DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY
Office for Scholarly Communication

Death by Survey: Estimating Adult Mortality without Selection Bias from Sibling Survival Data

Citation

Gakidou, Emmanuela, and Gary King. 2006. Death by survey: Estimating adult mortality without selection bias from sibling survival data. Demography 43(3): 569-585.

Published Version

doi:10.1353/dem.2006.0024

Permanent link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:4214912

Terms of Use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

DEATH BY SURVEY: ESTIMATING ADULT MORTALITY WITHOUT SELECTION BIAS FROM SIBLING SURVIVAL DATA*

EMMANUELA GAKIDOU AND GARY KING

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, as 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.

Demographers, social scientists, actuaries, public policy makers, and medical and public health researchers need accurate mortality data for many practical and research purposes. 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. These problems have generated 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 from household residents (Feeney 2001; Graham, Brass, and Snow 1989), siblings (Bicego 1997; Chiphangwi et al. 1992; Danel et al. 1996; Gakidou, Hogan, and Lopez 2004; Garenne and Friedberg 1997; Graham et al. 1989; Shahidullah 1995; Shiferaw and Tessema 1993; Walraven and van Dongen 1994; Wirawan and Linnan 1994; Zaba, Timœus, and Ali 2001), parents (Brass and Hill 1973; Hill and Trussell 1977; Timœus 1986, 1991a; Timœus and Jasseh 2004), and spouses (Malaker 1986; Singh 2000; Stanton, Noureddine, and Hill 2000; Timœus 1991b).

The Demographic and Health Surveys (DHS) program has invested heavily in collecting complete birth histories of nationally representative samples of women. This program has produced 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, through modules on sibling survival. However accurate these surveys may be, 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 a single survey of those who are alive only at the end of a period of interest makes the method inexpensive and thus feasible, but it leads to a serious

*Emmanuela Gakidou, Research Associate, Institute for Quantitative Social Science, Harvard University. Gary King, David Florence Professor of Government, Department of Government, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138; E-mail: King@Harvard.Edu. Data and software to replicate the results presented in this article are available at http://GKing.Harvard.edu. We thank Andrew Gelman, Michel Gillot, Ken Hill, Ian Timœus, 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.

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.

We introduce our method by way of a running example using data on sibling survival to estimate mortality. The same approach can be used to estimate mortality rates from other relatives and, in specialized cases, from acquaintances, friends, and neighbors. The method also applies to estimating rates of emigration and immigration (Zaba 1986) and other quantities of interest.

PROBABILITY OF DEATH IN A COHORT

Preliminary Notation

Let j ($j = 1, \ldots, N$) 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 in which 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. In the section Generalizations later in this article, we turn from cohort quantities to more practically useful period quantities.

We are interested in drawing a single sample of survivors at Time 2 to infer the mortality rate or other quantities from the full (unobserved) sample of interest at Time 1. That is, all the methods discussed in this article require only a single cross-sectional survey taken at the end of the period of interest, which we refer to as Time 2. For convenience, we assume 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 as 0 if j survives. The quantity of interest can be expressed in three equivalent ways:

$$q = \frac{\sum_{j=1}^{N} d_j}{N} \tag{1}$$

$$= \frac{\sum_{f=1}^{F} D_f}{\sum_{f=1}^{F} B_f} \tag{2}$$

$$= \frac{\sum_{j=1}^{N} M_j}{N}. \tag{3}$$

Eq. (1) 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 ($f = 1, \ldots, F$), where B_f and D_f are the numbers of births and deaths among the siblings of family f, respectively, and $\Sigma_{f=1}^{F} B_f = N$. 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. All three estimators are biased when applied to a sample drawn at Time 2; our goal is to develop an estimator without these biases.

To be clear about the notation, each individual who is 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., $M_j = M_{j'} = M_f = D_f / B_f$ for all j and j' that are members of the same sibship f). Thus, if we could sample at Time 1, each draw of an individual would be equivalent to a draw of a family selected with probability proportional to B_j. For example, a family with five siblings is represented in the population with five times the frequency, and thus has 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 refers to 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 ($i = 1, \ldots, n$) for respondents who 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:

$$\tilde{q} = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n} B_i}. \tag{4}$$

This estimator has an obvious and massive selection bias problem because 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. 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 because it is zero for all survey respondents):

$$\check{q} = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n} B_i - n}. \tag{5}$$

Trussell and Rodríguez (1990) pointed out three sources of bias in this method: the respondent (who is, of course, always alive) is not counted, which 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, causing mortality to be underestimated; and families with no survivors are not represented in the sample, which will bias mortality estimates downward. Trussell and Rodríguez then proved 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 in which 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 is positively correlated with sibship size—for example, if people in high-mortality areas have more children than those in low-mortality areas, or if children in large families have fewer resources and thus higher mortality than those in smaller families. Bias would also result in the reverse situation in which mortality is negatively correlated with sibship size. Any correlation between fertility and mortality, no matter the reason, will generate bias.

Although this unbiasedness condition likely applies only to real data in rare instances (Zaba and David 1996), Trussell and Rodríguez's mathematical result demonstrating unbiasedness when this condition holds is nonetheless vitally important. It demonstrates that there exist conditions under which it is possible to infer mortality in a population from a sample selected in a biased but convenient way. And what is more important, 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 only 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 S_i and B_i 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 Eq. (3) with the weighted average of M_i, using weight $W_i = B_i / S_i$. That is, the first part of our estimator, which applies to families with at least some survivors, is

$$\hat{q}_0 = \frac{\sum_{i=1}^{n} M_i W_i}{\sum_{i=1}^{n} W_i}. \tag{6}$$

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 in Eq. (6), means that the first problem with Time 2 sampling vanishes (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, 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 ζ and which needs to be added to both the numerator and denominator of the weighted average because for this group, $B_i = D_i$. With an estimator for ζ, which we denote as $\hat{\zeta}$, our estimator of the mortality rate will be

$$\hat{q} = \frac{\sum_{i=1}^{n} M_i W_i + \hat{\zeta}}{\sum_{i=1}^{n} W_i + \hat{\zeta}}. \tag{7}$$

Before discussing how to estimate ζ, we offer an alternative interpretation of Eq. (7) that is useful for intuition and for later generalizations. Thus far we have distinguished between two overlapping groups: the original population that is followed between Times 1 and 2 and the respondents who are drawn randomly at Time 2 from those who have survived. If we could apply one of the simple expressions in Eqs. (1)–(3) to the original population, we would recover the quantity of interest q because, of course, this is how we define q. If instead we had a random sample from this original population and could elicit information about each person's mortality during Times 1 and 2 and about his or her sibship, applying Eq. (1), $q = \Sigma_{j=1}^{N} d_j / N$, to the sample would yield an unbiased estimate of q.

Because 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 we apply the estimators in Eqs. (1)–(3). We do this by rewriting Eq. (7) as

$$\hat{q} = \frac{deaths}{deaths + survivors} \tag{8}$$

$$= \frac{\left[\sum_{i=1}^{n}\left(D_i / S_i\right) + \hat{\zeta}\right]}{\left[\sum_{i=1}^{n}\left(D_i / S_i\right) + \hat{\zeta}\right] + n}, \tag{9}$$

where *deaths* and *survivors* in Eq. (8) refer to the totals in the pseudo-sample. Note that this equation is the sample analogue to Eq. (1) applied to 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 constrains only the ratio in Eq. (8) 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. Eq. (9) 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, ζ. The Time 2 sample reveals the first component directly, and we need to estimate the second.

We now turn to estimating ζ, the final task of this section.[1] Although no certain or directly estimable information about ζ exists in a sample drawn at Time 2, 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 $s = 1, 2, \ldots$) 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 ζ.

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 $s = 1, \ldots, 7$. (We exclude death proportions from $s > 7$ because in our data sets 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 $Y_s \equiv \ln(\sum_{\{i:S=s\}} D_i / S_i)$ for $s = 1, \ldots, 7$ on a constant, s, and s^2, yielding

$$\hat{Y}_s = \hat{\alpha}_0 + \hat{\alpha}_1 s + \hat{\alpha}_2 s^2. \tag{10}$$

[1] 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. Establishing which survey respondents are from the same sibship is infeasible in most contexts and requires data that are not collected in any major national survey. We will therefore attempt to estimate ζ directly without this or any other decomposition.

We then transform the constant term, $\hat{\alpha}_0$ (which is the predicted value of the number of deaths in the pseudo-sample for $s = 0$), to obtain an estimate of ζ.[2]

Although we chose this simple quadratic model because it seemed to fit our real data well, nothing can guarantee that an extrapolation (i.e., an inference outside the range of observable data) will always be accurate (King and Zeng 2006). It is always possible that total deaths given s follow a completely different pattern for the unobserved point when $s = 0$ than for the observed points when $s > 0$. What makes us somewhat optimistic are experiments, discussed later in the section Empirical Evidence, with data from 24 countries in which we set aside data we observe in each country and try to predict them from the rest of the observed data in that country; these experiments work out well in a wide range of countries. For example, we fairly accurately predict the number of deaths from families with one survivor ($s = 1$) using only the data on deaths 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 $s = \{1, 3, 4, 5, 6, 7\}$ and so on.

We made several choices based on empirical evidence, and so these may need to be changed for other applications. For example, we chose a quadratic form because it was the most parsimonious model that fit our data well. We also dropped families with more than seven children in fitting the quadratic model because of data sparseness. Yet we know that the tails of quadratic formulas are sensitive to outliers and thus will not always be appropriate, and families with more than seven children may add important information in countries with higher mortality. Other possible approaches include using different functional forms following the strategy we adopted, using distributional assumptions (as in Zaba and David 1996), or collecting additional information outside the sample, such as reports from mothers or others on the number of deaths in sibships with zero survivors.

To emphasize the uncertain nature of extrapolation, we briefly discuss another approach, which is to regress the log of deaths in the observed sample, $\ln(\Sigma_{\{i:S = s\}}D_i)$, rather than the pseudo-sample, $\ln(S_{\{i:S = s\}}D_i \,/\, S_i)$, on a quadratic function of the number of survivors, s. This approach 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 than the approach we described above, 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 we usually stick with it, but which approach to use in any particular instance is, of course, an important substantive judgment.

Our ultimate estimator for the mortality rate from a cohort sample is then Eq. (7) with this estimate for ζ 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 ζ. Because the samples that are typically available are large, normally only the latter is a significant concern.

An advantage of our approach is that it isolates the piece of the problem that is not amenable to direct statistical estimation so that the extrapolation model cannot affect inferences about families with survivors. The same extrapolation issue in our estimator exists in both previously used 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 via a survey rather than a

[2] 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).

property of any one method. Standard errors or confidence intervals that are 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 for which ζ is likely to be the largest, such as groups with high mortality rates and low fertility rates.

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 all combinations of low (0.1), medium (0.2), and high (0.3) average mortality rates, with average fertility levels approximately representing Kenya (4.26 children), Turkey (3.07), and Kazakstan (2.56), 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

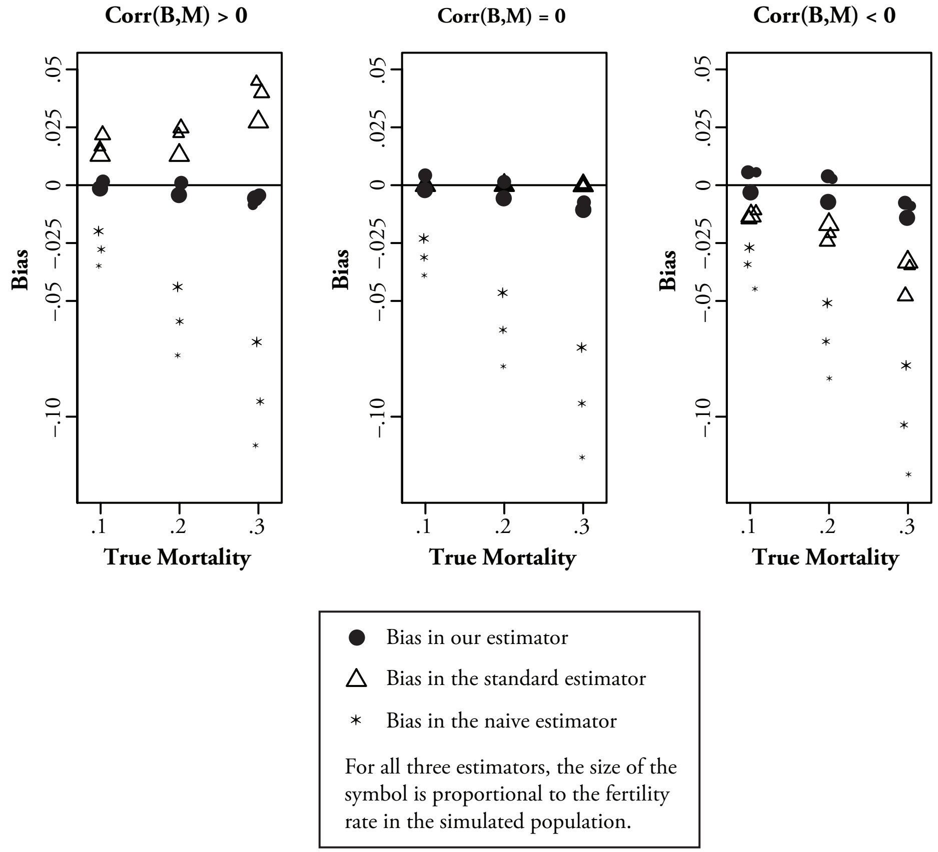
The degree of bias is defined for estimator $\hat{q}$ as $\text{Bias}(\hat{q}) = E(\hat{q} - q)$, where $E(\cdot)$ takes the average over repeated samples. We approximate this quantity and the bias for the other two estimators by subtracting the true mortality from the mortality estimated from each of the 1,000 data sets and then 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 is shown vertically, with a line drawn to denote zero bias. Higher fertility levels are represented with bigger symbols.

In all three graphs, we denote our estimator by a closed circle, all 27 of which are fairly near the zero-bias line. The deviation from zero bias is due only to estimating ζ, since it requires extrapolation. The other portion of the estimator is an exact correction (so that if we knew the true ζ 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 underestimates mortality, no matter what the correlation is. Finally, we plot the bias in the standard estimator widely used in the literature with a triangle. When sibship size is positively correlated with mortality, the standard estimator markedly overestimates mortality (note the positive bias portrayed above the line for all triangles 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 standard estimator is unbiased.

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 that is 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 smaller asterisks appear lower on the graph, indicating larger absolute bias) because 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.

Figure 1. Bias in Mortality Estimates: Simulations Drawn With Three Levels of Mortality (0.1, 0.2, and 0.3) and Three Different Relationships Between Mortality and Fertility (positive correlation, zero correlation, and negative correlation)



For all three estimators, the size of the symbol is proportional to the fertility rate in the simulated population.

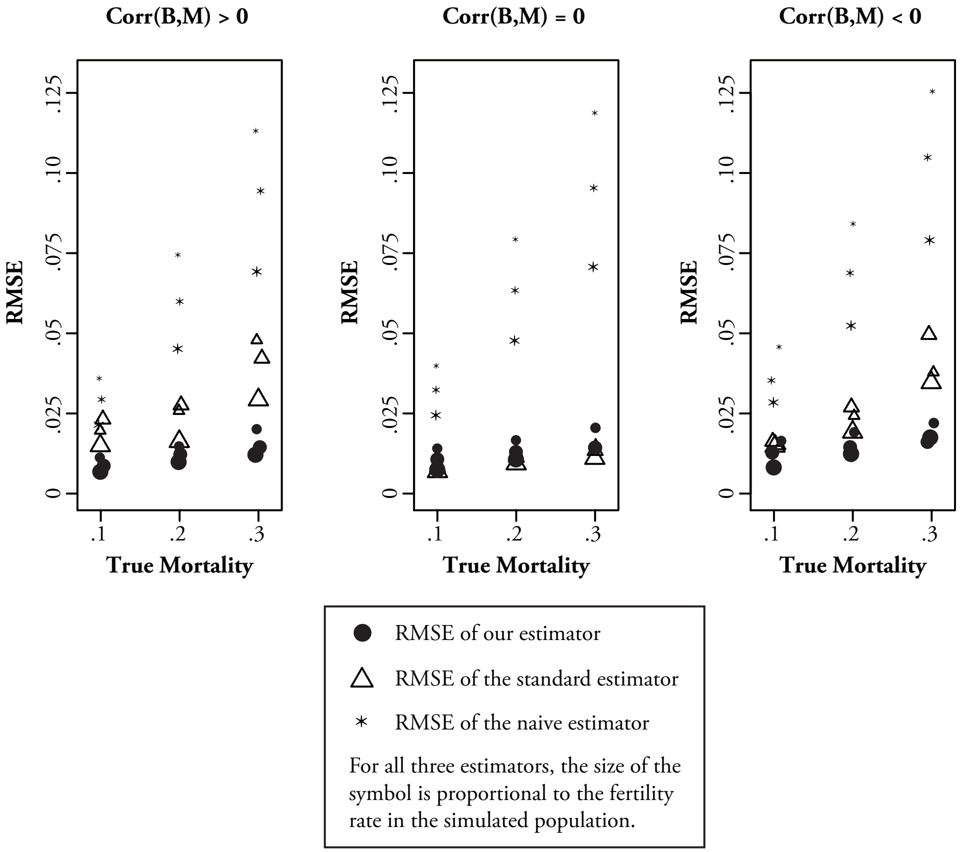
Mean Square Error

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

$$\begin{aligned}\mathrm{MSE}(\hat{q}) &= E\left[(\hat{q}-q)^2\right]\\ &= V(\hat{q}) + Bias(\hat{q})^2. \end{aligned}\tag{11}$$

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

Figure 2. Root Mean Square Error (RMSE) in Mortality Estimates: Simulations Drawn With Three Levels of Mortality (0.1, 0.2, and 0.3) and Three Different Relationships Between Mortality and Fertility (positive correlation, zero correlation, and negative correlation)



The results are presented in Figure 2, which is organized as Figure 1 is in terms of graphs, sizes, and symbols to represent correlation, average fertility, and estimators, respectively. In these graphs, RMSE is on the vertical axis, where symbols higher on the graph indicate larger values of RMSE and an inferior estimator. The results here parallel those 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.

Because 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.

Table 1. Percentage Bias Remaining Relative to the Absolute Bias in the Naive Estimator: Our Estimator Without Correcting for Families With No Survivors (i.e., with weighting only), $\hat{q}_0$, and With Full Corrections, $\hat{q}$

			Naive Absolute	Bias Remaining as a Percentage of the Naive Estimator	
Correlation	Fertility	Mortality	Bias	$\hat{q}_0$	$\hat{q}$
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
Average			0.06	27.5	8.1

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 corrections it includes. 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 shows the percentage 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

Table 2. Correlations Between Sibship Size and Mortality in 27 Country-Years

Country	Year	Coefficient	Country	Year	Coefficient
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			

weighting-only correction produces the largest reduction in bias in settings where the percentage of families with zero survivors is the lowest—that is, 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 by 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 showing our full estimator, which both weights and corrects for families with zero survivors. As shown in this column, the bias is now reduced, on average, to only about 8% of the naive estimator's bias.

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, are conducted 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 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, and so on, so that estimating ζ 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 .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 ζ (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 ζ, making it the crucial remaining source of uncertainty in estimating deaths by survey.

Figure 3 demonstrates that the quadratic model we use to estimate ζ fits the observed data in 27 different surveys from 24 different countries well. This finding provides considerable confidence, although not proof, that the only unobserved point would fit well too and thus that $\hat{\zeta}$ is probably accurate. We go another step and withhold one observed data point at a time to see how accurately we can predict the data point with the remainder of the observed data points; the highly accurate fit of the data in Figure 3 is, of course, a 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 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.)

GENERALIZATIONS

Maternal Mortality, Orphan Studies, and Other Family Data

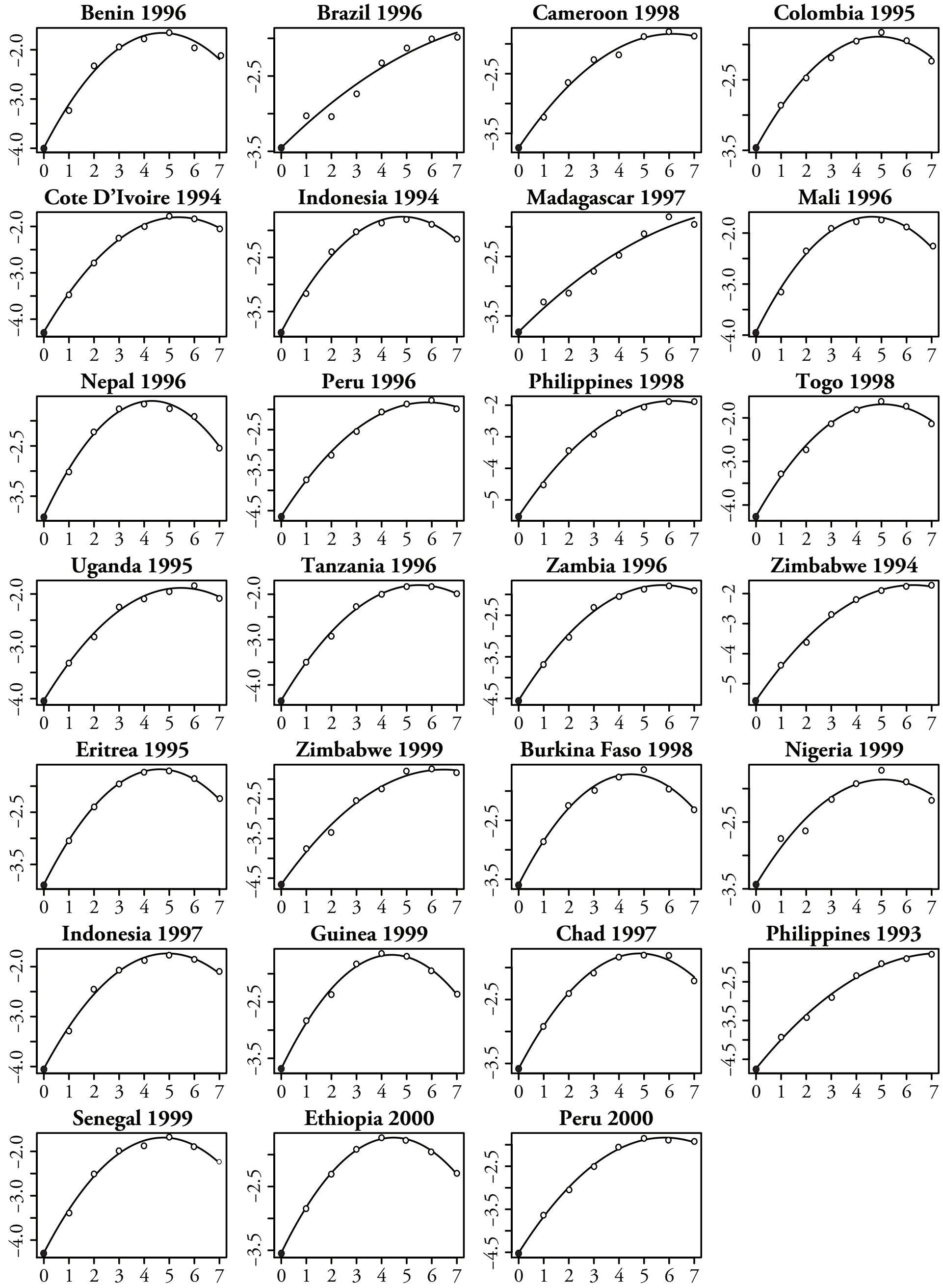
In the section Probability of Death in a Cohort, we assumed that respondents are asked about their siblings. However, the same sample typically includes respondents with information about relatives who are not in their group. For example, if we are interested in the male mortality rate, our procedure 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 use all the notation we used earlier 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. 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 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 that is analogous to that in our original estimator to avoid this selection bias. The procedure is 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 Eq. (8).

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

Figure 3. Quadratic Models Fit to Sibship Size (on *x*-axis), by Logged Proportion of Deaths (on *y*-axis), for 27 Country-Years



Notes: The filled circle at zero survivors is projected. The fit of the quadratic model to the observed points is represented by open circles.

do the latter, we first compute the number of males in the pseudo-sample represented by each female respondent sampled at Time 2, which is the ratio of the survival rate for males to the survival rate for females: $R_i = (S_i / B_i) \big/ (S_i^f / B_i^f)$, where S_i^f and B_i^f are the numbers of female siblings who survive and are born into the cohort, respectively, as reported by female respondents. In other words, if male survival is twice that of females, then each female who is interviewed informs us about two males and thus 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

$$survivors = n_m + \sum_{i=1}^{n_f} R_i. \tag{12}$$

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

$$deaths = \sum_{i=1}^{n_m} \frac{D_i}{S_i} + \sum_{i'=1}^{n_f} R_{i'} \frac{D_{i'}}{S_{i'}} + \hat{\zeta}, \tag{13}$$

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

To improve estimates of male mortality, we can ask females about their brothers, as in the example here, and we can gather information about male mortality from parents. A similar approach can be used to estimate maternal mortality or the mortality of parents from data on (adult) children or sisters (Timœus 1991a).

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- to 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 who were 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 20 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 turned 20 in 1990, we could include her in the sample by simply counting her as contributing "half of a respondent" to the sample because she was at risk of dying in the right age range for only 10 years, whereas those who were 20 years old in 1980 were at risk for 20 years. Any portion of a person's life for which he or she is not at risk of dying at age 20–40 and between Time 1 and 2 is 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

$$q^* = \frac{\sum_{j=1}^{N} M_j^*}{\sum_{j=1}^{N} B_j^*}, \tag{14}$$

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

$$\hat{q}^* = \frac{\sum_{i=1}^{n} M_i W_i + \hat{\zeta}}{\sum_{i=1}^{n} W_i^* + \hat{\zeta}^*}, \tag{15}$$

where in the denominator, ζ^* 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 Eq. (7), whereas the denominator is now the total number of person-years. One additional quantity, ζ^*, needs to be extrapolated from the sample, which we do in the same manner as for ζ, except that the variable being predicted by *s* is the total number of person-years in families with *s* survivors.

CONCLUDING REMARKS

The approach to estimating adult mortality from data on sibling survival described herein requires only a single representative cross-sectional survey with questions about the respondents' sibship size and survival histories. As with any inferential task, a larger data set is always better, or at least not worse, than a smaller one. For mortality estimation, larger sample sizes are especially useful when applied to cohorts or countries with low mortality rates, but no special features of the methods employed require data sets that are any larger than usual.

The method of estimation has two required features. The first is a weight variable that can be constructed from the variables that are already being collected and without any additional information. This is an unusual situation because although weighting functions are used often in statistics, weights are normally constructed using external information, such as sampling strata; with our approach 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 and is available either by assumption (as in the standard approach) or, as we suggest, by extrapolation from observed patterns in the data. Although the weighting is an exact correction, the extrapolation is by its nature more risky. It is, however, 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 approach introduced here for the standard method and its variants that are presently used in the literature. Unless a researcher happened to be certain that sibship sizes in a particular data set were unrelated to mortality, which seems unlikely in practice, the new approach would generally be preferable. Although the simulation presented here applies to a wide range of countries and types of mortality and fertility patterns, it will not apply to all; in those situations, it would be sufficiently easy to rerun our simulation with other parameter values.

The idea of using information on family members to estimate mortality rates of groups of which the respondent is a member can be extended to other, less well-defined groups based on, for example, surveys of neighbors, teachers, and coworkers. This seems like valuable information that would be easy to collect in most sample surveys, but then the equivalent of sibship size would not be known and thus would need to be estimated by other means, such as the numbers of friends and contacts or network centrality.

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, W. and K. Hill. 1973. “Estimating Adult Mortality in Africa From Orphanhood.” Pp. 111–23 in *Proceedings of the International Population Conference, Liege*. Liege: International Union for the Scientific Study of Population.

Chiphangwi, J.D., T.P. Zamaere, W. Graham, B. Duncan, T. Kenyon, and R. Chinayama. 1992. “Maternal Mortality in the Thyolo District of Southern Malawi.” *East African Medical Journal* 69:675–79.

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* 25:1017–22.

Feeney, G. 2001. “The Impact of HIV/AIDS on Adult Mortality in Zimbabwe.” *Population and Development Review* 27:771–980.

Gakidou, E., M. Hogan, and A.D. Lopez. 2004. “Adult Mortality: Time for a Reappraisal.” *International Journal of Epidemiology* 33:710–17.

Garenne, M. and F. Friedberg. 1997. “Accuracy of Indirect Estimates of Maternal Mortality: The Sisterhood Methods.” *Studies in Family Planning* 28:132–42.

Graham, W., W. Brass, and R.W. Snow. 1989. “Estimating Maternal Mortality: The Sisterhood Methods.” *Studies in Family Planning* 20:125–35.

Hill, K. and J. Trussell. 1977. “Further Developments in Indirect Mortality Estimation.” *Population Studies* 31:313–34.

King, G., M. Tomz, and J. Wittenberg. 2000. “Making the Most of Statistical Analyses: Improving Interpretation and Presentation.” *American Journal of Political Science* 44:341–55.

King, G. and L. Zeng. 2006. “The Dangers of Extreme Counterfactuals.” *Political Analysis* 14: 131–59.

Malaker, C.R. 1986. “Estimation of Adult Mortality in India: 1971–1981.” *Demography India* 15:126–36.

Preston, S.H., P. Heuveline, and M. 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. 1963. “Maternal Mortality in Rural Communities of Illubabor, Southwestern Ethiopia: As Estimated by the ‘Sisterhood Method.’” *Ethiopian Medical Journal* 31:239–49.

Singh, R. 2000. “Estimation of Adult Mortality From Widowhood Data for India and Its Major States.” Report. International Institute for Population Sciences, Mumbai, India.

Stanton, C., A. Noureddine, and K. Hill. 2000. “An Assessment of DHS Maternal Mortality Indicators.” *Studies in Family Planning* 31:111–23.

Timœus, I. 1986. “An Assessment of Methods for Estimating Adult Mortality From Two Sets of Data on Maternal Orphanhood.” *Demography* 23:435–50.

———. 1991a. “Estimation of Adult Mortality From Orphanhood Before and Since Marriage.” *Population Studies* 45:455–72.

———. 1991b. “Measurement of Adult Mortality in Developing Countries: A Comparative Review.” *Population Index* 57:552–68.

Timœus, I.M. and M. Jasseh. 2004. “Adult Mortality in Sub-Saharan Africa: Evidence From Demographic and Health Surveys.” *Demography* 41:757–72.

Trussell, J. and G. Rodríguez. 1990. “A Note on the Sisterhood Estimator of Maternal Mortality.” *Studies in Family Planning* 21:344–46.

Walraven, G.E.L. and P.W.J. van Dongen. 1994. “Assessment of Maternal Mortality in Tanzania.” *British Journal of Obstetrics and Gynaecology* 101:414–17.

Wirawan, D.N. and M. Linnan. 1994. “The Bali Indirect Maternal Mortality Study.” *Studies in Family Planning* 5:304–309.

Zaba, B. 1986. *Measurement of Emigration Using Indirect Techniques: Manual for the Collection and Analysis of Data on Residence of Relatives*. Belgium: Ordina Editions.

Zaba, B. and P.H. David. 1996. “Fertility and the Distribution of Child Mortality From Data on Adult Siblings.” Pp. 43–66 in *Brass Tacks: Essays in Medical Demography,* edited by B. Zaba and J. Blacker. London: Athlone.

Zaba, B., I. Timœus, and M. Ali. 2001. “Estimation of Adult Mortality From Data on Adult Siblings.” Pp. 43–66 in *Brass Tacks: Essays in Medical Demography,* edited by B. Zaba and J. Blacker. London: Athlone.