bristol.ac.uk
www.bristol.ac.uk/cmipo/

Sibling-Linked Data in the Demographic and Health Surveys

Sonia Bhalotra

1. Introduction

For this Special Issue, I want to highlight and illustrate an aspect of the enormous and little-exploited potential of the Indian National Family Health Survey (NFHS) and its sister Demographic and Health Surveys (DHS). My objective is to encourage researchers to develop the range of uses of the DHS data in other contexts and across academic disciplines. The NFHS is one of a family of about 200 DHS surveys conducted in some 75 developing and transition economies (see www.measuredhs.com). The aspect of unexploited potential that I focus on in this paper pertains to the use of sibling-linked data.¹ These can be extracted from the retrospective birth histories of mothers, a centrepiece of the DHS surveys. Section 2 describes in more detail how the data can be constructed and what use they can be put to. Sections 3 and 4 illustrate two applications of these data, both of which use the NFHS. One uses sibling data to control for the (endogenous) composition of births in studying the effects of business cycle fluctuations on infant mortality. The other is designed to understand the clustering of infant mortality amongst siblings and, in particular, to identify the extent to which this reflects causal processes like birth spacing over which policy interventions may have some sway, as opposed to predetermined and thereby less amenable family-level traits. Section 5 discusses potential limitations of the data and Section 6 concludes.

2. Data Structure and Uses

The DHS interview women of reproductive age, most often, age 15-49. The women record a complete retrospective history of their births and siblings are easily identified by virtue of having a common mother identifier. Since there are sequential observations of births for each mother, there is effectively a “panel” of children within mother. We may think of the mother as the cross-sectional unit and of her children, born in different years, as presenting the time dimension of the panel. The panel is “unbalanced” in the sense that the time window of the panel, which is the span of birth years of children, varies across mothers. These sorts of data are immensely useful in identifying causal effects.

A significant advantage is that they can be used to net out the effects of potentially confounding unobservables at the mother (or family) level such as ability, frailty, tastes or

¹ In the archive of DHS publications at <http://www.measuredhs.com/pubs/articles/start.cfm?selected=3>, I found 511 papers. As far as I can see none of these exploit the information on siblings.

attitudes (*unobserved heterogeneity*). The statistical model can be specified to use either *random effects* or *fixed effects*; for a discussion of these alternatives, see Hsiao (2003), for example. As these procedures rely upon variation that is common to siblings, they do not use information from mothers in the sample who record just one child. It therefore becomes relevant to confirm that the effective selection of mothers with at least two children is not biasing the results. Developing countries, which dominate the DHS sample, have relatively low age at first birth and relatively high fertility rates. Pooling information on births across the available DHS surveys, I find that the average age at first birth is 19.52 (s.d. 3.93) and that 80.5% of mothers have more than one child. The mean [median] number of children in the sample of families with at least two children is 4.26 [4]. As a result, panel data estimates are relatively well determined. I am aware of three studies that have exploited the DHS data to control for mother-level heterogeneity. Bhalotra (2007a) uses the Indian NFHS to identify the effects of aggregate income shocks on infant mortality; a more detailed discussion is below. Kudamatsu (2008) uses the African DHS data sets to identify the effect of democracy on infant mortality. Bhalotra and Steele (work in progress) use a handful of DHS data sets to examine the effects of business cycle variation on the timing of fertility. The first two studies use mother fixed effects and the third uses mother random effects.

A further advantage of linked data on siblings is that they can be used to investigate correlations in outcomes amongst siblings. Rajan and James (in this issue) look at the correlation in nutritional outcomes of siblings as an indicator of data quality. Bhalotra (2008) studies the correlation of gender of births within mother. Oettinger (2000) identifies causal effects of an individual's schooling on the schooling attainment of his or her younger sibling, after allowing for shared traits amongst siblings. In a series of papers, Arulampalam and Bhalotra (2006, 2008) and Bhalotra and van Soest (2008) analyse causal and correlated effects in the clustering of death amongst siblings. They estimate dynamic models with unobserved heterogeneity that involve isolating from the correlated traits that siblings share on account of having the same mother and environment (unobserved heterogeneity), the causal effect of an outcome for one child on the outcome for his or her succeeding sibling (this dynamic effect is referred to as *state dependence* or *scarring*). In consistent estimation of such models, an especially useful feature of the sibling linked data in the DHS is that they include information on *every* birth, including the first. This is relevant because it helps researchers address the "initial conditions problem" (Heckman 1981). This is best explained with an illustration (see section 4).

3. Removing Confounding Mother-Level Unobservables

This section illustrates the usefulness of mother fixed effects estimation in analysing the effects of economic fluctuations on infant survival; for more detail, see Bhalotra (2007a). The common expectation is that infant mortality is higher in recessions (when incomes fall) and lower in booms (when incomes rise). But what if some people anticipate this and avert birth in recessions? If at all, it seems that the people most likely to behave in this way are poor people whose ability to buffer a dip in income is limited. Indeed richer women may behave in the opposite way, preferring to give birth in recessions, when their job market opportunities are more scarce. In a scenario in which recessions are associated with a decline in the share of births contributed by poor women, the average birth will face a lower risk of infant death, other things equal.² As a result, infant mortality may appear to decline in recessions, contrary to intuition. This contrary effect is in fact a compositional (not causal) effect. There is another more autonomous reason that the data may show a decline in infant mortality in recessions. To understand this, note that infant mortality is typically measured as the proportion of *live* births in a year that do not survive to the age of one. If recessions affect maternal health and result in an increased risk of miscarriage or stillbirth, and if this effect is largest amongst poorer women then again poor women will contribute disproportionately fewer births in a recession.

The upshot is that if we want to identify the causal effect of income changes on infant mortality then part of the estimation procedure must involve controlling for the composition (or *selection*) effects described. To the extent that fertility preferences or miscarriage risk are constant (fixed) for a given woman, this can be done by use of linked sibling data. In particular, we can compare the risk of infant death of siblings, one born in a recession and one not.

A boom [recession] is defined as a positive [negative] annual change in net state domestic product per capita deflated by the consumer price index for agricultural workers. The reason to study the effects of annual *changes* in income is that longer-range growth is often entangled with a lot of other change- in technology, education, infrastructure and political and social institutions. These other things evolve sluggishly so that by looking at annual changes we can expect to difference them out. I constructed data on infant mortality for linked siblings using NFHS-2, conducted in 1998/9. The sample I analysed contains information on more than 152,000 children born to around 50,000 mothers in 15 Indian states during 1970-97. I merged these data by state and birth-year of the child with time-series data on state income, social expenditure, rainfall, etc. The length of the state panel (28 years, 1970-97) aids identification, making it more likely that there are independent macroeconomic fluctuations within states;

² This holds as long as births to poor women are more likely to die in infancy. I use the expression “poor women” somewhat loosely. The DHS do not contain data on individual wages or household incomes. The poor may be identified instead as, for example, rural or uneducated.

indeed the standard deviation of within-state income variation in these data is almost identical to that of the between-state variation.

In the rural sample, 12% of mothers have only one child and these children account for 3.7% of all children in the sample. Using a specification that includes observable mother traits, I compared the effect of income on mortality risk in the full sample with the effect obtained in a sample restricted to mothers with at least two births, and found that they were identical. This confirms that the effective selection of mothers with at least two children is not, in this case, biasing the results.

I find some evidence that births from high risk women- in particular, uneducated and scheduled tribe women in rural areas- are under-represented in recessions. Controlling for the composition of births by using mother fixed effects, I find that recessions increase rural infant mortality in India. The estimates imply that a negative income shock of median size (4.4%) will raise infant mortality by 0.136 percentage-points. This is almost half of the total annual decline in mortality in India in 1970-99 (which I estimate at about 0.3%-points p.a.). The median positive income shock in these data, at 5.8%, is larger, so the simulated beneficial effects of booms on infant survival chances are accordingly larger. The effects of income shocks on *lifetime health* will tend to be even greater since, where children survive income shocks in childhood, early exposure to poor living conditions has lasting adverse effects on their health (van der Berg et al, 2006, Banerjee et al. 2007, Deaton 2007, Bhalotra, 2007b).

The effects of recessions are not evenly distributed. The most vulnerable are rural households in which the mother is uneducated or had her first birth when she was a teenager. Within households, girls are much more likely to die in a downturn than their brothers, reinforcing previous findings that girls' welfare is put second to that of boys in lean times (Behrman and Deolalikar 1989, Rose 1999).

I investigated potential mechanisms using data on recent births in the first two rounds of the NFHS. Less than a fifth of mothers in these data contribute more than one child to the sample, and they are a select group who have shorter birth intervals than average. For this leg of the analysis, I therefore simply pool the cross-sectional data and cluster the standard errors by mother to account for non-independence of the residuals for the subset of siblings in the data. I find that delivery outside the home, antenatal care, child vaccinations and the probability of treatment for infectious diseases amongst children are lower in economic downturns. This holds even after I control for the fact that the supply of public services (proxied by data on state health and development expenditure) declines in downturns, suggesting that the demand for health services is lower. This is consistent with lower earnings in a downturn but I show that it is also

because mothers are working harder and do not have as much time to seek healthcare. So it seems that households use maternal labour supply as an insurance mechanism (with most of the additional work taken on by women being in agriculture). This imposes a cost in terms of the health of their children that has not been sufficiently recognised.

There is an interesting contrast here with recent results for richer countries, where women's work is thought to be procyclical, and this is hypothesised to contribute to the seemingly counter-intuitive finding that infant mortality is procyclical (higher in upturns); see Dehejia and Lleras-Muney (2004). Overall, the results suggest a need for mechanisms that shield the vulnerable from temporary falls in wages or increases in unemployment (of main earners) that we find have irreversible consequences. They show that (temporary) increases in labour force participation amongst relatively poor women may signal distress rather than "liberation" and have unintended consequences for child survival and health.

The analysis illuminates a question of long-standing policy interest concerning the extent to which income (or poverty) is an ultimate cause of childhood death in poor countries. This question is informative of the welfare effects of economic growth, and of the effectiveness of cash transfers made to poor households. We may expect growth in aggregate income (GDP) to lower mortality if it (a) raises private incomes, especially of the poor, so that they can acquire more nutrition and other health inputs and, (b) raises state social expenditure. However, the evidence on the effectiveness of income in improving survival is not overwhelming. Historical evidence suggests that secular improvements in medical technology, public services and education were more important than income growth in bringing about sustained mortality decline (Cutler et al. 2006). And as discussed, recent studies of the United States and other OECD countries show that mortality risks –for adults and children- are *lower* in recessions (Ruhm 2000). Against this backdrop, the analysis of income and substitution effects in the poorer setting of India is pertinent.

4. Identifying Dynamic Effects

In this section, I consider an application in which the question of interest concerns the effect on a child of an outcome for his or her preceding sibling (what we earlier referred to in section 2 as state dependence or scarring). The outcome of interest is infant or neonatal death. The phenomenon of interest is the clustering or concentration of death risk amongst siblings (Curtis et al. 1993, Zenger 1993). To illustrate the importance of this phenomenon, consider the following figures derived from NFHS-2 for the state of Uttar Pradesh. Amongst second and higher-order children, the average probability of neonatal death is 5.2% in the sub-sample in

which the previous sibling survived the neonatal period. In contrast, in the sub-sample in which the previous sibling suffered neonatal death, this probability is a remarkable 18.80%. So the death of a preceding sibling is associated with a more than three-fold increase in mortality risk.

Some of the reason for sibling death clustering is no doubt that families are different, with some being more effective at averting child death than others. For example, mothers in the more effective families are probably more educated, more aware and innately healthier than in the less effective families. The DHS (and most data sets) record education but they do not record awareness or frailty. To comprehensively capture these “unobservables”, we can use mother-level fixed or random effects, and we choose to use random effects (see Arulampalam and Bhalotra (2006, 2008), Bhalotra and van Soest (2008)).

Now consider how and why dynamics come in to this picture. An especially interesting question for policy is the extent to which the event of death of a child *causes* a higher risk of death for his or her younger sibling. A potential mechanism is as follows. The death of a neonate or infant will result in the cessation of breastfeeding. Since breastfeeding delays the return of fecundity following birth, this will tend to reduce the birth interval to the next child. And there is a lot of evidence to suggest that short preceding birth intervals elevate mortality risk. So in an appropriate model, neonatal or infant mortality risk for child j of mother k will depend upon the realised neonatal or infant mortality of child $j-1$ of mother k [state dependence] as well as on all traits specific to mother k [mother-level unobserved heterogeneity]. A classical problem with consistent estimation of this sort of model is that the realised mortality of the preceding sibling is necessarily correlated with mother-level heterogeneity. This is called the initial conditions problem (Heckman 1981), and a good way to attempt to address it is to use information on the first-born child in each family. 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, while the first child presumably has no history. So it is important to use information on whether the first child died or not. In this respect, the availability of complete birth histories of married women in the DHS surveys is a valuable resource. In other applications of dynamic models with unobserved heterogeneity, data on the start of the process are often 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 retrospective data, that is, data that do not include the first spell of unemployment for each individual.

The main findings of this research are as follows. 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. In a model that allows for scarring, the proportion of the error variance attributable to mother-level unobservables is estimated to be 11% in Uttar Pradesh, 21% in West Bengal and 7% in Kerala. We estimate that eliminating scarring 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 (Arulampalam and Bhalotra 2006). Scarring is large and significant in thirteen of the fifteen major states of India. The two states in which evidence of scarring is weak are Punjab, the richest, and Kerala, the socially most progressive. The size of the scarring effect depends upon the gender of the previous child in three of the fifteen states, in a direction consistent with son-preference (Arulampalam and Bhalotra 2008). The only other covariate with an effect of a similar magnitude to scarring is mother's (secondary or higher) education. While there is considerable evidence of the effects of maternal education on infant survival, the literature has paid scarce attention to scarring effects. Evidence of scarring implies that policies targeted at reducing infant mortality will have social multiplier effects through helping avoid the death of subsequent siblings. Mechanisms underlying scarring were further investigated for the state of Uttar Pradesh in Bhalotra and van Soest (2008), who exploit the panel nature of the data to model birth spacing and fertility jointly with (neonatal) mortality. The tendency for neonatal death to shorten the birth interval to the next child explains about a fourth of the total scarring effect. We speculate that some of the residual effect may involve maternal depression.

We find direct evidence of replacement behaviour: a child death results in a shortening of the interval to the next birth, and also increases the probability of a next birth. Model simulations imply that, accounting for direct and indirect effects, 37 in 100 children who die during the neonatal period are replaced by new births. Of these, about 30 survive. There is no evidence that couples practice hoarding, that is, that they have higher fertility in anticipation of the risks of neonatal death. The estimates of fertility behaviour are consistent with son-preference. The estimates suggest that fertility decline in India started in 1981. Despite this, birth intervals have got shorter since about then.

5. Potential Problems

Having elucidated some of the less recognised or less used advantages of the retrospective fertility histories in the DHS, it is relevant to point out some of their limitations.

We have highlighted in section 4, the advantage of having information on all births to a mother, including her first, as this helps address the initial conditions problem. A consequence of this is that the range of birth years in the data is wide, and varies by mother. The long time span for which cohort infant mortality is available is highlighted as an advantage in section 3, as it assists in identifying the effects of within-state changes in income on mortality. However, we may be limited in the range of variables that can be associated with information on births in a time-consistent fashion. For instance, even if we have data on infant mortality rates in 1970 from a survey conducted in 1992, we do not have detailed information on the mother's (or family's) circumstances (e.g. assets, location, labour force participation) over time, rather, this information pertains to the year of the survey.

Location is of particular interest if we want to match state level factors like political or macroeconomic indicators to outcomes for children born across long periods of time. This is because if the mother migrated between births then we cannot assume that all of her children were born in the state that she was in when interviewed. This problem can be assessed if not addressed by using answers to a question in the DHS that asks how long the mother has lived in her current place of residence (i.e. place at time of survey). Using NFHS-2, I estimate that 86% of children born in 1970-97 in the 15 major states were born in the mother's current place of residence. This refers to her local location and will record, for example, movement between villages within a state. So it is a much stricter condition than required. If migration is exogenous to the question of interest, then the parameter of interest can be re-estimated on the sub-sample of children for whom the state of birth is known with certainty and the problem is solved. However if migration is endogenous then it is not, although this exercise may nevertheless be informative. And it may be useful to estimate the model on the full sample and the restricted sample (of non-migratory mothers) and compare the estimates.

Another potential problem with retrospective data is recall bias: the concern that mothers are more likely to forget the incidence or dates of events the further back in time they are. This may be especially pertinent if mothers with different characteristics (e.g. age, education) do this to different extents. The DHS surveys do have numerous checks built in to ensure the quality of birth history data (see ORC Macro 2006, p.14) but researchers might nevertheless assess the extent of this problem by, for example, checking the sensitivity of their results to left-truncation of the data- that is, to restricting the sample to windows of more recent years. This is done, for example, in Arulampalam and Bhalotra (2006) and in Bhalotra (2007a). Another recall-related issue is rounding off in age. For example, data on age at death for children reveal age-heaping at six-month intervals. If one is studying a discrete event such as death by the age of one (infant

mortality), a natural sensitivity check would be to conduct the analysis with mortality defined inclusive and exclusive of deaths in the twelfth month.

There are other problems relating to how representative the retrospective sample is (e.g. Rindfuss, *et. al.*, 1982). For example, as we go back in calendar time, the births captured in the sample are not only fewer but also disproportionately of relatively young mothers. For example, a woman who gave birth at age 15 in 1965 will be 49 in 1999, and her birth will be recorded. However, births to women any older than 15 years in 1965 will not be recorded since then women will be older than 49 years in 1999. If young mothers are more prone to infant mortality risk then these data will over-estimate mortality rates in the earlier years and, consequently, under-estimate the trend decline in mortality. In multivariate analyses, this problem can be addressed by conditioning upon maternal age at birth. Another problem is that the survey does not, of course, record births of mothers who died before the date of the interview. If it is the frail or poor mothers who die early then we will have a selectively low-risk sample of children, especially for older cohorts of mothers.

6. Concluding Remarks

This paper has argued that the DHS offer a wealth of sibling data that can be used either to control for family-level unobserved heterogeneity that might confound the relationship of interest, or to study correlations in sibling outcomes. These data emerge from retrospective fertility histories of mothers, as a result of which they face problems of “time consistency” of covariates, recall and selectivity. We have discussed some of these problems and offered some tentative solutions or indications of the direction of likely bias where relevant.

References

- Arulampalam, Wiji and Sonia Bhalotra (2006), Sibling death clustering in India: State dependence vs unobserved heterogeneity, *Journal of the Royal Statistical Society, Series A* 169(4): 829-848.
- Arulampalam, Wiji and Sonia Bhalotra (2008), The linked survival prospects of siblings: Evidence from the Indian states. *Population Studies*, July.
- Banerjee, A., E. Duflo, G. Postel-Vinay and T. Watts (2007), Long run health impacts of income shocks: Wine and phylloxera in 19th century France, Mimeograph, January.
- Behrman, Jere and Anil Deolalikar (1989), Seasonal Demands for Nutrient Intakes and Health Status in Rural South India, in David Sahn, ed., *Seasonal Variability in Third World Agriculture: Consequences for Food Security*, Baltimore: Johns Hopkins University Press.

Bhalotra, Sonia (2007a), Fatal fluctuations: Cyclicity in infant mortality in India. IZA Discussion Paper 3086, Bonn. Revised version available from the author on request.

Bhalotra, Sonia (2007b), Wuthering heights: Birth shocks and stature in India, Mimeograph, University of Bristol. Conference paper presented at the Platinum Jubilee Conference of the Indian Statistical Institute, Delhi, December 2007 (with the title Little Women).

Bhalotra, Sonia (2008), Sisters, brothers and clusters. Mimeograph, University of Bristol.

Bhalotra, Sonia and Arthur van Soest (2008), Birth-spacing, fertility and neonatal mortality in India: Dynamics, frailty and fecundity, *Journal of Econometrics*, April 2008.

Bhalotra, Sonia and Fiona Steele, The effects of income shocks on the timing of fertility. Work in progress, 2008.

Curtis, Sian L., Ian Diamond and John W. McDonald. 1993. Birth interval and family effects on post-neonatal mortality in Brazil, *Demography* 33(1): 33-43.

Cutler, David, Angus Deaton and Adriana Lleras-Muney (2006), The determinants of mortality, *Journal of Economic Perspectives*, 20(3), Summer.

Deaton (2007) 'Height, health and development' *Proceedings of the National Academy of Sciences of the United States of America* Vol. 104, No. 33, pp13232-13237

Dehejia, Rajeev and Adriana Lleras-Muney (2004), Booms, Busts, and Babies' Health, *Quarterly Journal of Economics* 119(3): 1091-1130.

Hsiao, Cheng (2003), *Analysis of Panel Data*, 2nd edition, Cambridge University Press (Econometric Society Monographs, No. 34).

Kudamatsu, Masayuki (2008), Has Democratization Reduced Infant Mortality in sub-Saharan Africa? Evidence from Micro Data. Mimeograph, IIES, Stockholm University.

Oettinger, Gerald S. 2000. Sibling similarity in high school graduation outcomes: causal interdependency or unobserved heterogeneity?, *Southern Economic Journal* 66(3): 631-648.

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.

Rose, Elaina (1999), Consumption smoothing and excess female mortality in rural India, *Review of Economics and Statistics*, 81(1), February, 41-49.

Ruhm, Christopher (2000), 'Are Recessions Good for your Health?', *Quarterly Journal of Economics* 115(2): 617-50.

van der Berg, Gerard, Maarten Lindeboom and France Portrait (2006), 'Individual Mortality and Macro-economic Conditions from Birth to Death', *American Economic Review* 96(1): 290-302.

Zenger, Elizabeth. 1993. Siblings' neonatal mortality risks and birth spacing in Bangladesh, *Demography* 30(3): 477- 88.