

# Death by Survey: Estimating Adult Mortality without Selection Bias<sup>1</sup>

Emmanuela Gakidou<sup>2</sup> and Gary King<sup>3</sup>

July 4, 2005

<sup>1</sup>The current version of this paper is available at <http://GKing.Harvard.edu>. Our thanks to Andrew Gelman, Michel Gillot, Ken Hill, Ian Timaeus, and Langche Zeng for helpful comments and the National Institutes of Aging (P01 AG17625-01) and the National Science Foundation (SES-0318275, IIS-9874747) for research support.

<sup>2</sup>Research Associate, Institute for Quantitative Social Science, Harvard University (34 Kirkland Street, Harvard University, Cambridge MA 02138; [Emmanuela.Gakidou@Harvard.edu](mailto:Emmanuela.Gakidou@Harvard.edu), (617) 496-6132).

<sup>3</sup>David Florence Professor of Government, Department of Government, Harvard University (Institute for Quantitative Social Science, 34 Kirkland Street, Cambridge MA 02138; <http://GKing.Harvard.edu>, [King@Harvard.Edu](mailto:King@Harvard.Edu), (617) 495-2027).

## **Abstract**

The widely used methods for estimating adult mortality rates from sample survey responses about the survival of siblings, parents, spouses, and others depend crucially on an assumption that we demonstrate does not hold in real data. We show that when this assumption is violated — so that the mortality rate varies with sibship size — mortality estimates can be massively biased. By using insights from work on the statistical analysis of selection bias, survey weighting, and extrapolation problems, we propose a new and relatively simple method of recovering the mortality rate with both greatly reduced potential for bias and increased clarity about the source of necessary assumptions.

# 1 Introduction

Public policy makers, and medical and public health researchers, require reliable mortality data to understand, evaluate, and eventually ameliorate problems in national health systems. Yet, only a minority of the world's countries have complete vital registration systems, and demographic surveillance systems are only occasionally feasible and then only in a few isolated areas. Political scientists often try to explain and predict elite decisions to go to war rather than the more ultimate dependent variables like human misery or mortality in part because measures of such concepts are often unavailable or flawed and because vital registration systems rarely operate in wartime or in war-prone regions (see Murray et al. 2002 and Ghobarah, Huth and Russett 2003). For these and other purposes, a wide variety of social scientists and policy makers need high quality mortality data and have similar problems obtaining it in the places and times when it is needed most.

These problems have led to extensive efforts to develop and apply methods of estimating mortality rates from sample surveys of relatives or acquaintances. Hundreds of applications of these methods have appeared in demography, epidemiology, sociology, public health, and medicine, with scholars creating and using methods to estimate mortality (and other vital rates) from information collected about deaths among household residents (Feeney, 2001; Graham, Brass and Snow, 1989), siblings (Bicego, 1997; Chipangwi et al., 1992; Danel et al., 1996; Gakidou, Hogan and Lopez, 2004; Garenne and Friedberg, 1997; Graham, Brass and Snow, 1989; Shahidullah, 1995; Shiferaw and Tessema, 1993; Timaeus and Ali, 2001; Walraven and van Dongen, 1994; Wirawan and Linnan, 1994), parents (Brass and Hill, 1973; Hill and Trussell, 1977; Timaeus, 1991b, 1986), and spouses (Malaker, 1986; Stanton, Nouredine and Hill, 2000; Singh, 2000; Timaeus, 1991).

The Demographic and Health Surveys (DHS) program has invested heavily in collecting complete birth histories of nationally representative samples of women. This program has led to accurate child mortality estimates for a large number of countries without vital or sample registration systems since the 1980s. Some relevant data for measuring adult mortality are also being collected through many household surveys, such as the DHS and the World Health Survey, but current methods of using this information suffer from selection bias. The task now is to develop a method that uses these data to produce accurate estimates, at which point the incentives may be in place to begin to collect more

extensive and accurate information on the survival of adult relatives.

Collecting data from those alive only at the end of a period of interest makes the method inexpensive and thus feasible, but it leads to a serious selection bias problem: Individuals from high mortality families are less likely to appear in a survey as respondents. The current literature approaches this issue by making a mathematical assumption that avoids selection bias only if the assumption holds empirically. We show that this assumption — that mortality does not differ by sibship size — is violated in practice in most available data sets. We therefore develop a new approach that avoids this assumption altogether, by dividing the task into a component that can be corrected exactly via weighting and one that requires extrapolation from observable patterns. We offer theoretical, simulation, and empirical evidence that the new method is to be preferred in all known situations to the existing approach in the literature.

Section 2 introduces our notation, and defines the existing approaches and our method of estimating mortality rates from a survey about sibling survival in the simplified context of a fixed cohort. Section 3 then demonstrates how our approach works in a variety of types of data, including those situations that generate biased results under the standard approach in the literature. Section 4 demonstrates empirically that the assumption of the standard method is invalid in a wide range of countries, and that our estimation method seems to work in real data. Section 5 then generalizes these results to apply to information collected about relatives other than siblings, and to period-based quantities of more practical interest. Section 6 concludes.

## 2 Probability of Death in a Cohort

**Preliminary Notation** Let  $j$  ( $j = 1, \dots, N$ ) denote an index for an individual in a population of interest at time 1. Denote by  $B_j$  the number of siblings in the family of respondent  $j$  (including respondent  $j$ ) at the Beginning of the period (or “Born” into the group at time 1), by  $S_j$  the number of siblings in the family of respondent  $j$  who Survive to time 2, and by  $D_j$  the number who Die between times 1 and 2, so that  $B_j = S_j + D_j$ . The proportion of those who die in this family is the Mortality rate, calculated as  $M_j = D_j/B_j = (B_j - S_j)/B_j$ . This notation thus applies to a cohort where all group members begin at time 1 at the same age (such as 40 year old men in Uganda) and end in time 2 at

the same age (as, e.g., 45 year old Ugandan men) if they do not die in the interval between times 1 and 2. Section 5 changes from cohort to more practically useful period quantities.

We are interested in drawing a sample of survivors at time 2 to infer the mortality rate or other quantities from the full sample identified at time 1. We assume for convenience the absence of measurement error. Of course, in applications researchers will need to follow all the standard techniques of survey design, such as pretesting and cognitive debriefing, to avoid recall bias and other potential sources of error, none of which are addressed by the methods discussed here.

**Quantity of Interest** We now define the quantity of interest  $q$ , the probability of death (or the proportion of those in the population who die) for people in the interval from time 1 to time 2. To do this in an informative way, we first define  $d_j$  as 1 if individual  $j$  dies between time 1 and 2 and 0 if  $j$  survives. This quantity can be expressed in three equivalent ways:

$$q = \frac{\sum_{j=1}^N d_j}{N} = \frac{\sum_{f=1}^F D_f}{\sum_{f=1}^F B_f} = \frac{\sum_{j=1}^N M_j}{N} \quad (1)$$

The first expression is perhaps the most obvious definition of the mortality rate. The second defines  $q$  for information collected at the family level or with one respondent per family ( $f = 1, \dots, F$ ). The third is defined for the family mortality rate at the individual level, for all individuals in the population.

If it were possible to draw a random sample from individuals at time 1, the first and the third definitions would provide unbiased and consistent estimators; the second would be consistent if one could sample families (or one person from each family) at time 1. None would be unbiased if applied to a time 2 sample of survivors, the correction of which is our goal.

To be clear about the notation, each individual surveyed provides information about members of his or her entire sibship or, in other words, family-level information about  $B$ ,  $S$ , and  $M$  (e.g.,  $M_j = M_{j'}$  for all  $j$  and  $j'$  that are members of the same sibship). Thus, if we could sample at time 1, each draw of an individual is equivalent to a draw of a family selected with probability proportional to  $B_j$ . For example, families with five siblings are represented in the population with five times the frequency, and thus have five times the sampling weight, as a family with one sibling. The problem posed here is to estimate  $q$

from the “biased” random sample of those surviving to time 2 rather than the definition, which is representative of the full population at time 1.

**Existing Mortality Estimators** We now define in our notation two existing estimators of mortality from the sample available at time 2. To do this, we introduce index  $i$  ( $i = 1, \dots, n$ ) for respondents that have survived to time 2 and thus appear in the time 2 sample and are observed ( $n \leq N$ ).

The first existing estimator is what we refer to as the *naive* estimator and is merely the ratio of the total number of deaths to births reported by the survey respondents:

$$\dot{q} = \frac{\sum_{i=1}^n D_i}{\sum_{i=1}^n B_i}. \quad (2)$$

This estimator has an obvious and massive selection bias problem since respondents from families with high mortality are underrepresented in the sample. Respondents from families with no survivors have zero probability of making it into the sample and so are not counted at all. In addition, by design, every sample contains information on  $n$  people (the respondents themselves) about whom we have no uncertainty and thus learn nothing, since they would not have been selected unless they survived. As a result of the selection bias problem, the naive estimator will under most circumstances underestimate the true mortality rate.

Second is the *standard* estimator. This approach eliminates data that contain no information by omitting self-reports from the denominator (no modification of the numerator is necessary since it is zero for all survey respondents):

$$\check{q} = \frac{\sum_{i=1}^n D_i}{\sum_{i=1}^n B_i - n}. \quad (3)$$

Trussell and Rodriguez (1990) point out three sources of bias in this method: The respondent (who is of course always alive) is not counted at all, which presumably biases mortality estimates upward; the mortality experience of the respondent’s siblings may be counted multiple times if they are all interviewed, and so families with low mortality may be overrepresented, resulting in mortality being underestimated; and families with no survivors are not represented in the sample at all, which will bias mortality estimates downward. Trussell and Rodriguez (1990) then prove the remarkable result that *if mortality does not vary with sibship size* these biases cancel out and  $\check{q}$  is itself unbiased.

The assumption is critical: The estimator  $\check{q}$  will be biased when applied to data where any predictable relationship exists between sibship size and mortality. A causal relationship between these two variables may be the reason for the relationship, but even a noncausal, spurious relationship will generate bias. Thus, bias would result if mortality were positively correlated with sibship size, for example if people in high mortality areas have more children than in low mortality areas, or if children in large families have fewer resources and thus higher mortality. Bias would also result in the reverse situation where mortality is negatively correlated with sibship size. Any correlation between fertility and mortality will generate bias, no matter the reason.

Although this unbiasedness condition likely only applies to real data in rare instances (Zaba and David, 1996), Trussell and Rodriguez’s mathematical result demonstrating unbiasedness when this condition holds is nonetheless vitally important. It demonstrates that there exist conditions where it is possible to infer mortality in a population from a sample selected in a biased but convenient way. And more importantly, once an assumption is highlighted and clarified, it is often possible to eliminate it altogether, a task to which we now turn.

**Our Estimator** We now build an estimator that requires no assumption about the relationship between fertility and mortality. The key is to recognize that sampling at time 2 generates two separate problems. The first is that selecting respondents at time 2 with equal probability is equivalent to sampling families proportional to  $S_i$  (the number of siblings surviving to time 2 for person  $i$ ) rather than  $B_i$  (the number of siblings at time 1 for person  $i$ ). Fortunately, both quantities are known for all observations sampled and so, to return to the desired  $B_i$  weighting, we replace the simple average of  $M_j$  in the last expression in (1) with the weighted average of  $M_i$ , using weight  $W_i = B_i/S_i$ :  $\sum_{i=1}^n M_i W_i / \sum_{i=1}^n W_i$ . The weighting solves this portion of the problem with no uncertainty except for the usual sampling variability and measurement error. That is, using weights as we suggest here means that the first problem with time-2 sampling vanishes entirely (so that the estimate is exactly equal to the quantity of interest) in a census, in a sample as  $n$  increases, or on average for any fixed sample size.

The second problem with the sample drawn at time 2 is that families with no survivors ( $S_i = 0$ ) are not represented at all, and so weighting to recover the full information is

impossible. To be more precise, the missing information is the total number of siblings in families with zero survivors, which we denote with the Greek letter zeta,  $\zeta$ , and which needs to be added to both the numerator and denominator of the weighted average, since for this group  $B_i = D_i$ . With an estimator for  $\zeta$ , which we denote as  $\hat{\zeta}$ , our estimator of the mortality rate will be

$$\hat{q} = \frac{\sum_{i=1}^n M_i W_i + \hat{\zeta}}{\sum_{i=1}^n W_i + \hat{\zeta}}. \quad (4)$$

Before discussing how to estimate  $\zeta$ , we offer an alternative interpretation of Equation 4 useful for intuition and for later generalizations. Thus far we have distinguished between two overlapping groups: The original population followed between times 1 and 2, and the respondents drawn randomly at time 2 from those who have survived. If we could apply one of the simple expressions in Equation 1 to the population, we would recover the quantity of interest  $q$ , since of course this is how we define  $q$ . If instead we had a random sample from this population, and could elicit information about each person's mortality during times 1 and 2, and that of his or her sibship, applying the first simple expression in Equation 1,  $q = \sum_{j=1}^N d_j/N$ , to the sample would yield an unbiased estimate of  $q$ .

Since bias would result if we applied the same uncorrected estimator to the observed time 2 sample (and we do not observe the time 1 population or a sample from it), we construct a *pseudo-sample* of the time 1 respondents, from the information in our time 2 sample. The pseudo-sample contains data that would not result in bias when applying the estimators in Equation 1. We do this by rewriting Equation 4 as

$$\hat{q} = \frac{\text{Deaths}}{\text{Deaths} + \text{Survivors}} = \frac{\left[ \sum_{i=1}^n (D_i/S_i) + \hat{\zeta} \right]}{\left[ \sum_{i=1}^n (D_i/S_i) + \hat{\zeta} \right] + n}, \quad (5)$$

where “Deaths” and “Survivors” in first expression refer to the totals in the pseudo-sample. The entire first expression in this equation is the sample analogue to the first expression in Equation 1 for the population.

So far, the only constraint we have put on the pseudo-sample is that the simple estimators would yield unbiased estimates, but this only constrains the ratio in the first expression above to be correct, not Deaths or Survivors alone, or their sum. To make real calculations, we need to constrain one of these (although the specific constraint is arbitrary and will not affect our estimate of  $q$ ). Thus, we add the arbitrary constraint that the number of survivors in the pseudo-sample equals the number of respondents in

our observed time 2 sample, so that “Survivors” =  $n$ .

The remaining task is to compute the number of deaths in the pseudo-sample, adjusted to be relative to the fixed number of survivors. The last expression in Equation 5 shows how to do this by decomposing the number of deaths into the sum of two parts: “Deaths” =  $\sum_{i=1}^n (D_i/S_i) + \hat{\zeta}$ . To get the first component, we need to know the number of deaths for each survivor, which is  $D_i/S_i$ , and to add these up for all  $n$  survivors. The second component is the number of deaths in families with zero survivors,  $\zeta$ . The time 2 sample reveals the first component directly, and we need to estimate the second.

We now turn to estimating  $\zeta$ , the final task of this section.<sup>1</sup> Although no certain or directly estimable information about  $\zeta$  exists in a sample drawn at time 2, it turns out that informative statistical information does appear to exist. We thus extrapolate to these quantities from information in the sample. To do this, we first compute the total number of deaths in the time 1 pseudo-sample from families with  $s$  survivors (for  $s = 1, 2, \dots$ ) and fit a model predicting this with  $s$ . We then use the same model to extrapolate these back to the (unobserved) number of deaths from families with  $s = 0$  survivors, which gives us an estimate of  $\zeta$ .

One approach is to regress the log of total deaths from families with  $s$  survivors in the time 1 pseudo-sample on a quadratic function of  $s$  for  $s = 1, \dots, 7$ . (We exclude death proportions from  $s > 7$  because they are based on too few respondents and are thus noisier and less useful for extrapolating all the way back to  $s = 0$ .) That is, we run a linear regression of  $\ln(\sum_{\{i:S=s\}} D_i/S_i)$  for  $s = 1, \dots, 7$  on a constant,  $s$ , and  $s^2$ . We then transform the constant term,  $\hat{\alpha}$  (which is the predicted value of the number of deaths in the pseudo-sample for  $s = 0$ ), to obtain an estimate of  $\zeta$ .<sup>2</sup>

Although this simple quadratic model seems to fit real data well, nothing can guarantee that an extrapolation (i.e., an inference outside the range of observable data) will always

---

<sup>1</sup>We might think about factoring this number as  $\zeta = F\bar{\tau}$ , where  $\bar{\tau} = \sum_b b\tau_b$  is the expected number of siblings in families without survivors, and  $F$  is the number of families. However, although we could estimate  $\bar{\tau}$  from data collected at time 2, the only way to estimate  $F$  from survey data would be to have some idea of how many people were interviewed from the same sibship. However, establishing which survey respondents are from the same sibship is infeasible in most contexts and requires data not collected in any major national survey. We will therefore attempt to estimate  $\zeta$  directly without this or any other decomposition.

<sup>2</sup>One might think that we could merely exponentiate the constant term,  $e^{\hat{\alpha}}$ , to remove the log scale but this procedure is biased because the expected value of the log (which the regression estimates) is not equal to the log of the expected value (which this calculation would produce). A better procedure is either to simulate or to use the simple analytical solution based on the expected value of a log-normal density:  $e^{\hat{\alpha} + \hat{\sigma}^2/2}$ , where  $\hat{\sigma}$  is the standard error of the regression (see King, Tomz and Wittenberg, 2000).

be accurate (King and Zeng, 2004). It is always possible that total deaths given  $s$  follows a completely different pattern for the unobserved point where  $s = 0$  than for the observed points where  $s > 0$ . What makes us somewhat optimistic are experiments discussed in Section 4 with data from 24 countries where we set aside data we observe in each and try to predict it from the rest of the observed data in that country; these experiments work out well in a wide variety of countries. For example, we fairly accurately predict the number of deaths from families with one survivor ( $s = 1$ ) using only the death data for families with more than one survivor ( $s > 1$ ). We also predict accurately the number of deaths from families with two survivors ( $s = 2$ ) using deaths observed at  $s = \{1, 3, 4, 5, 6, 7\}$ , etc.

To emphasize the uncertain nature of extrapolation, we briefly discuss another approach, which is to regress the log of deaths in the observed sample,  $\ln(\sum_{\{i:S=s\}} D_i)$  rather than the pseudo-sample,  $\ln(\sum_{\{i:S=s\}} D_i/S_i)$ , on a quadratic function of the number of survivors  $s$ . This would appear wrong except that the last observed point before extrapolation occurs at  $S_i = 1$ , where the two are equivalent. This approach is slightly closer to the observed data and we still may be extrapolating to the number of deaths in the time 1 pseudo-sample. We find that this approach fits the data slightly better and so usually stick with it, but what to do in any instance is of course an important substantive judgment.

Our ultimate estimator for the mortality rate from a cohort sample is then Equation 4 with this estimate for  $\zeta$  from the quadratic extrapolation substituted in. The uncertainties in this approach are due to sampling error, which vanishes as the sample size increases, and specification uncertainty due to the model used for the extrapolation necessary to estimate  $\zeta$ . Typically available sample sizes mean that normally only the latter is a significant concern.

An advantage of our approach is that it isolates the piece of the problem not amenable to direct statistical estimation so that the extrapolation model cannot affect inferences about families with some survivors. The same extrapolation issue in our estimator exists in both previously existing approaches, the only differences being that the extrapolation is hidden in other calculations in those approaches, does not adapt to changes in the observed data, and affects inferences about all families. Of course, the necessity of extrapolation is a property of the problem of estimating mortality by survey rather than a property

of any one method. Standard errors or confidence intervals intended to represent these uncertainties in the current approach other than model dependence can be computed via bootstrapping. We should expect model dependence to be larger in groups where  $\zeta$  is likely largest, such as with high mortality and low fertility rates.

### 3 Simulation Evidence

We now compare the naive and standard estimators with our new estimator in the usual way by evaluating bias and mean square error. We do this by Monte Carlo simulation. We create 27 scenarios by cross-classifying low (0.1), medium (0.2), and high (0.3) average mortality with average fertility levels approximately representing **Kenya** (4.26 children), **Turkey** (3.07), and **Kazakhstan** (2.56) (with colors corresponding to our graphs, below), and positive, zero, and negative correlations between family size and mortality. For each of these 27 scenarios, we create 1,000 data sets, each with  $n = 1,000$  randomly drawn time 2 survey respondents. For each data set, we compute each of the three estimators and evaluate bias and mean square error. We then quantify how much each of the two corrections contributes to the bias reduction in our estimator.

**Bias** The degree of bias is defined for estimator  $\hat{q}$  as  $\text{Bias}(\hat{q}) = E(\hat{q} - q)$ , where  $E(\cdot)$  take the average over repeated samples. We approximate this quantity, and the bias for the other two estimators, by subtracting the true mortality from the mortality estimated from each of the 1,000 data sets and averaging. Figure 1 portrays these results. The three graphs in this figure portray data from positive (left graph), zero (middle graph), and negative (right graph) correlations between sibship size and mortality. True mortality is displayed horizontally and bias in an estimator vertically, with a line drawn to denote zero bias. The three fertility levels are displayed in different colors.

In all three graphs (and thus correlation levels), we denote our estimator by a filled circle, all 27 of which are fairly near the zero-bias line. The deviation from zero bias is due only to estimating  $\zeta$ , since it requires extrapolation. The other portion of the estimator is an exact correction (so that if we knew the true  $\zeta$  and used it, our estimator would be exactly on the zero-bias line). As expected, bias in the naive estimator, which we denote in the graph by an asterisk, is always well below the line, indicating that it is underestimating mortality, no matter what the correlation is. Finally, we plot the

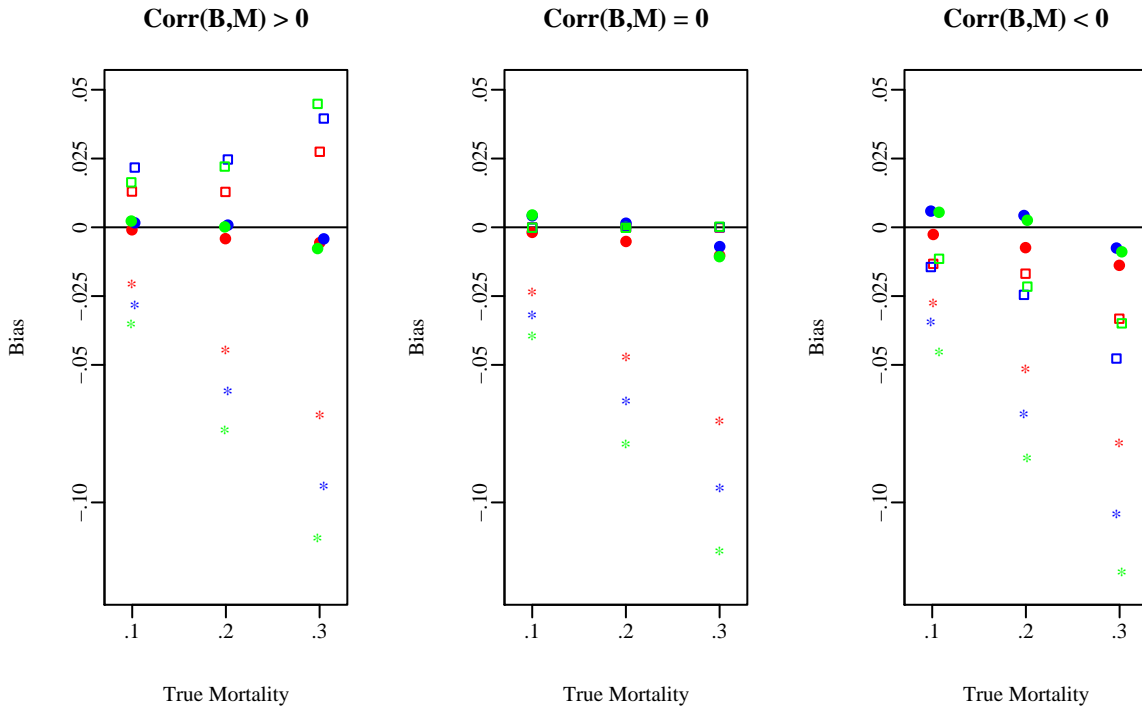


Figure 1: Bias in Mortality Estimates. Closed circles represent our estimator (all near zero bias), open squares represent bias in the standard estimator, and asterisks are for the naive estimator. Simulations were drawn with a correlation between mortality and fertility that is positive (left graph), zero (middle), and negative (right). The figure shows simulations drawn with three levels of mortality (0.1, 0.2, and 0.3, on the horizontal axis) and **low** (green), **medium** (blue), and **high** (red) fertility.

bias in the standard estimator widely used in the literature via a square. The result shows that for data where sibship size is positively correlated with mortality, the standard estimator markedly overestimates mortality (note the positive bias portrayed above the line for all squares in the left graph). When sibship size and mortality are negatively correlated (portrayed in the right graph), the standard estimator is substantially biased in the opposite direction, indicating that it underestimates mortality. As expected, when the assumptions of the technique happen to hold in the data (i.e., when sibship size and mortality are uncorrelated, as in the middle graph) the estimator is unbiased.

Note that for all three estimators, the absolute level of bias increases to some degree with the true level of mortality. This pattern exists because higher average mortality normally signifies a larger number of deaths and variance in the possible numbers of deaths from families with zero survivors — which is information not directly represented in the sample.

Bias in the standard estimator and our approach do not appear to increase or decrease systematically with different fertility levels. In contrast, the naive estimator is clearly worse with lower levels of fertility (the green asterisks appear on the graph lower, indicating larger absolute bias, than the blue which are in turn more biased than the red), since in these situations the apparent information from the respondent’s self-reports of their own (lack of) mortality represent a larger fraction of the data used in the naive estimator.

**Mean Square Error** When approximate unbiasedness is used as a criterion in statistical inference, it pays to check that unbiasedness is not being achieved at a cost in higher variance. For this purpose, mean square error is normally used. The mean square error is the squared difference between the estimator and the truth, and it factors into the squared bias plus the variance. For example, the mean square error of our estimator  $\hat{q}$  is:

$$\begin{aligned} \text{MSE}(\hat{q}) &= E[(\hat{q} - q)^2] \\ &= V(\hat{q}) + \text{Bias}(\hat{q})^2. \end{aligned} \tag{6}$$

We follow convention in presenting results in terms of the square root of mean square error (or RMSE) since it is on the scale of mortality.

The results are presented in Figure 2, where we use the same setup as in Figure 1 of graphs, colors, and symbols to represent correlation, average fertility, and estimators, respectively. In these graphs, RMSE is on the vertical axis, where higher on the graph indicates larger values of RMSE and an inferior estimator. The results here parallel that for bias: The naive estimator has the highest (worst) RMSE for all scenarios. Our estimator has lower RMSE than the other two for negative and positive correlations. Only when the standard estimator’s assumption of zero correlation happens to hold, as in the middle graph, does the standard estimator have about the same RMSE as our approach.

Since we cannot know *ex ante* what the correlation is between sibship size and mortality, our estimator, which does not require an assumption about this relationship, is clearly a better choice for real applications than the standard approach. In statistical language, we say that the new estimator “dominates” the standard approach.

**Sources of Bias Reduction** Now that we have established the advantages of our estimator in simulated data, we briefly show how much bias is reduced by each of the two

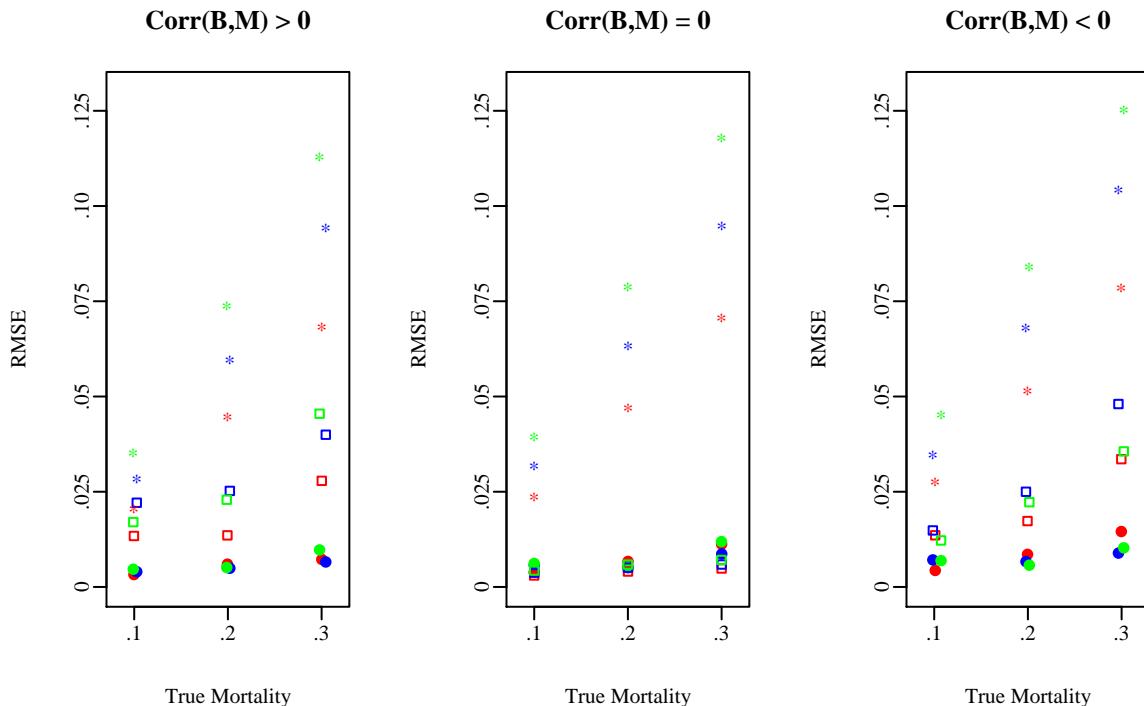


Figure 2: Root Mean Square Error (RMSE) in Mortality Estimates. Closed circles represent our estimator (with small RMSE, near the bottom of the graphs), open squares represent RMSE in the standard estimator, and asterisks are for the naive estimator. Simulations were drawn with a correlation between mortality and fertility that is positive (left graph), zero (middle), and negative (right). The figure shows simulations drawn with three levels of mortality (0.1, 0.2, and 0.3, on the horizontal axis) and **low** (green), **medium** (blue), and **high** (red) fertility.

corrections it includes. Thus, Table 1 lists the absolute bias of the naive estimator for each of our 27 data sets, each randomly drawn from a different set of starting parameters. The penultimate column of the table gives the percent reduction in bias from the naive estimator via weighting without correcting for families with zero survivors. Clearly weighting eliminates most of the bias — with only 27.5% of the bias left on average, and never more than half of the bias left. This weighting-only correction produces the largest reduction in bias in settings where the percent of families with zero survivors is lowest — in populations with high fertility and low mortality — and where there is a positive correlation between sibship size and mortality. Although most of the bias is eliminated via weighting alone in all simulated data sets, correcting also for families with zero survivors adds significantly to the bias reduction. The final column of the table demonstrates this by giving our full estimator that both weights and corrects for families with zero survivors. In this column,

| Correlation | Fertility | Mortality | Naive    | Bias as % of Naive  |           |
|-------------|-----------|-----------|----------|---------------------|-----------|
|             |           |           | abs bias | $\hat{q}_{\zeta=0}$ | $\hat{q}$ |
| Negative    | High      | Low       | 0.03     | 25.9%               | 9.5%      |
| Negative    | High      | Medium    | 0.05     | 27.4                | 13.9      |
| Negative    | High      | High      | 0.08     | 31.1                | 17.4      |
| Negative    | Medium    | Low       | 0.03     | 29.2                | 16.5      |
| Negative    | Medium    | Medium    | 0.07     | 33.4                | 6.0       |
| Negative    | Medium    | High      | 0.10     | 41.5                | 7.0       |
| Negative    | Low       | Low       | 0.05     | 35.7                | 12.5      |
| Negative    | Low       | Medium    | 0.08     | 40.3                | 3.3       |
| Negative    | Low       | High      | 0.13     | 46.3                | 7.1       |
| Zero        | High      | Low       | 0.02     | 19.8                | 8.3       |
| Zero        | High      | Medium    | 0.05     | 21.2                | 11.3      |
| Zero        | High      | High      | 0.07     | 23.5                | 14.7      |
| Zero        | Medium    | Low       | 0.03     | 23.6                | 13.5      |
| Zero        | Medium    | Medium    | 0.06     | 27.7                | 2.6       |
| Zero        | Medium    | High      | 0.09     | 32.1                | 7.7       |
| Zero        | Low       | Low       | 0.04     | 30.7                | 11.9      |
| Zero        | Low       | Medium    | 0.08     | 35.7                | 0.2       |
| Zero        | Low       | High      | 0.12     | 40.1                | 9.5       |
| Positive    | High      | Low       | 0.02     | 10.8                | 4.6       |
| Positive    | High      | Medium    | 0.04     | 15.9                | 8.9       |
| Positive    | High      | High      | 0.07     | 12.8                | 8.0       |
| Positive    | Medium    | Low       | 0.03     | 14.7                | 5.9       |
| Positive    | Medium    | Medium    | 0.06     | 19.4                | 1.7       |
| Positive    | Medium    | High      | 0.09     | 21.5                | 4.2       |
| Positive    | Low       | Low       | 0.04     | 23.7                | 6.1       |
| Positive    | Low       | Medium    | 0.07     | 28.8                | 0.1       |
| Positive    | Low       | High      | 0.11     | 28.9                | 7.2       |

Table 1: Percent Bias Reduction Relative to Absolute Bias in the Naive Estimator: Our Estimator without correcting for families with zero survivors  $\hat{q}_{\zeta=0}$  (i.e., with weighting only), and with full corrections,  $\hat{q}$ .

the bias is now reduced on average to only about 8 percent of the naive estimator’s bias.

## 4 Empirical Evidence

Unfortunately, publicly available data do not exist to make extensive validation tests. It would be ideal to be able to compare estimates based on survey data to a gold standard such as mortality calculated from a reliable vital registration system, but these data are not available. The DHS, for example, conducts its surveys in countries with incomplete or nonexistent vital registration systems. In this section, we therefore focus on two empirical issues that are nevertheless crucial.

First, we estimate in real data the correlation between sibship size and mortality. If

|              |      |      |               |      |       |
|--------------|------|------|---------------|------|-------|
| Peru         | 2000 | 0.97 | Guinea        | 1999 | 0.80  |
| Indonesia    | 1997 | 0.96 | Zimbabwe      | 1994 | 0.76  |
| Burkina Faso | 1998 | 0.95 | Nepal         | 1996 | 0.75  |
| Benin        | 1996 | 0.95 | Cameroon      | 1998 | 0.75  |
| Peru         | 1996 | 0.95 | Cote D'Ivoire | 1994 | 0.75  |
| Nigeria      | 1999 | 0.93 | Togo          | 1998 | 0.74  |
| Philippines  | 1998 | 0.93 | Eritrea       | 1995 | 0.70  |
| Chad         | 1997 | 0.93 | Ethiopia      | 2000 | 0.71  |
| Brazil       | 1996 | 0.92 | Zimbabwe      | 1999 | 0.69  |
| Indonesia    | 1994 | 0.91 | Colombia      | 1995 | 0.52  |
| Senegal      | 1999 | 0.90 | Zambia        | 1996 | 0.47  |
| Philippines  | 1993 | 0.88 | Uganda        | 1995 | -0.06 |
| Mali         | 1996 | 0.86 | Madagascar    | 1997 | -0.19 |
| Tanzania     | 1996 | 0.82 |               |      |       |

Table 2: Correlations between sibship size and mortality in 27 country-years.

this correlation is always near zero, then the method used in the literature would pose little risk of bias. To estimate this correlation, we apply our estimator of mortality separately to survey respondents with two siblings, three siblings, etc., so that estimating  $\zeta$  is not necessary. Then we simply compute the zero-order correlation between mortality  $\hat{q}$  and sibship size,  $B$ . We apply this procedure in 27 separate DHS surveys covering 24 countries. The countries, the year in which the survey was conducted, and the estimated correlation appear in Table 2.

Table 2 demonstrates unambiguously that mortality is not empirically independent of sibship size, as the standard estimator assumes. In the vast majority of surveys, the correlation is very high, often above 0.90. In two surveys, the correlation is negative. Any deviation from a zero correlation invalidates the standard estimator, but this table does not even suggest a tendency toward a zero correlation.

Finally, we offer empirical evidence that our procedure for estimating  $\zeta$  (the number of deaths in families with zero survivors) works well. The estimation weights in  $\hat{q}$  eliminate all bias from information obtained from families with one or more survivors, and so any bias that remains is solely a function of bias in estimating  $\zeta$ , making it the crucial remaining source of uncertainty in estimating deaths by survey.

Figure 3 demonstrates that the quadratic model we use to estimate  $\zeta$  fits the observed data in 27 different surveys from 24 different countries well. This provides considerable confidence, although not proof, that the only unobserved point would fit well too and so  $\hat{\zeta}$  is probably accurate. We also go another step and withhold one observed data point at

a time and see how accurately we can predict it with the remainder of the observed data points; the highly accurate fit of the data in Figure 3 is of course good indication that this exercise (which we do not show here) also reveals high quality predictions. Finally, although one should not make too much of a close fit of a model with three parameters to seven data points, the fact that the estimated relationship between sibship size and mortality is highly similar across this long list of diverse countries estimated independently is additional strong evidence that we have found a persistent, stable pattern that may be useful in extrapolating to deaths in families with zero survivors. (Indeed, although we do not pursue it here, a hierarchical model that shrinks these patterns toward a common mean or geographic neighbors might improve these estimates further.)

## 5 Generalizations

### 5.1 Maternal Mortality, Orphan Studies, and Other Family Data

In Section 2, we assume that respondents are asked about the family group in which they are a member. However, the same sample typically includes respondents with information about relatives not in their group. For example, if we are interested in the male mortality rate, the procedure above requires asking males about their male siblings. However, asking females about their male siblings would provide additional valuable information about the same quantity. We now develop a method to improve our estimator by using this readily available additional information. Without loss of generality, we continue with this example and assume the quantity of interest is the male mortality rate. We also use all notation above to refer to information about males from either male or female respondents, and add other notation when necessary.

In this alternative data collection scheme, each male respondent reporting about his male siblings in the time 2 sample still counts for  $D_i/S_i$  deaths in the time 1 pseudo-sample. The information reported by our female respondents must be treated somewhat differently. For one, females without male siblings convey no information about the male mortality rate and thus should be dropped. We might think about using the information provided by the remaining female respondents in the same way as we do with male respondents. However, this would assume that the mortality rate of the male siblings of females who survive to time 2 and appear in our sample is the same as the mortality rate of male siblings

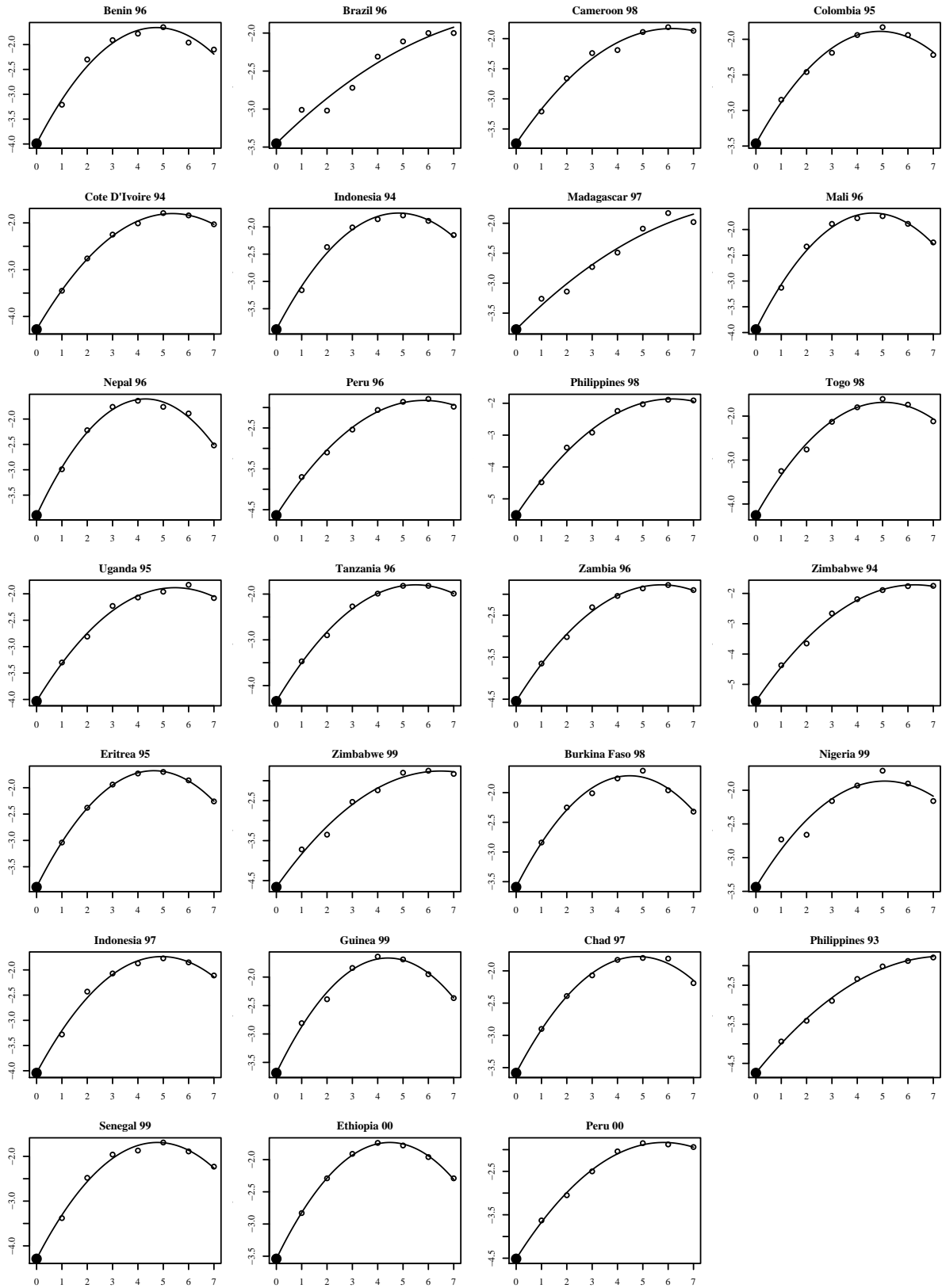


Figure 3: Quadratic Models Fit to Sibship Size (horizontally) by Logged Proportion of Deaths (vertically). The filled circle at zero survivors is projected. Note how well the quadratic fits the observed points (represented by open circles) and follows similar patterns across countries. The country and last two digits of the year of the survey are in the title of each graph.

of males who also appear in our sample. If instead, for example, males in families with many female siblings have lower mortality rates than males in families with fewer female siblings, then using the additional information provided by randomly selected surviving females would introduce a new form of selection bias.

Fortunately, we can use a weighting mechanism analogous to that in our original estimator to avoid this selection bias. The procedure is, first, to weight the sample of females in the same proportion as the males they represent, and then to apply the original weight to transform the result to the time 1 pseudo-sample. To do this, we need to define “Deaths” and “Survivors” in the first expression in Equation 5.

We first fix the number of “Survivors” to be the number of male respondents,  $n_m$ , plus the number of female respondents  $n_f$  weighted to represent an equivalent sample of males. To do the latter, we first compute the number of males in the pseudo-sample represented by each time 2 sampled female respondent, which is the ratio of the survival rate for males to the survival rate for females:  $R_i = (S_i/B_i)/(S_i^f/B_i^f)$ , where  $S_i^f$  and  $B_i^f$  are the numbers of female siblings surviving and born into the cohort, respectively, as reported by female respondents. In other words, if male survival is twice that of females then each female interviewed informs us about two males and so needs to be weighted with  $R = 2$ . If male and female survival rates are the same, then  $R = 1$  and no weighting is needed. Then we define

$$\text{Survivors} = n_m + \sum_{i=1}^{n_f} R_i. \quad (7)$$

Finally, for each survivor, we determine the number of deaths in the time 1 pseudo-sample. This is  $D_i/S_i$  for each male respondent and  $R_i(D_i/S_i)$  for each female respondent, as well as  $\zeta$  which is not represented in the sample. Thus, we write

$$\text{Deaths} = \sum_{i=1}^{n_m} \frac{D_i}{S_i} + \sum_{i'=1}^{n_f} R_{i'} \frac{D_{i'}}{S_{i'}} + \hat{\zeta} \quad (8)$$

where we use  $i$  ( $i = 1, \dots, n_m$ ) for male respondents and  $i'$  ( $i' = 1, \dots, n_f$ ) for female respondents.

To improve estimates of male mortality, we can ask females about their brothers as in the example here, but we can also gather information about male mortality from parents, neighbors, teachers, coworkers, etc. The same approach can be used to estimate maternal mortality or the mortality of parents from data on (adult) children or sisters, etc.

## 5.2 From Cohort to Period Estimation

For simplicity, we have until now presented our approach using quantities of interest defined for a cohort of people with fixed attributes, such as 20–40 year-old women, defined over the same interval of time, say 1980-2000. However, unless we draw a special sample that includes only 20-year-old women all born on the same day, our surveys will include many women who are at least 20 and no older than 40 for only part of the twenty years from 1980 to 2000. The cohort approach would not make use of this information. This problem is typically addressed in demographic studies by counting person-years, and measuring period rather than cohort quantities (Preston, Heuveline and Guillot, 2001). If a woman turns 20 in 1990, then we could include this person in the sample by simply counting her as contributing “half of a respondent” to the sample since she is at risk of dying in the right age range for only 10 years whereas those 20 years old in 1990 would be at risk for 20 years. Any portion of a person’s life for which they are not at risk of dying when aged 20–40 and between time 1 and 2 are thus removed from the sample and not counted, but those who appear in the sample for part of the period would contribute to our estimates. Following person-years in this way means that we are estimating the mortality of the period defined by the interval from time 1 to time 2 and are not following a specific cohort over this interval.

To formalize this idea, we first define  $B_j^*$  as the person-years lived and for simplicity define the period of interest as one “year” (or we could equivalently refer to person years lived as “person-periods lived”). The new quantity of interest, denominated in person-years, is thus

$$q^* = \frac{\sum_{j=1}^N M_j^*}{\sum_{j=1}^N B_j^*} \quad (9)$$

where  $M_j^*$  is the number of deaths in the family of respondent  $j$  (including  $j$ ) in the group of interest, divided by the number of people who contribute any positive number of person years to the analysis in family  $j$  in the group of interest.

Our conditional estimator is then

$$\hat{q}^* = \frac{\sum_{i=1}^n M_i W_i + \hat{\zeta}}{\sum_{i=1}^n W_i^* + \hat{\zeta}^*} \quad (10)$$

where, in the denominator,  $\zeta^*$  is the total number of person-years lived in families with zero survivors and  $W_j^* = B_j^*/S_j$  denotes an alternative weight. The numerator of the

conditional estimator is thus identical to that in (10), whereas the denominator is now the total number of person years. One additional quantity,  $\zeta^*$ , needs to be extrapolated from the sample, which we do in the same manner as for  $\zeta$ , except that the variable being predicted by  $s$  is the total number of person-years in families with  $s$  survivors.

## 6 Concluding Remarks

The approach developed here has two required features. The first is a weight variable that can be constructed from the variables already being collected and without any additional auxiliary information. This is an unusual situation since, although weighting functions are used often in statistics, the information necessary to construct the weights normally comes from external information, such as sampling strata whereas here the weight is constructed directly from the variables of interest. The second required feature is an estimate of the number of people who have died in families with zero survivors. This is information not represented directly in the sample at all, and is available either by assumption (as in the standard estimator) or by extrapolation from observed patterns in the data. Although the weighting is an exact correction, the extrapolation is by its nature more risky, although it is the only portion of the estimator that contains uncertainty beyond that typically involved in sample surveys.

For applied work, researchers should easily be able to substitute the method introduced here for the standard method and its variants presently used in the literature. Unless a researcher happened to be certain that sibship sizes in a particular dataset was unrelated to mortality, the new approach would generally be preferable.

Research should now turn to improving the quality of data on survival of relatives collected through surveys. The DHS surveys became successful in measuring child mortality by steadily improving both survey question instrumentation and interviewer training. For the mortality experience of adult relatives, both areas could use similar sustained attention from the research community. Data quality might be improved by techniques used in other areas, such as adding prompting questions, like the time of last contact with and physical distance to relatives, and using memorable events such as wars, famines, etc. to improve temporal recall. To validate these approaches, these types of survey data also need to be collected in countries with valid vital registration systems or in areas with established

demographic surveillance.

## References

- Bicego, G. 1997. "Estimating adult mortality rates in the context of the AIDS epidemic in sub-Saharan Africa: analysis of DHS sibling histories." *Health Transition Review* 7(S2):7–22.
- Brass, William and Kenneth Hill. 1973. Estimating Adult Mortality in Africa from Orphanhood. In *Proceedings of the International Population Conference Liege*. International Union for the Scientific Study of Population.
- Chiphangwi, J.D., T.P. Zamaere, W Graham, B. Duncan, T. Kenyon and R. Chinyama. 1992. "Maternal mortality in the Thyolo district of southern Malawi." *East African Medical Journal* 69:675–679.
- Danel, I., W. Graham, P Stupp and P. Castillo. 1996. "Applying the sisterhood method for estimating maternal mortality to a health facility-based sample: a comparison with results from a household-based sample." *International Journal of Epidemiology* 25:1017–1–22.
- Feeney, G. 2001. "The Impact of HIV/AIDS on Adult Mortality in Zimbabwe." *Population and Development Review* 27(4):771–980.
- Gakidou, Emmanuela, Margaret Hogan and Alan D Lopez. 2004. "Adult Mortality: Time for a Reappraisal." *International Journal of Epidemiology* 33(4):710–717.
- Garenne, M. and F. Friedberg. 1997. "Accuracy of indirect estimates of maternal mortality: a simulation model." *Studies in Family Planning* 28:132–142.
- Ghobarah, Hazem, Paul Huth and Bruce Russett. 2003. "Civil Wars Kill and Maim People—Long after the Shooting Stops." *American Political Science Review* 97(2, May):189–202.
- Graham, W., W. Brass and R.W. Snow. 1989. "Estimating Maternal Mortality: The Sisterhood Methods." *Studies in Family Planning* 20(125):125–135.
- Hill, Kenneth and J Trussell. 1977. "Further Developments in Indirect Mortality Estimation." *Population Studies* 31:313–334.
- King, Gary and Langche Zeng. 2004. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." <http://gking.harvard.edu/files/counterf.pdf>.
- King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical

- Analyses: Improving Interpretation and Presentation.” *American Journal of Political Science* 44(2, April):341–355. <http://gking.harvard.edu/files/abs/making-abs.shtml>.
- Malaker, CR. 1986. “Estimation of Adult Mortality in India: 1971–1981.” *Demography India* 15:126–136.
- Murray, Christopher J.L., Gary King, Alan D. Lopez, Niels Tomijima and Etienne Krug. 2002. “Armed Conflict as a Public Health Problem.” *BMJ (British Medical Journal)* 324(February 9):346–349. <http://gking.harvard.edu/files/abs/armedph-abs.shtml>.
- Preston, Samuel H., Patrick Heuveline and Michel Guillot. 2001. *Demography: Measuring and Modeling Population Processes*. Oxford, England: Blackwell.
- Shahidullah, M. 1995. “The sisterhood method of estimating maternal mortality: the Matlab experience.” *Studies in Family Planning* 26:101–106.
- Shiferaw, T. and F. Tessema. 1993. “Maternal mortality in rural communities of Illubabor, Southwestern Ethiopia: as estimated by the ‘sisterhood method’.” *Ethiopian Medical Journal* 31:239–249.
- Singh, R. 2000. “Estimation of Adult Mortality from Widowhood Data for India and its Major States.” Mumbai, India: International Institute for Population Sciences.
- Stanton, C., A. Nouredine and K. Hill. 2000. “An Assessment of DHS Maternal Mortality Indicators.” *Studies in Family Planning* 31:111–123.
- Timaeus, Iain. 1991. “Measurement of Adult Mortality in Developing Countries: A Comparative Review.” *Population Index* 57(4):552–568.
- Timaeus, Ian. 1986. “An Assessment of Methods for Estimating Adult Mortality from Two Sets of Data on Maternal Orphanhood.” *Demography* 23:435–450.
- Timaeus, Ian. 1991b. “Estimation of Adult Mortality from Orphanhood Before and Since Marriage.” *Population Studies* 45:455–472.
- Timaeus, Ian M. and Mohammed Ali. 2001. Estimation of Adult Mortality from Data on Adult Siblings. In *Brass Tacks: Essays in Medical Demography*, ed. B. Zaba and J. Blacker. Athlone pp. 43–66.
- Trussell, J. and G. Rodriguez. 1990. “A Note on the Sisterhood Estimator of Maternal Mortality.” *Studies in Family Planning* 21(6, Nov-Dec):344–346.
- Walraven, G.E.L. and P.W.J. van Dongen. 1994. “Assessment of maternal mortality in Tanzania.” *British Journal of Obstetrics and Gynaecology* 101:414–417.

- Wirawan, D.N. and M. Linnan. 1994. "The Bali indirect maternal mortality study." *Studies in Family Planning* 5:304-309.
- Zaba, Basia and Patricia H. David. 1996. "Fertility and the Distribution of Child Mortality Risk Among Women: An Illustrative Analysis." *Population Studies* 50:263-278.