Imputing Birth Histories to Construct Direct Estimates of Child Mortality

Livia Montana, Kenneth Hill
Harvard School of Public Health

Abstract
We present a new methodology of imputing full birth histories from one sample population to another with only summary birth histories (SBH), by matching on unique combinations of exposure, total number of children ever born, and number of dead children. We control for exposure by matching on the age of mother, aggregated into 3-year groups. We randomly match each woman with a given SBH to multiple comparable women with the same SBH and a full birth history, and we repeat this *m* times. Once the matches are made, direct mortality estimates are derived from the imputed full birth histories. We make one assumption: controlling for exposure or duration of childbearing, the birth history of a woman given her CEB and CD does not change over time. We present results for five countries and show that direct estimates from imputed data match or are better than indirect estimates.

Introduction
Monitoring levels, trends and differentials in child mortality has become a high priority in the run-up to the Millennium Development Goals end year of 2015, and with large flows of development assistance into child health programs, the need for accurate child mortality estimates is ever greater. Most developing countries lack the complete civil registration data from which measures of child mortality are derived in developed countries, relying instead on reports from women about the survival of their children in censuses or household surveys.

There are two main approaches to estimating child mortality: direct and indirect methods. Direct methods are based on full birth histories (FBH) collecting dates of birth and ages at death for each child a woman has given birth to. Indirect methods are based on summary birth histories (SBH - numbers of children ever born and children dead). SBHs are widely used in censuses, where complete population coverage can support estimates for small geographic areas, and are sometimes used in sample surveys as well (for example, Unicef's MICS surveys), since they are much less expensive to collect than full birth histories. Methods of estimating child mortality from SBHs require assumptions, and may also be affected by selection bias.

Perhaps the most widely used indirect method for child mortality was developed by Brass (1975) using proportions of children dead by age group of mother, and has been extended by others including Sullivan (1972); Trussell (1975), Preston and Palloni (1978). The method relies on two pieces of information: the number of children ever borne, and the number of children surviving. Hill and Figueroa (2001) proposed using the time since first birth instead of age to reduce selection bias. Using this information along with assumptions of an underlying distribution of age specific fertility (or time since first birth), child mortality rates can be estimated. The indirect method does not, however, provide estimates of age patterns of mortality or mortality for specified time periods, distinct disadvantages of the method.

For developing countries, direct measures of child mortality are calculated from full birth histories which include dates of birth and death for each child born to every woman. These

histories have been used in most World Fertility Surveys, and in the Demographic and Health Surveys. The detailed data can be used to construct life table measures of mortality, and confidence intervals can also be estimated. Furthermore, full birth histories allow for individual-level data analysis, and these data have proven valuable in studies of household and individual determinants of mortality. The main drawbacks of full birth history data is the cost of data collection, and accuracy of dates (see for example Pullum 2006). Interviewers must be carefully trained to probe adequately to obtain accurate dates of birth and death. Answering the questions can be time consuming for women with many births. And in many settings, dates are not recorded or remembered, and it can be difficult to get respondents to respond precisely. Both summary and full birth history data collection may underreport deaths as it is a traumatic memory for women to recall, and omission may be likely. Nonetheless, direct estimates from full birth histories are generally considered the gold standard for mortality rates in absence of accurate and reliable vital registration data.

Multiple Imputation

The problem of missing data affects all types of data gathering in any discipline. Respondents may not answer questions in a survey, interviewers may mistakenly skip a question, data entry clerks may make mistakes when entering data, and data processing and editing procedures could intentionally or unintentionally cause the loss of information. Many methods have been developed to impute missing information. For large public use datasets, missing data is generally imputed by the agency sponsoring the data collection, and the imputed values are made available to researchers. For such datasets, it is important to impute values in a transparent and preferably simple fashion, in a consistent manner.

Multiple imputation, originally proposed by Rubin (1977, 1978), is a method which replaces each missing value with two or more acceptable values which represent a distribution of possibilities. Rubin (1987) describes the main strengths of multiple imputation: 1) it is suitable for large surveys analyzed by many users; 2) survey data analysis is highly specific, and missing data can have substantial consequences for survey estimations; appropriate adjustments for missing data may not be straightforward or easy to implement; 3) missing values do not generally occur at random, so the study of differences between respondents and nonrespondents can be of special value; and finally 4) various methods of imputation are already common in census and survey practice, so that adapting these single-imputation methods to multiple-imputation can be a simple modification. As Rubin (1987) describes, in real-life applications where missing data are a nuisance rather than a the primary focus of an analysis, “an easy, approximate solution with good properties can be preferable to one that is more efficient but problem-specific and complicated to implement.”

The method presented here relies on these general principles, but diverges in a few key ways. First, this method imputes a complete dataset rather than some missing values. There is no systematic nature of, selection effect, or bias in the missing values; they are missing for the entire sample selected at random. The method proposed here differs from traditional imputation in that it does not actually predict missing values, it uses real values (full histories) from actual women. The prediction process would be extremely cumbersome as many values would have to be estimated for the same individual woman.

Another divergence relates to how to treat the imputed values once they have been imputed *m* times. Rubin's method calls for averaging the imputed values across *m* datasets. When imputing values such as income, or body weight, averaging makes sense. But for dates of birth and death, averaging month or year of death across multiple imputations would skew the histories in unpredictable ways, distorting exposure times, length of pregnancies and birth intervals. In this case, the imputed estimates are appended and reweighted by 1/*m* in order to adjust for the repeat cases.

Assuming that completely missing data meets the criteria of missing at random, we can then try to determine how many times we will need to impute to get a reliable and efficient estimate. The rate of missing information, together with the number of imputations m, determines the relative efficiency of the MI inference. In datasets where only a small fraction of data is missing, Rubin and others have shown that only a few imputations are needed, usually only three to five. Rubin's formula for large sample coverage probabilities follows this formula:

$$\text{coverage proportion} = \left(1 + \frac{a}{m}\right)^{-1}$$

where *a* is equal to the fraction of data which is missing and *m* refers to the number of imputations (Rubin 1987). The following table demonstrates the coverage proportion across a range of values of *a*, the fraction of missing data, and *m* imputations.

Data

The data used in this analysis come from the Demographic and Health Surveys (DHS). The DHS data contain full birth histories for all eligible women in the survey between the ages of 15 and 49. For four test countries, we use pairs of DHS surveys, where the first survey contains the full birth history, and the second is treated as if it had only a summary birth history. We first used pairs of surveys five years apart from Kenya, Bolivia, Egypt and Indonesia (see Table 1). We then matched women from the 2000-01 Zambia DHS with women in the 1999 Zimbabwe DHS.

Method

The basic idea is to assume the fertility pattern of a woman in the first survey (time one) of the pair for a woman with the same characteristics (age or exposure, CEB, CD) in the second (time two) survey. We assume women in the time one survey provide the full birth histories, and women in the time two survey have only summary birth histories. Each woman in the time two survey is matched randomly, with replacement, to a woman with the same summary birth history in the time one survey. The match is repeated 500 times until each woman in the time one survey has 500 matches. Some women from the time one survey did not match on exposure, CEB and CD. These women were then matched to women from the same geographic region in a pool of full birth histories from all recent DHS surveys.

A simple example illustrates the method. In the first step, each woman in the time two survey with a summary birth history is given a type, consisting of the total number of children ever born, number died, and her years since first birth (grouped into 5 year categories). Emily from the 2004 DHS has 1 child, 0 dead, and less than 5 years have elapsed since her first birth. Emily's "type" is 01-00-01. In the second step, all of the women in the time 1 survey with type 01-00-01 are identified. One is chosen at random, Natalie, and she is matched with Emily. They

each have the same number of children, same number died, and same years since first birth. Natalie's distribution of the date of birth of her only child is then imputed or appended onto Emily's record, and it becomes Emily's full birth history. Natalie is back in the pool of possible matches. The same process is then followed for Emma; she is matched with Julia and assumes her full birth history. For Isabella and Ashley who have each had one birth and one child death, and 10-14 years have elapsed since their first (only) birth, they both get matched to Chloe at random. Chloe's birth history is then appended to both Isabella and Ashley.

2004 DHS					2000 DHS										
									Month of Birth	Year of Birth	Age at Death (Months)	Month of Birth	Year of Birth	Age at Death (Months)
Caseid	Age Group	YSFB	CEB	CD	Caseid	YSFB	CEB	CD	Birth 1			Birth 2		
Emily	20-24	0-5	1	0	*Natalie*	0-5	1	0	5	94				
Emma	25-29	0-5	1	0	*Julia*	0-5	1	0	11	96				
Madison	20-24	0-5	1	0	*Nicole*	0-5	1	0	1	97				
Abigail	25-29	10-14	1	1	Katherine	10-14	1	1	3	90	214			
Isabella	30-34	10-14	1	1	*Chloe*	10-14	1	1	7	90	203			
Ashley	30-34	10-14	1	1	*Chloe*	10-14	1	1	7	90	203			
Samantha	25-29	10-14	1	1	*Rebecca*	10-14	1	1	1	91	107			
Elizabeth	35-39	10-14	1	1	*Sophia*	10-14	1	1	11	90	214			
Erin	40-44	25+	10	2	*Isabel*	25+	10	2	10	96		6	95	
Caroline	35-39	25+	10	2	*Allison*	25+	10	2	10	95		6	93	
Jacqueline	45-49	25+	10	2	*Mary*	25+	10	2	1	86		10	84	

Results and Discussion

We present in figure 1 initial results for Bolivia, on a self-imputed sample stratified on sampling strata. Imputed estimates (m=500) match almost exactly to the true direct estimates of mortality for the Bolivia 2003 population. Figure 2 shows imputation results for Bolivia, matching birth histories from 1998 to women surveyed in 2003 across a range deviations in the original method. We carried out weighted matching, accounting for the differing probabilities of a given woman appearing in the DHS sample, and we carried out unweighted matching using rural/urban, and rural/urban/region as additional matching criteria. We also compared using 3 and 5 year exposure intervals in our matching typology. Results for all four countries will be presented in the final paper. Advantages and disadvantages of the variations on the original method will be discussed.

Figure 1. Self-imputation, matching within sampling strata, m=500.

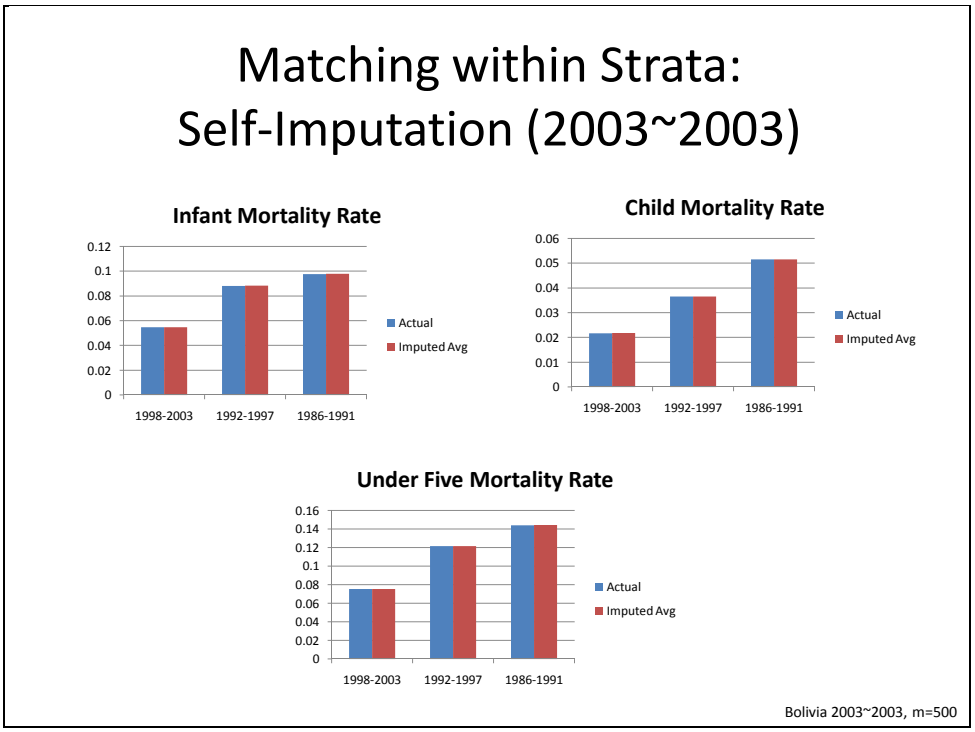
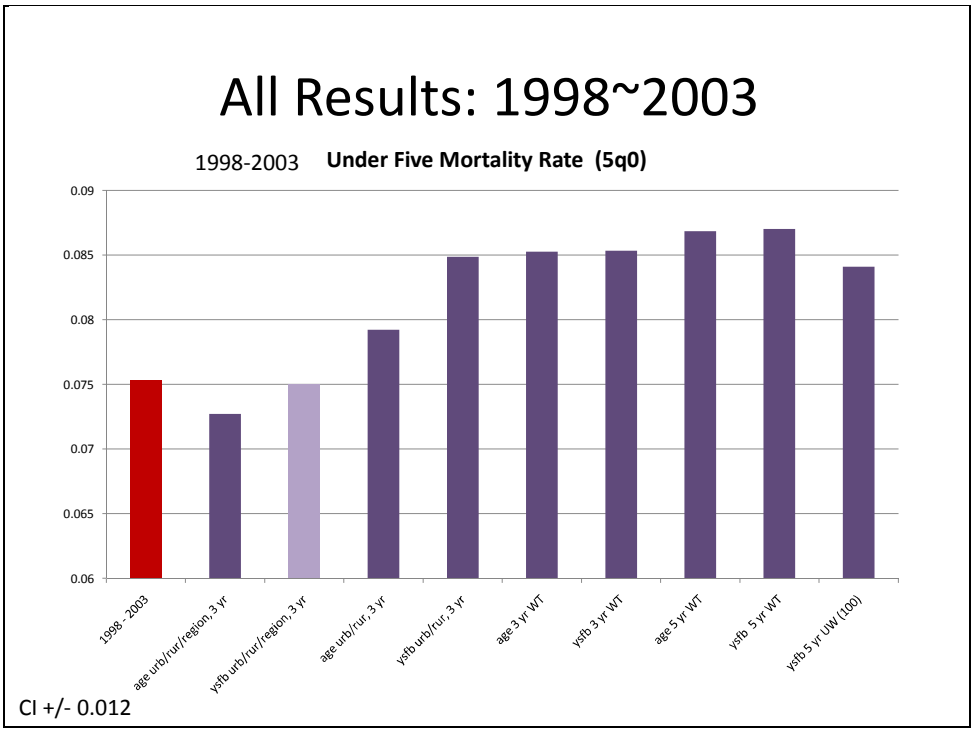


Figure 2. Bolivia actual and imputed estimates from nine iterations of matching, m=500.



The results suggest the method works reasonably well within pairs of countries (Bolivia, Egypt, Kenya and Indonesia) and is relatively unbiased. Not surprisingly, agreement is closest for the broadest measure of child mortality (U5MR) and better for narrower exposure intervals. Initial

results suggest the method works less well across countries (ie. Zambia/Zimbabwe, not shown). We will present comparisons of results for imputations based on age and time since first birth. If further results continue to be encouraging, we believe that this new method will provide the opportunity to extract more informative estimates of under-5 mortality from summary birth histories, unlocking the possibility of obtaining estimates for small areas from population censuses and of conducting expensive full birth history surveys less frequently with little loss of estimation quality.