Estimating HIV Prevalence and Incidence in Africa From Mortality Data

Emily Oster*
University of Chicago and NBER
First Draft: March 25, 2007

This Draft: January 16, 2009

Abstract

An estimated 33 million people are infected with the HIV virus, 67% of them in Sub-Saharan Africa. Despite this, knowledge about HIV prevalence in Africa is limited and imperfect. Although population-based testing in recent years has provided reliable information about *current* prevalence in the general population, we have little reliable data on prevalence in early years of the epidemic. This paper suggests a new methodology for estimating HIV prevalence and incidence using inference from mortality data. This methodology can be used to generate prevalence estimates from early in the epidemic. This information is valuable for understanding how the epidemic has evolved over time and is also likely to be helpful in analyses that explore how policy affects the epidemic or how HIV affects other country-level outcomes.

1 Introduction

An estimated 33 million people, 67% of them in Sub-Saharan Africa, are currently infected with the HIV virus. Despite the enormity of the epidemic, until recent years researchers have had only limited and imperfect knowledge about HIV rates. Prior to the early 2000s, HIV testing was not done systematically among the general population, and was typically limited to pregnant women and high-risk groups such as IV drug users. In recent years, better data has become available on HIV prevalence, based on testing of random samples in the general population, which has led (among other things) to a dramatic reduction in UNAIDS estimates of the magnitude of the epidemic (UNAIDS, 2008).

Although the recent population based testing has improved information about the current magnitude of the epidemic, we still know relatively little about the historical time path in infection

*Ward Cates, Jane Fortson, Daniel Halperin, Norman Hearst, Emir Kamenica, Lawrence Katz, Michael Kremer, Steven Levitt, Kevin Murphy, Ben Olken, James Poterba, Jesse Shapiro, Andrei Shleifer, Rebecca Thornton and participants in seminars at the Gates Foundation, the University of Chicago, UCL, Northwestern, Ohio State and Berkeley provided helpful comments.

rates. This information deficit includes the period before 2000, when testing was typically not done among the general population, *and* knowing even less about the period before the mid-1980s, when no test was available at all. UNAIDS estimates of the time path of the epidemic, to the extent they exist, use modeling assumptions along with the available testing data to provide estimates. Although this can give a sense of the movement in prevalence over time, the estimates tend to be quite sensitive to the assumptions in the model, and ultimately they are only as good as the available testing data, which is very weak.

In this paper I suggest a new methodology for estimating HIV rates in Africa which relies on inference from mortality data. In brief, this methodology begins with the observation that all individuals who die of HIV in a given year must have been infected at some point in the past. Given data on HIV deaths over time, along with information on time to death from HIV, it is possible to infer infection rates in the past. The methodology is similar to "backcalulation", which has been used for this type of analysis in the developed world (Deuffic-Burban and Costagliola, 2006; Ong et al, 1998; Liao and Brookmeyer, 1995). A methodology of this type has been suggested, but not implemented, by Weinreb (1999). The major advantage of this methodology is that – as long as the data on HIV deaths is consistent – it is equally reliable for estimating prevalence in early years of the epidemic as in the present. While the testing data which is available for the present is clearly preferable to the data generated by this inference-based methodology, I believe that these mortality-based estimates dominate historical estimates based largely on modeling assumptions.

I begin in Section 2 of the paper by providing details on the methodology and how estimates are produced. [1] Section 3 then describes the data used in the estimation. The primary data intput is information on HIV deaths, by country and year. Unfortunately, data on AIDS deaths in Africa are extremely unreliable and, as an alternative, we estimate deaths from AIDS by getting data on *total* mortality and subtracting expected non-HIV mortality (based on comparison countries). The intuition behind this strategy can be seen in Figure 1, which graphs deaths by age for Botswana (heavily affected by HIV) and Egypt (mostly unaffected). This figure demonstrates that, among prime-age individuals, death rates in an unaffected country are extremely low. The age pattern of HIV is such that the highest death rates are experienced in age groups with these otherwise very low death rates; this makes the existence of HIV clear in the data. By subtracting "expected deaths" from actual deaths we get an estimate of deaths from HIV/AIDS. The data issues are made more difficult by the lack of reliable data even on *total* mortality rates in Africa over this

[1] Disease incidence refers to number or rate of new infections; prevalence refers to the stock of infections.

period. To get around this I use data from sibling mortality histories to estimate death rates. These histories, and their reliability, are discussed in more detail in Section 3.

Section 4 then presents a set of estimates of HIV prevalence generated using this methodology. Specifically, I show prevalence estimates over time for two age cohorts (individuals 25-35 in the final year of the sample, and those 35-45 in the final year); it is important to note that, in principle, this methodology can provide estimates of prevalence within an age cohort, across age cohorts, within a particular age group over time, etc. The estimates I show here are simply an example of the type of estimates which could be generated; the programs used to generate these data are available from the author. I provide some evidence on the validity of the methodology by comparing the prevalence data generated here to estimates from random population-based testing. I generate prevalence estimates for the most recent years in the sample by country, gender and five-year age group, and compare these with matched samples from population-based testing. I find a strong link between the two: even controlling for country, gender and age-group fixed effects there is a very strong relationship between prevalence from testing and the prevalence estimated based on mortality data. In addition to this basic validation, I provide evidence on how the prevalence estimates differ if we change the assumptions used to generate prevalence.

First and foremost, these data provide new information on the path of the HIV epidemic over time in Africa. In Section 5 I compare trends in the epidemic over time based on these data to trends estimated by UNAIDS in the same countries. I find that, for the period in which both datasets are available, the trends are similar overall. On average, however, the UNAIDS estimates are higher than the mortality-based estimates, particularly in high prevalence areas. This could reflect understatement by the mortality-based data, or it could reflect continued overstatement by UNAIDS, despite their recent revision of the estimates. In addition, it should be noted that the mortality-based data provide a much clearer picture of the beginnings of the epidemic, since they go back significantly further in time; the growth of the epidemic does not seem to be identical in different countries, and this is reflected in the data.

In addition to providing information on trends over time, these data may also be useful in analyzing what factors affect the growth of the HIV epidemic, or what variables are affected by HIV. Oster (2008) uses these data, alongside data from UNAIDS, to estimate the impact of exports on HIV. These data may also be useful in other academic work estimating, for example, the effect of HIV on fertility (Kalemi-Ozcan, 2006; Young, 2006; Forston, 2006) or economic growth (Werker, Ahuja and Wendell, 2006). These papers have typically relied on data from UNAIDS, the US Census

HIV/AIDS Surveillance Database, or they have simply inferred trends over time. The data generated here provide a real alternative to either inconsistent testing data or model-generated prevalence for analyses requiring a time series of HIV rates.

2 Methodology

The methodology used here begins with deaths from HIV in a given year. The output is incidence or prevalence rate, sometime in the past. Section 3.1 below describes how deaths from HIV are calculated; this section starts from the assumption that deaths are known. Denote the number of deaths from HIV among age group i in year t as $\mu_{i,t}$.

Time to death from HIV is a central input to the calculations here. Detailed data on time to death are difficult to generate, particularly in developing countries, since it requires knowing (roughly) the time of infection. For this reason, the best available data are drawn from developed countries, from time periods before HIV treatment was available, and I will use these data. It is possible, of course, that the time to death is different (in particular, perhaps faster) in Africa, but there is no empirical evidence suggesting this is true. In the robustness section I will explore the robustness of the export results to different assumptions about time to death.

Figure 2 shows a graph of time to death from infection, drawn from existing work on South Africa, by age group (Collaborative Group on AIDS Incubation and HIV Survival, 2000; Stover, 2003; Statistics South Africa, 2004). This distribution shows death is common at virtually all periods between 4 and 20 years after infection. Since 90% of deaths occur during this period, I will focus on those 17 years.[2] The figure also demonstrates that time to death is longer for individuals who are infected at younger ages. The analysis across age groups will incorporate these differences. One important note is that in an era with HIV treatment, these paths of time to death are likely to be very different. However, in the period covered by these data, in Africa, treatment levels were extremely low (UNAIDS, 2008), indicating that it is reasonable to use this time path for a situation of no treatment.

Assume that a share d_α of infected people die α years after infection, where d_α is known (for example, taken from Figure 2) and, for the moment, that no one dies from anything else (this assumption is relaxed below, and in generating the estimates). Denoting $b_{i,t}$ as the number of

[2]Extending the time frame to 1-20 years does not make much difference; however, since very early deaths are rare and poorly measured, we begin at 4 years where the share of deaths is large enough to be more reliable. Similarly, truncating the end at 17 years also makes no difference.

infections among age i in year t, we can write the total HIV deaths among age group i in year t ($\mu_{i,t}$) as:

$$\mu_{i,t} = d_4 b_{i-4,t-4} + d_5 b_{i-5,t-5} + ... + d_{20} b_{i-20,t-20} \tag{1}$$

This is a single equation with seventeen unknowns, and it is therefore not identified. Adding another year of data on mortality generates another equation (see below), but also another unknown ($b_{i-21,t-21}$).

$$\mu_{i,t} = d_4 b_{i-4,t-4} + d_5 b_{i-5,t-5} + ... + d_{20} b_{i-20,t-20} \tag{2}$$

$$\mu_{i-1,t-1} = d_4 b_{i-5,t-5} + d_5 b_{i-6,t-6} + ... + d_{20} b_{i-21,t-21} \tag{3}$$

The system of equations will be identified only given data such that

$$\mu_{i-x,t-x} = d_4 b_{i-x-4,t-x-4} \tag{4}$$

That is, we must observe a year in which the only deaths are from people infected four years ago. Assuming that d_4 is known, equation (4) will identify $b_{i-x-4,t-x-4}$ and we can solve the system of equations backwards to solve for the entire $\mathbf{b}$ vector.

There are two ways to fulfill the requirement detailed in equation (4). One is to get data from far enough back in time that there were no infections more than four year ago. The other is to get data from far enough back in the life cycle that this holds (i.e. from age groups who are likely to have had their sexual debut four or fewer years ago). Either procedure (I will focus on the first) will make it possible to solve the system by backward induction.

The discussion above focuses on the case in which individuals infected with HIV die only from HIV, which is a simplification. In reality, some of these individuals would have died anyway, even without HIV infection. If the expected non-HIV death rates are known (as they will be from the data used to calculate deaths from HIV, described in the subsection below) then adjusting the procedure to take this into account is relatively straightforward. Define the probability that an individual would have died even without HIV m years after infection as q_m. In the first year (when all deaths from HIV are due to infections four years before), equation (4) becomes

$$\mu_{i-x,t-x} = d_4 \prod_{i=1}^{3} (1 - q_i)\, b_{i-x-4,t-x-4} \tag{5}$$

This adjustment increases the estimate of HIV infections, since more initial infections are necessary to produce a given number of excess deaths when we assume that some infected people died before

they were killed by HIV. In later years, when deaths depend on more than one year of infection data, the adjustment follows the same principles (although with more elements of the q_i vector).

Having solved for the **b** vector, calculating incidence and prevalence is straightforward. Number of new infections (incidence) is the elements of the **b** vector. Incidence rate among age i in year t is $b_{i,t}$ divided by the population of age i in year t. The calculation of prevalence relies on the **b** vector and the vector of death rates by time from infection. If we denote c_i as the share of individuals still alive i years after infection, then total infections among age i in year t ($p_{i,t}$) is

$$p_{i,t} = b_{i,t} + c_1 b_{i-1,t-1} + c_2 b_{i-2,t-2} + ... + c_{20} b_{i-20,t-20}$$

Prevalence rate is generated by dividing $p_{i,t}$ by the population of age i in year t.

With noiseless data on mortality rates, solving the system of equations for the **b** vector will work perfectly. In practice, however, data on deaths will be noisy. This produces inappropriately high volatility across years – intuitively, a small amount of noise is translated into very large deviations between periods because of the very limited additional information in each of the simultaneous equations.[3] To avoid this issue, I approach the problem as a minimization with a smoothness restriction, rather than strictly as a set of simultaneous equations. Denoting the matrix of death probabilities (d_α) as Θ and the vector of death rates ($\mu_{i,t}$) as Λ, I solve the following minimization.

$$\min_{b_1 ... b_n} \left((\Lambda - \Theta \mathbf{b})' (\Lambda - \Theta \mathbf{b}) \right) + \gamma \left(\sum_{i=2}^{n-1} \left(b_i - \frac{b_{i-1} + b_{i+1}}{2} \right)^2 \right) \tag{6}$$

Without the second part (the expression after γ), this is simply minimized by the solution to the system of equations $\Lambda = \Theta \mathbf{b}$. The second element imposes a cost on the function if any given element of the **b** vector differs from the two surrounding values, which will smooth out the elements of **b**; the degree of smoothness will depend on the size of γ. In the primary analysis, I use a γ value of 0.05; the estimates are not very sensitive to movements in γ, and look similar for values between $\gamma = .01$ and $\gamma = .5$; this is discussed in more detail when I present the estimates.[4]

All estimates of prevalence and incidence are generated by a FORTRAN program, which is available from the author. As we note in the discussion of the data in the following section, the sample sizes used to generate mortality rates are not infinite. This means there is some (observable)

[3] This issue is discussed in more detail, including a simple example, in Appendix A.

[4] Related to the issue of smoothness is the possibility that the function may be minimized when some of the elements of the **b** vector are negative. Since it is not possible for there to be negative HIV infections, I solve the minimization subject to the constraint that all values of $b_1 ... b_n$ are positive. Obviously, if the model was frequently suggesting negative values for elements of the **b** vector, this might call into question the technique. In reality, this is extremely unusual, and the unconstrained results are very similar to the constrained results. To the extent that there are differences, they seem to arise only in early years of the epidemic, and the unconstrained results generally do not allow us to reject zero.

standard error in our estimates of the μ vector. To calculate standard errors of the prevalence estimates, we therefore begin each run of the program by sampling the values of μ from a distribution with the observe mean and standard errors. We then calculate 10,000 estimates of prevalence and incidence, and observe the mean and standard error from the distribution of the results.

3 Data: Estimating HIV Deaths and Comparison HIV Prevalence

This section describes the data used in the paper. The first subsection discusses how I estimate HIV deaths, which are used as inputs in the calculations. The second subsection introduces the comparison data on HIV prevalence rates which will be used to validate these estimates.

3.1 Calculating Deaths from HIV

The methodology described above requires data on deaths from HIV. Because of poor reporting, there are no direct data in Africa on HIV deaths, so deaths from HIV will be calculated by comparing total mortality to expected mortality (based on a non-HIV environment). That is, we will rely on the intuition in Figure 1: because HIV deaths have a very unusual age pattern, if we observe total deaths in an HIV-infected country and total deaths in a country with no HIV but a similar level of development, we can subtract to estimate the number of deaths from HIV.

Total Mortality in HIV Affected Countries One possible source for total mortality is official mortality statistics – for example, those reported in the United Nations Demographic Yearbook. Unfortunately, there are no consistent official statistics for Sub-Saharan Africa. As an alternative, I use information from the Demographic and Health Surveys (DHS) sibling mortality histories to calculate death rates. The DHS are household surveys run in a number of African countries beginning in the late 1980s and early 1990s. The focus of the surveys is fertility behavior and maternal and child health. A number of the surveys included "sibling history" modules in which women were asked to list all of their siblings and give information about their gender, their date of birth and date of death (if deceased). Using these reports, it is possible to construct mortality rates. For example, if we are interested in the mortality rate of 25 year old men in the last year, we can use these data to figure out how many living 25 year old male siblings the survey participants had one year ago, and then observe how many have died in the last twelve months.

Other researchers have used these data to construct mortality profiles in Africa and argued that, on average, the sibling mortality history data match relatively well with official data on

mortality (Bicego, 1997; Timaeus and Jasseh, 2004; Stanton, Abderrahim and Hill, 2000). However, at least one paper has argued that relative to United Nations life tables, sibling histories underestimate mortality (Gakidou, Hogan and Lopez, 2004), although it is possible this reflects *overstatement* by the UN life tables. In Appendix B I discuss in more detail the validity of these data and provide evidence that, where we can compare directly (the Philippines and Zimbabwe), the sibling mortality histories line up very closely with official mortality data. Together with the support from the existing work, this should provide confidence in the use of these data.

A list of the DHS surveys used appears in Table 1. In most cases there are two or three surveys used to generate the death data, typically starting in the early 1990s and ending in the late 1990s or early 2000s. It is clear in the methodology section that it is necessary to have estimates of death rates in each year, starting at the beginning of the epidemic, in order to calculate prevalence and incidence. This means that it is not sufficient to simply estimate death rates in the survey years. However, since individuals are asked about the date of death, it is possible to use these data to estimate mortality rates in the past. Death rates among men aged 25 five years ago are calculated by observing how many living 25 year old male siblings the survey participants had five years ago, and then observe how many have died between four and five years ago. Column 3 of Table 1 reports, for each DHS survey used, the range of years for which that survey generates the data. Although DHS surveys are run in more than the twelve countries (Burkina Faso, Cameroon, Ethiopia, Kenya, Malawi, Mali, Mozambique, Namibia, Uganda, Tanzania, Zambia and Zimbabwe) used here, the specific choice of countries is determined by data availability: I use all countries for which there are DHS sibling histories, and therefore mortality data are available starting in the 1990s.

In the results section I show a variety of prevalence estimates – estimates for two particular age cohorts over time, estimates by gender and five-year age group and estimates over time for all adults. When generating estimates for a particular group, the mortality rate used is the average in that group. For example, to generate an estimate of prevalence for women aged 15-20, I begin by calculating the mortality rate in that entire age group during the relevant period, and using that to estimate overall prevalence. This is in contrast to, for example, estimating prevalence for each single-year age group and averaging them.

The number of individuals used to calculate death rates will vary, both across countries (since the DHS sample size varies) and across different types of estimates. Table 2 gives a sense of the average sample size per cell for the different sets of estimates generated. The smallest sample size is for the five-year age and gender groups, where we observe an average of 1500-3500 individuals

per year; samples are larger for the analysis of larger cohorts. As I noted above, the sample size is not infinite, and standard errors on the mortality rate estimates will be used to generate standard errors on prevalence.

Expected Non-HIV Mortality The other important component of the calculation of HIV deaths is expected mortality from non-HIV causes. I draw these comparison data on mortality from the United Nations Demographic Yearbook Historical Supplement. I rank country-years from those data based on their child mortality, and use the highest-mortality one-third as the comparison set. I assume that adult mortality in the HIV-affected countries in the absence of HIV would be the same as the adult mortality in this comparison set.[5]

This procedure effectively ignores country-specific non-HIV mortality causes, assuming instead that all countries in the sample have similar non-HIV mortality. Of course, in reality we would expect some differences in the non-HIV mortality rates across areas. This analysis relies on the fact that, on average, total adult mortality in areas not affected by HIV seems to be very low. For prime age adults, the expected death rate is generally between 2 and 4 in 1000, which is very small compared to the estimated death rates in HIV-affected countries (as high as 20 in 1000 in some country-years). This large difference means that small perturbations in the non-HIV mortality rate are unlikely to make a large difference in the estimates.[6]

3.2 Existing HIV Prevalence Data

For the primary validation test, I compare the HIV rates estimated from the mortality data to existing HIV prevalence estimates from the DHS surveys themselves, based on testing done in six of the twelve countries in the most recent survey year. The testing is of the general population (all survey respondents are asked to submit to a test), so the estimates should be much more representative and comparable to the mortality-based estimates. The major drawback is that the testing data represent prevalence from a slightly later period than the mortality data (an average of 5 years later) so it is possible that the mortality-based prevalence rates will be somewhat lower. In addition, although these data are clearly more representative than antenatal clinic testing, they do

[5] An alternative to using the United Nations data would be to use DHS sibling histories from the few countries with sibling history data that are unaffected by HIV (only Brazil and the Philippines). Although this may have some advantages, it has the disadvantage that the results could be driven by events in those two countries, which is not ideal. The procedure here smoothes such idiosyncratic events over a much larger number of observations.

[6] One thing that could, in principle, make a large difference here is deaths from war. If the countries in the sample were actively involved in major wars during this time period, this could skew the estimates of HIV rate up, since some of the deaths of prime-age adults would, in fact, have happened in war. Empirically, however, in the countries in this sample, during the sample period, the data suggest there are not many deaths from war (available at http://users.erols.com/mwhite28/warstat3.htm).

suffer from relatively high refusal rates (as high as 40% in some groups). This could bias the estimates either upward or downward, depending on the source of the refusals. Despite this, the comparability between the mortality-based estimates and the DHS is high.

In addition, I compare the estimates generated here to the model-based estimates generated by UNAIDS. These UNAIDS data come from a recent release of country-level data on trends in HIV prevalence (UNAIDS, 2008). Although, in general, the methodology used to generated the UNAIDS estimates is somewhat opaque, these data rely on population-based testing, some antenatal clinic data, conversations with country-level HIV organizations and epidemic modeling. These data are available for all countries in the sample, for the period from 1990 to the present.

4 Results: Estimates of HIV Prevalence

Table 3 shows a baseline set of prevalence estimates. In this table I report HIV prevalence, over time, for two age cohorts: individuals 25-35 in the final year of the sample, and individuals 35-45 in the final year. Standard errors are also reported. This is, of course, just one way to look at variation over time and across countries.

In general, the data generated seem to line up well with levels and trends observed in the epidemic over time. The countries which are known to be high HIV in the sample (for example, Zambia and Uganda) have higher prevalence estimates than those known to be low-prevalence (for example, Burkina Faso and Mali). Most countries show increasing prevalence through the 1980s and early 1990s, and then there is some flattening out, consistent with, for example, modeling of the epidemic by UNAIDS. The data also show a decline in prevalence in Uganda in the early 1990s, which echoes what we see in consistent testing data from that country (i.e. as described in DeWalque (2005)).

Before moving on to validating the estimates with comparison to testing data, it seems useful to get a visual sense of where the identification is coming from – how the data generated relates to the death rates which are the source of variation. To get a sense of this, Figures 3a-3c graph, for three high prevalence countries (Uganda, Zambia and Zimbabwe) excess deaths and estimated incidence rates over time. It is clear from these figures how excess deaths translate into incidence. The shape of the two series are very similar; the difference is that excess deaths in the current year translate into incidence in the past, so the incidence series trails the death series. If we look at Uganda, which has a very striking time path of increased and then decreased incidence, we

see that comes directly from the death series, which increase sharply and then falls down toward the end of the period.

4.1 Validation

The broad observations above provide support for the validity of the mortality-based procedure. As a more formal test, I focus on comparing these mortality-based estimates to data from DHS population testing. In particular, I look at the match between age-gender groups in the two samples. This match will only be possible for six of the countries in the sample with DHS-based testing (Burkina Faso, Cameroon, Kenya, Mali, Zambia and Zimbabwe). I generate estimates by five-year age group (15-20, 20-25, 25-30, 30-35, 35-40 and 40-45) for men and women, and match these with the appropriate estimate from the DHS. I generate data for the most recent three years available, and take the average, to eliminate some of the noise. It is important to recall that the DHS estimates therefore come from an average of 5 years later than the mortality-based estimates so they may, on average, be higher.

The comparison of estimates is shown in Figure 4, which graphs the mortality-based estimates against the DHS estimates and includes the 45 degree line. For a large share of the sample, the mortality-based estimates lie very close to the 45 degree line. The primary exception comes at very high HIV rates, where the DHS estimates are generally somewhat higher than those based on mortality data. This may reflect growth in the epidemic over time (since the testing data are from four years later), or it may reflect underestimates by the mortality data at these high levels (or, less likely, overestimates by the DHS).

Even with the slight mismatching levels at the highest rates, the correlation between the two series is extremely high. This can be seen statistically in Table 4, which shows a regression of the mortality-based prevalence on DHS testing rates by group. Columns 1 and 2 include the whole sample: Column 1 controls for age and gender fixed effects, and Column 2 includes country fixed effects. In both cases, the relationship is positive and highly significant (t-statistic between 10 and 20). Columns 3 and 4 limit the sample to women and men, respectively, and also include age group and country fixed effects. The significant coefficients in these latter columns suggest that the match between the datasets is close enough that it can be identified off of differences in the age pattern of the epidemic in different countries. From the perspective of the export analysis, this close link is quite important. It suggests that the mortality-based estimates are accurately picking up even relatively small variations in HIV prevalence.

A second check on these data relies on a falsification test: what prevalence do the data generate for a country with virtually no HIV? Similar sibling mortality data are available for the DHS surveys in the Philippines and Brazil. Using the same mortality-based procedure, I generate HIV prevalence over time for these two countries. These estimates are reported in Table 5. As in Table 3, these estimates are for a single age cohort, in this case individuals age 25-45 in the final year of the sample. In contrast to the estimates from Africa, the estimated prevalence rates in these control countries are very close to zero in these estimates; indeed, in no year can we distinguish these rates from zero.

The close match found with the DHS and the falsification test above both support the validity of these mortality-based estimates. The section below now briefly discusses the sensitivity of these estimates to the assumptions which generate them.

4.2 Sensitivity to Assumptions Generating HIV Rates

Figures 3a-3c give some sense of how the procedure detailed here converts excess deaths into estimates of HIV rates – HIV incidence tracks death rates, but with a lag. However, there are a number of important assumptions which go into this conversion, and it seems worthwhile briefly exploring how the estimates vary with changes in these assumptions. In this section I explore how the estimates change if we change two of the most important assumptions: time to death from HIV, and the smoothing parameter used to address noise in the mortality data (the use of this parameter is discussed more in Appendix A).

I begin with varying the assumptions about time to death. As noted, in the primary analysis data on time to death from HIV is taken from existing literature (Collaborative Group on AIDS Incubation and HIV Survival, 2000; Stover, 2003; Statistics South Africa, 2004). In addressing robustness, I consider two alternative time paths to death: faster and flatter. The faster time path assumes that everyone has the time to death profile of the oldest age group. The flatter time path assumes that the speed is similar, but deaths are more concentrated in the middle years: 8% of people die every year between 5 and 14 years after infection, rather than the more peaked shape observed in Figure 1. Both alternative paths are illustrated in Appendix Figure 1.

Figure 5a uses data from one country (Uganda) to illustrate how the time path of incidence varies with these different assumptions. The exact path varies somewhat, although the basic pattern in the estimates is consistent across all assumptions. The flatter time to death assumption makes a larger difference, with a higher peak in the late 1980s and lower in the mid-1990s. Columns 1-3 of

Table 6 then reports, under the varying assumptions about time to death, average HIV prevalence for individuals 25-45 (the same cohorts represented in Table 3) in the final year of the sample for each country. Consistent with the evidence from Uganda, changing the time to death has only a limited effect on the final prevalence estimate. On average, the results with flatter death rates are slightly lower (again, consistent with the Uganda evidence), but all three estimates are very similar.

As the second robustness analysis we consider how the results vary if we vary the smoothing parameter, γ, used to generate the estimates. The primary analysis uses a smoothing parameter of $\gamma = .05$; we consider a value of $\gamma = .01$, which means less smoothing, and $\gamma = .5$, which means more smoothing. Figure 5b shows, again for Uganda, how the time path of incidence varies with these assumptions. Again, these changes make relatively little difference. When there is less smoothing the incidence estimates are higher in the late 1980s than with more smoothing, but the overall pattern is very similar.

Columns 4 and 5 of Table 6 reports, under various smoothing assumptions, average HIV prevalence for individuals 25-45 (the same cohorts represented in Table 3) in the final year of the sample for each country. Column 1 of Table 6 includes the estimates with the standard smoothing. As with the changes in time to death, the different smoothing parameters make little difference in the estimates.

These results suggest that the estimate are at least broadly robust to changes in the assumptions used to generate them. The basic patterns in the data are not driven by the exact choice of smoothing or of time to death. However, it should be noted that very large changes in these parameters would, of course, make a difference in the estimates.

5 Comparison with UNAIDS Trend Estimates

Since early in the epidemic, UNAIDS has been publishing periodic reports on epidemic levels, typically inferring prevalence rates from testing of pregnant women, any other testing done in the country, some epidemic modeling and reports from HIV workers within the country. Although prevalence estimates are available at different time periods from these reports, UNAIDS often cautions that rates in different reports are not comparable over time (UNAIDS, 2008). This has made it difficult to infer trends.

In the most recent report (UNAIDS, 2008), new data was provided on estimated trends in the epidemic over time. These data cover the period from 1990 to the present and, again, were

generated using a combination of data from HIV testing, modeling assumptions and other information. The data generated here provide an interesting comparison for the UNAIDS data; allowing us to look at how the trends and levels compare and ask what additional information is contributed by observing the earlier period.

Figure 6 shows, for each country-year in the sample, the UNAIDS prevalence rate and the mortality-based rate for individuals between 25 and 45 graphed over time. The mortality-based data are generated separately for each year in the sample, which means that smoothness between years is not imposed on the data. There are several things to note in these data. First, overall the trends in the two datasets match relatively well. By construction, the UNAIDS data is very smooth, and the mortality-based data need not be. Nevertheless, the overall trend picture over the post-1990 period is very similar in nearly all countries. One interesting exception is Uganda, where the UNAIDS data has a consistent but small decline in prevalence over this period, whereas the mortality-based data indicates a much larger decline, followed by some rebound in the post-1995 period.

In terms of levels, the story is more mixed. In a number of countries – Burkina Faso, Cameroon, Ethiopia, Malawi, Mali, Uganda, Zambia – the levels match quite well. In others, it seems that the UNAIDS estimates are much higher. This is most evidence in Kenya, Tanzania and Zimbabwe, where the UNAIDS data suggests similar trends, but the estimates are two or three times as high. Given the recent revision of the estimates by UNAIDS it seems plausible that this disparity reflects continued over-estimation by UNAIDS, although this is difficult to say conclusively.

Finally, we can see from these graphs that the trends in the epidemic *before* the start of the UNAIDS data in 1990 are actually quite different across countries. For example, if we compare Uganda and Zambia, both fairly high prevalence countries, we can see that the epidemic clearly started much earlier in Uganda; prevalence rates take off there in the early 1980s, and not until the late 1980s in Zambia. The evidence on Kenya suggests extremely low prevalence rates until the 1990 period, whereas in Malawi the epidemic was beginning by the early 1980s. The fact that there is significant variation in these pre-period trends points to the value of being able to observe the whole time period of the epidemic, rather than just a recent subset.

6 Conclusion

In this paper I suggest a new methodology for estimating historical HIV prevalence in Africa using inference from mortality data. I argue that, where it is possible to compare, these data match well to

data from population-based testing. To my knowledge, this methodology generates among the first estimates of prevalence in the early period which are *not* based solely on modeling assumptions. These data provide new insight into how quickly the epidemic began, where it spread earliest and how it has changed over time. In addition, these data provide a check on the new data on trends in prevalence over time reported by UNAIDS. I find that the trends in the UNAIDS data line up well with the trends in the mortality-based, although the UNAIDS levels – despite the adjustment downward in the 2008 report – are still higher that what we would calculate based on deaths.

As noted in the introduction, these data may also be useful in doing analyses which rely on variation in HIV over time – for example, analyses which consider what factors impact the growth of HIV (Oster, 2008) and those which consider the effects of HIV on other variables (Fortson, 2008; Kalemi-Ozcan, 2006). To this end, the programs which generate these estimates are available from the author, and using these it is possible to generate estimates for other age groups, gender groups or time periods.

References

Bicego, George, "Estimating adult mortality rates in the context of the AIDS epidemic in sub-Saharan Africa: analysis of DHS sibling histories," *Health Transition Review*, 1997, *7*.

Collaborative Group on AIDS Incubation and HIV Survival, "Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis," *The Lancet*, 2000, *355*, 1131–1137.

Deuffic-Burban, Sylvie and Deminique Costagliola, "Including Pre-AIDS Mortality in Back-Calculation Model to Estiamte HIV Prevalence in France, 2000," *European Journal of Epidemiology*, 2006, *21*, 389–396.

DeWalque, Damien, "How Does the Impact of an HIV/AIDS Information Campaign Vary With Educational Attainment: Evidence from Rural Uganda," 2005. Mimeo, World Bank.

Editorial Desk, "The State of AIDS," *The New York Times*, 2005, *December 1, 2005.*

Feeney, Griffith, "The Impact of HIV/AIDS on Adult Mortality in Zimbabwe," *Population and Development Review*, 2001, *27* (4), 771–780.

Forston, Jane, "Mortality Risks and Human Capital Investment: The Impact of HIV/AIDS in Sub-Saharan Africa," 2006. Mimeo, Princeton University.

Gakidou, Emmanuela and Gary King, "Death by Survey: Estimating Adult Mortality without Selection Bias from Sibling Survival Data," *Demography*, 2006, *43* (3), 569–585.

___ , Margaret Hogan, and Alan D Lopez, "Adult mortality: time for a reappraisal," *International Journal of Epidemiology*, 2004, *33*, 1–8.

Kalemi-Ozcan, Sebnam, ""AIDS, Reversal of the Demographic Transition and Economic Development: Evidence from Africa"," *NBER Working Paper, No. 12181*, 2006.

Liao, J and R Brookmeyer, "An Empirical Bayes Approach to Smoothing in Backcalculation of HIV Infection Rates," *Biometrics*, 1995, *51* (2), 579–588.

Maugh, Thomas, "AIDS Growth Slowing Worldwide, U.N. Finds," *The Los Angeles Times*, 2006, *May 31, 2006.*

Murphy, Kevin and Finis Welch, "Empirical Age-Earnings Profiles," *Journal of Labor Economics*, 1990, *8* (2), 202–229.

Ong, HC, SH Quah, and HC Low, "An Application of the Backcalculation Method to Estimate Past HIV Infection Rates in Malaysia," *Medical Journal of Malaysia*, 1998, *53* (4), 385–391.

Oster, Emily, "Routes of Infection: Exports and HIV Incidence in Sub-Saharan Africa," 2008. Mimeo, University of Chicago.

Stanton, Cynthia, Noureddine Abderrahim, and Kenneth Hill, "An Assessment of DHS Maternal Mortality Indicators," *Studies in Family Planning*, 2000, *31* (2), 111–123.

Statistics South Africa, "Mid-year population estimates, South Africa," Technical Report, Statistics South Africa 2004.

Stover, James, "AIM version 4. A computer program for HIV/AIDS projections and examining the social and economic impacts of AIDS," *Spectrum system of Policy Models. The Futures Group International.*, 2003.

Timaeus, Ian and Momodou Jasseh, "Adult Mortality in Sub-Saharan Africa: Evidence from the Demographic and Health Surveys," *Demography*, 2004, *41* (4), 757–772.

Trussell, James and German Rodriguez, "A Note on the Sisterhood Estimator of Maternal Mortality," *Studies in Family Planning*, 1990, *21* (6), 344–346.

UNAIDS, *2008 Report on the global AIDS epidemic*, Joint United Nations Program on HIV/AIDS, 2008.

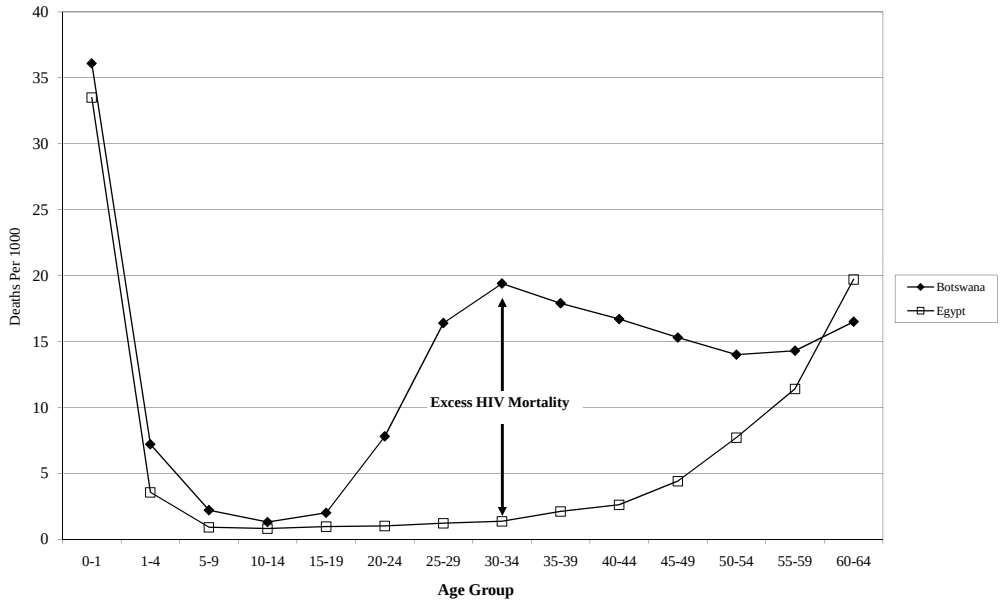
United Nations, *Demographic Yearbook*, United Nations, 2001.

Weinreb, Alex, "Estimating HIV+ Incidence from Mortality Rates: A Method and Agenda," 1999. Presentation, Durbin South Africa.

Werker, Eric, Amrita Ahuja, and Brian Wendell., "Male Circumcision and AIDS: The Macroeconomic Impact of a Health Crisis," 2006. Harvard Business School Working Paper.

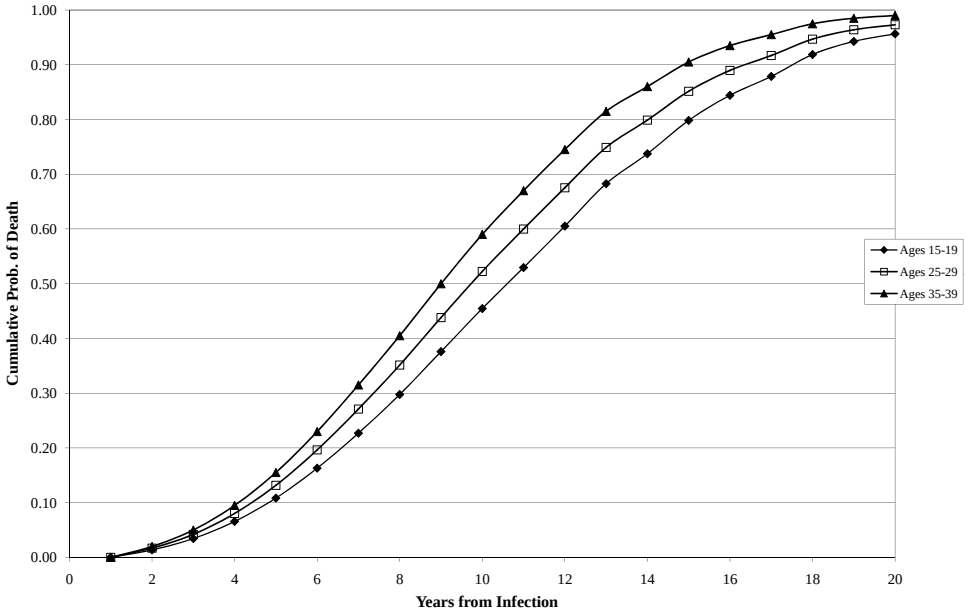
Young, Alwyn, "In Sorrow to Bring Forth Children: Fertility amidst the Plague of HIV," 2006. Mimeo, University of Chicago.

Figure 1:
Death Rates by Age, Botswana in 2001 and Egypt, 1990s



Notes: Death rates are from the United Nations Demographic Yearbook, 2002 edition and Historical Supplement. The Botswana data is for 2001, and the Egypt data is an average for the 1990s. Death rates are deaths (all individuals, both genders) per 1000 people in the age group. Botswana is affected by HIV; Egypt is largely not in this period.

Figure 2:
Distribution of Time from HIV Infection to Death



Notes: Figure presents the cumulative probability of death as a function of years from infection with HIV, by age group. The figure is based on data from Collaborative Group on AIDS Incubation and HIV Survival, 2000.

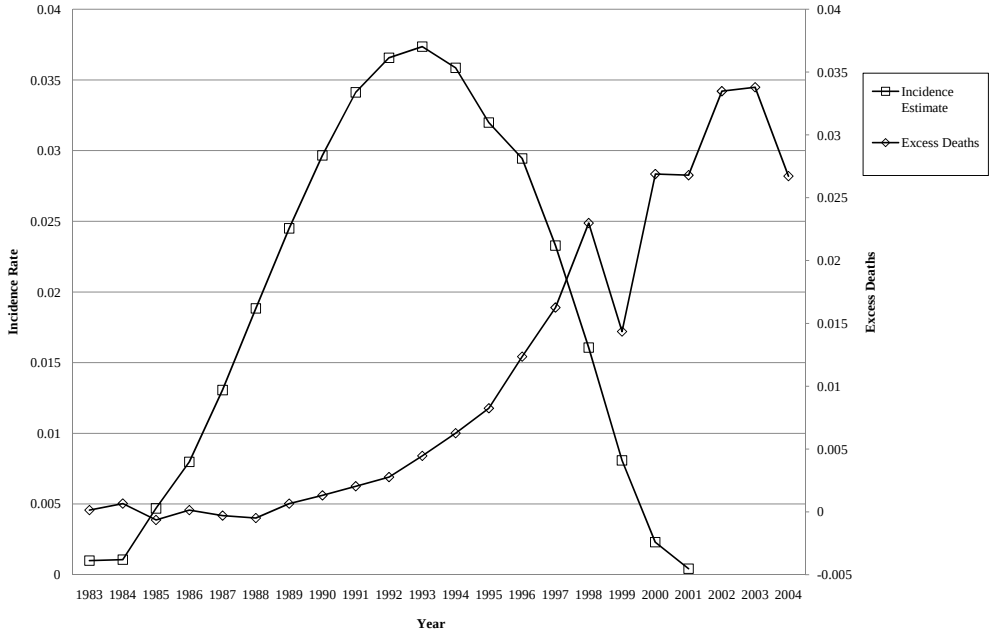
Figure 3a:
Excess Deaths and Incidence Estimates, Uganda

Notes This figure shows the excess deaths and the estimate of incidence rate, by year, for Uganda.

Figure 3b:
Excess Deaths and Incidence Estimates, Zambia

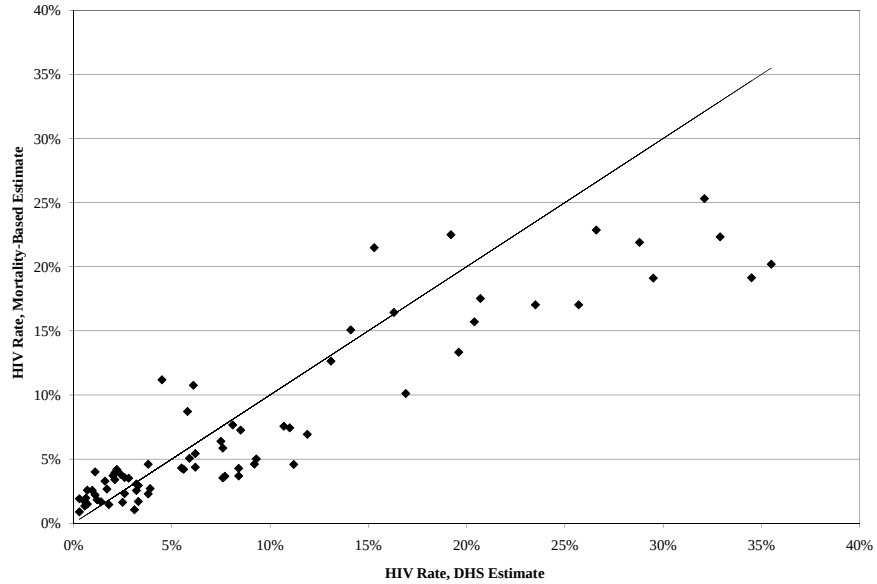
Notes This figure shows the excess deaths and the estimate of incidence rate, by year, for Zambia.

Figure 3c:
Excess Deaths and Incidence Estimates, Zimbabwe



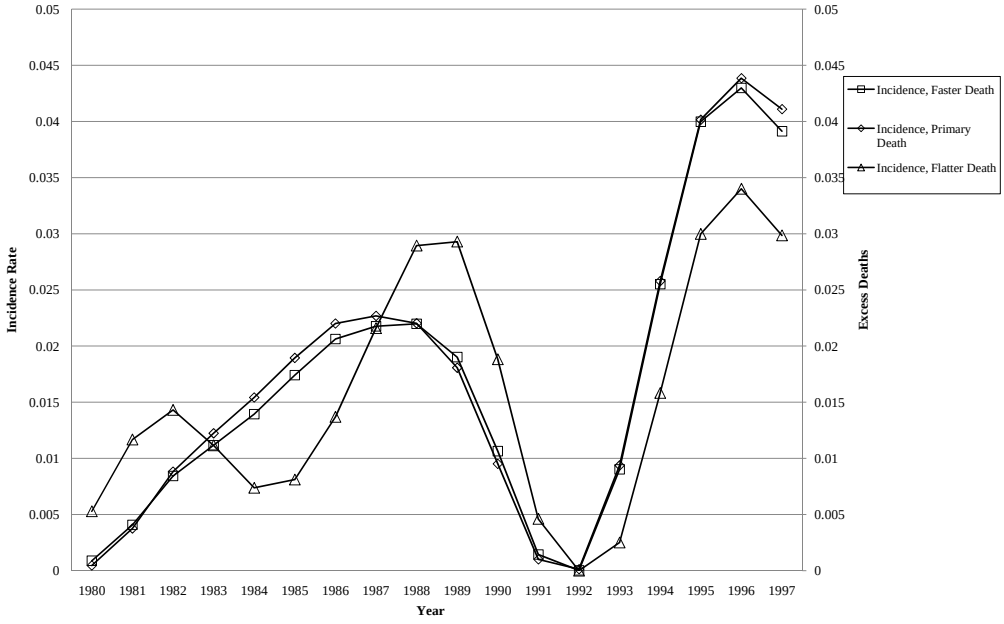
Notes This figure shows the excess deaths and the estimate of incidence rate, by year, for Zimbabwe.

Figure 4:
Mortality-Based Estimates and DHS Testing Estimates by Country-Gender-Age Group



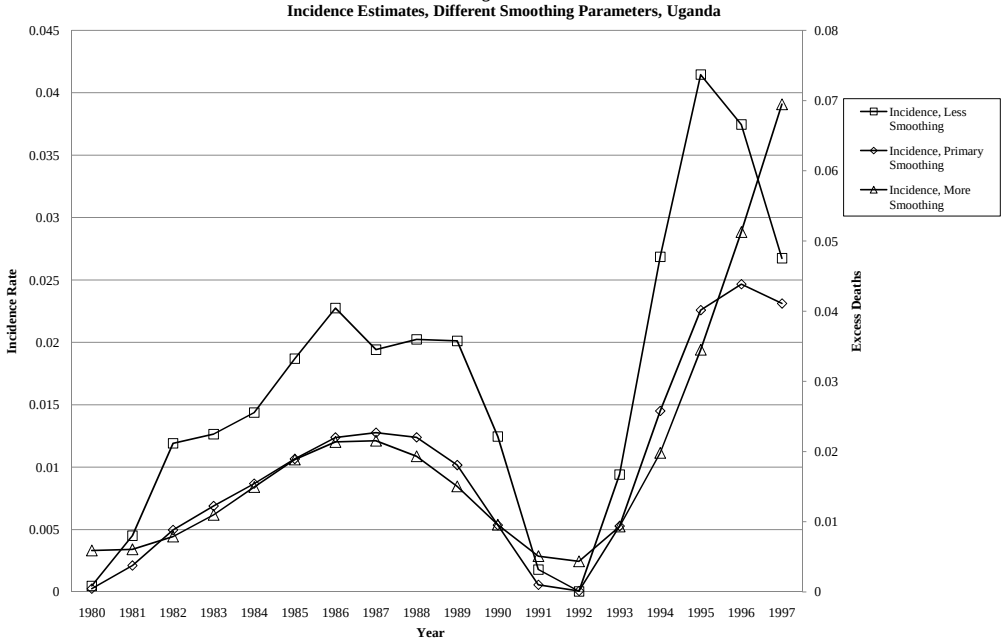
Notes: This figure shows HIV prevalence estimated by the mortality data for each country-gender-age group (5-year age groups) for the most recent year in the sample graphed against the DHS estimated prevalence from an average of five years later (average of three years of mortliaty estimates). Countries include: Burkina Faso, Cameroon, Kenya, Mali, Zambia and Zimbabwe.

Figure 5a:
Incidence Estimates, Different Time to Death, Uganda



Notes This figure shows estimates of incidence, by year, for Uganda with different assumptions about time to death. The different time paths to death can be seen in Appendix Figure 1.

Figure 5b:
Incidence Estimates, Different Smoothing Parameters, Uganda

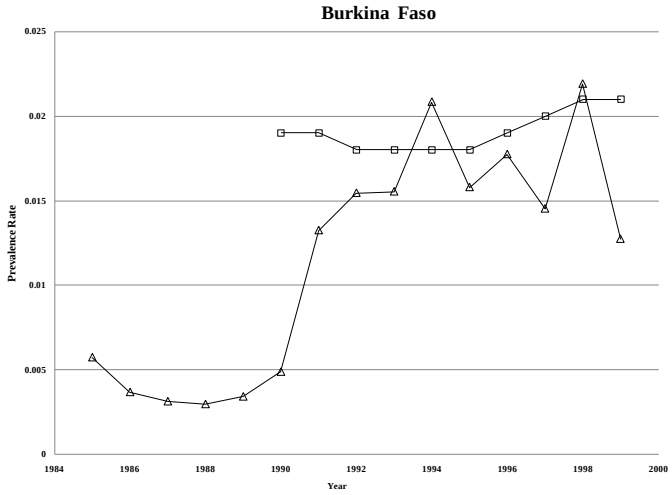


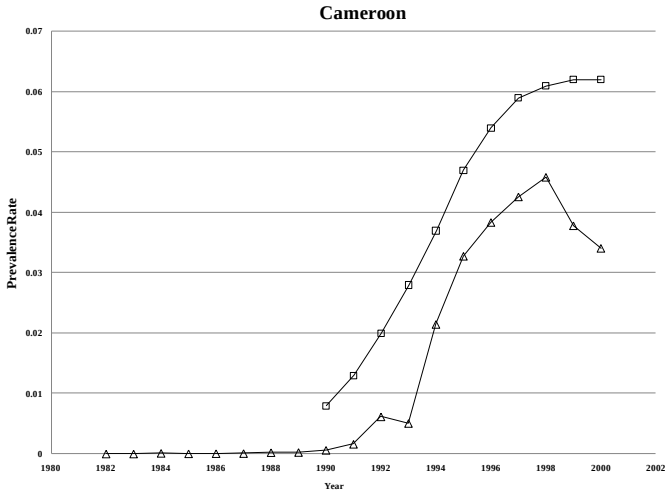
Notes This figure shows estimates of incidence, by year, for Uganda with different smoothing parameters. The standard smoothing is gamma=.05; less smooth is gamma=.01; more smooth is gamma=.5

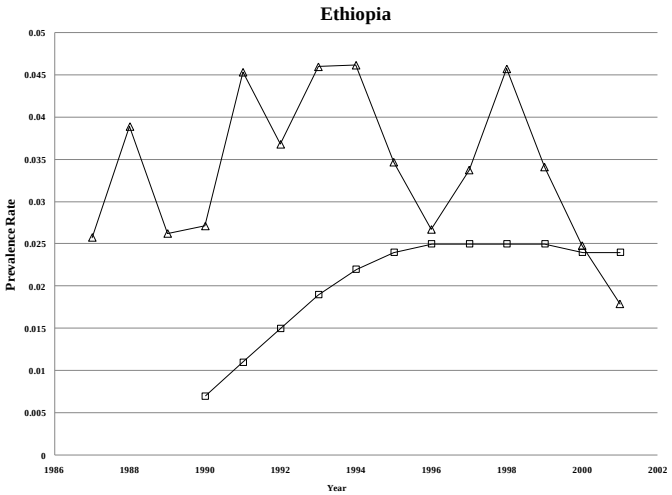
Figure 6:

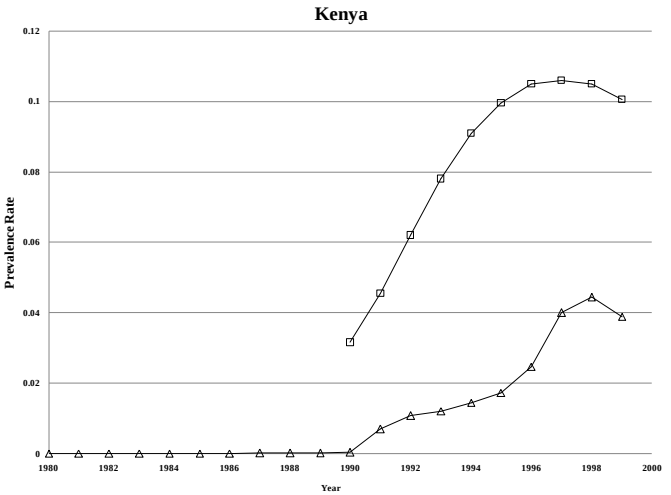
UNAIDS and Mortality-Based Estimates

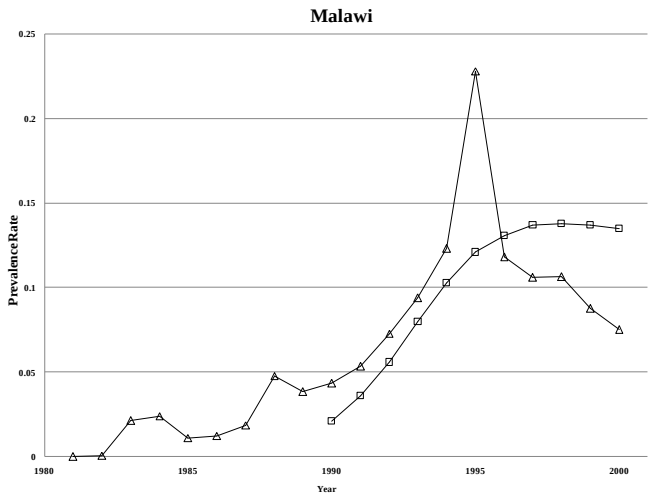
□ UNAIDS Estimate △ Mortality-Based Estimate

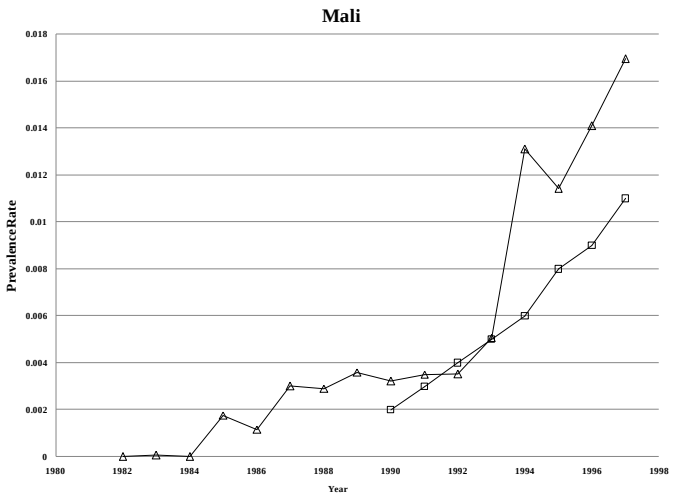


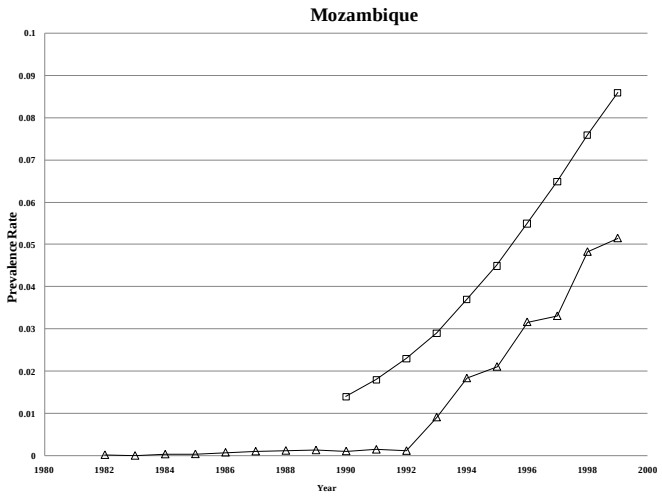


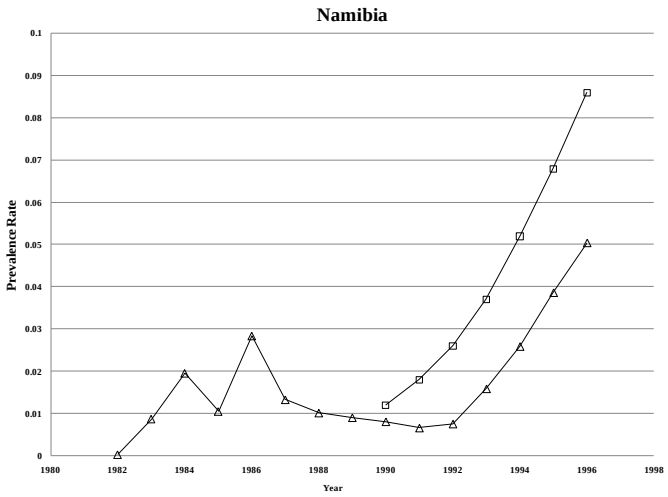


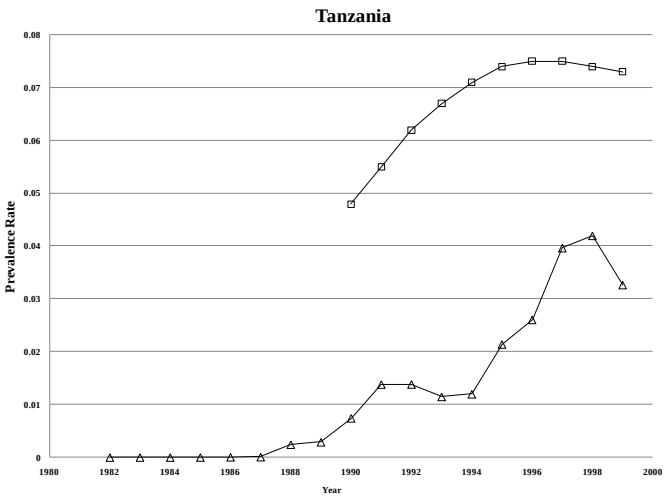




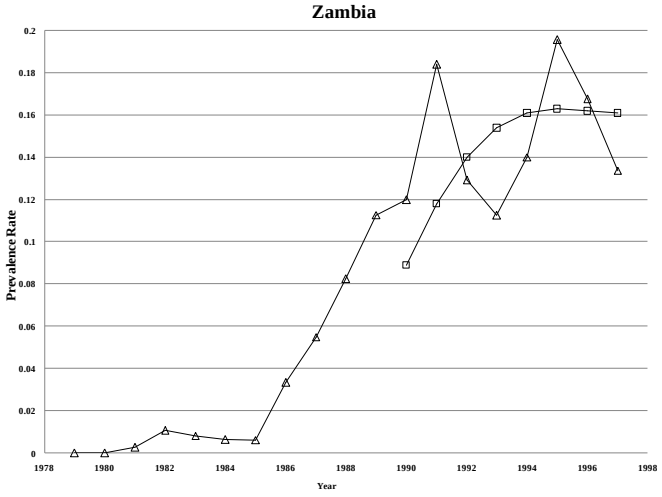



Mozambique
Prevalence Rate
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
Year


Namibia
Prevalence Rate
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
Year


Tanzania
Prevalence Rate
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
Year


Uganda
Prevalence Rate
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
Year


Zambia
Prevalence Rate
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1978
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
Year

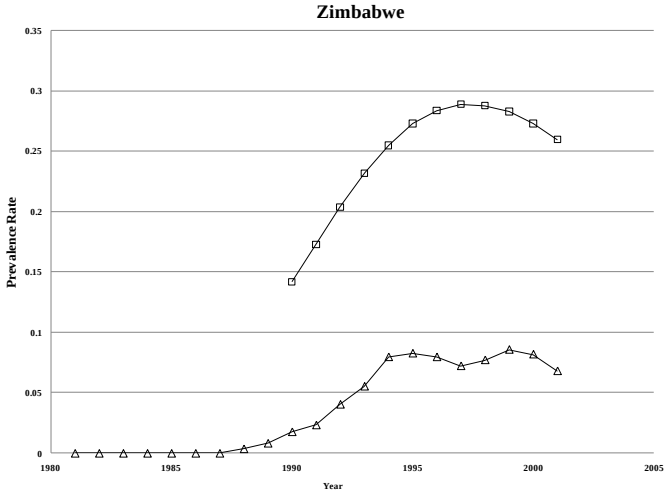

Zimbabwe
Prevalence Rate
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1980
1985
1990
1995
2000
2005
Year

Table 1. ***Demographic and Health Surveys Used***

Country	Year of Survey	Years of Mortality Data Generated
Burkina Faso	2003	1999-2003
Burkina Faso	1998	1984-1998
Cameroon	2004	1999-2004
Cameroon	1998	1985-1998
Ethiopia	2005	2001-2005
Ethiopia	2000	1984-2000
Kenya	2003	1999-2003
Kenya	1998	1982-1998
Mali	2001	1997-2001
Mali	1996	1984-1996
Malawi	2004	2001-2004
Malawi	2000	1993-2000
Malawi	1992	1982-1992
Mozambique	2003	1998-2003
Mozambique	1997	1983-1997
Namibia	2000	1993-2000
Namibia	1992	1986-1992
Tanzania	2004	1997-2004
Tanzania	1996	1981-1996
Uganda	2001	1996-2001
Uganda	1995	1980-1995
Zambia	2001	1997-2001
Zambia	1996	1980-1996
Zimbabwe	2005	2000-2005
Zimbabwe	1999	1995-1999
Zimbabwe	1994	1983-1994

Notes: This table lists the DHS surveys used in generating the prevalence and incidence estimates.

Table 2. ***Demographic and Health Surveys Sample Size***

Country	**Average Siblings Observed, Per Year, for Calculations**		
	Five Year Age Group (Figure 4)	**Cohort Analysis (Table 3)**	**UNAIDS Comparison (Figure 6)**
Burkina Faso	1,896	6,631	13,009
Cameroon	1,939	6,768	13,248
Ethiopia	3,565	12,750	24,932
Kenya	2,646	9,854	19,276
Malawi	2,295	8,172	15,939
Mali	2,566	9,238	18,140
Mozambique	2,369	8,597	16,811
Namibia	1,757	6,438	12,622
Tanzania	2,657	10,059	19,737
Uganda	2,148	7,757	15,262
Zambia	2,353	8,199	16,026
Zimbabwe	2,099	8,169	15,793

Notes: This table reports the average sample size (i.e. number of siblings) used to calculate mortality rates for the three sets of estimates generated in the paper. The numbers are an average across all years in the sample.

Table 3. ***HIV Prevalence Estimates***

Country	Year	Prevalence Age Cohort 25-35 in Final Year	Prevalence Age Cohort 35-45 in Final Year
Burkina Faso	1980	0.459%(0.220%)	0.284%(0.111%)
Burkina Faso	1981	0.951%(0.453%)	0.737%(0.294%)
Burkina Faso	1982	0.951%(0.453%)	0.737%(0.294%)
Burkina Faso	1983	1.256%(0.615%)	1.565%(0.617%)
Burkina Faso	1984	1.217%(0.595%)	1.907%(0.713%)
Burkina Faso	1985	1.152%(0.561%)	2.068%(0.780%)
Burkina Faso	1986	1.052%(0.511%)	2.254%(0.857%)
Burkina Faso	1987	1.065%(0.486%)	2.515%(0.957%)
Burkina Faso	1988	1.284%(0.508%)	2.729%(1.063%)
Burkina Faso	1989	1.725%(0.581%)	2.903%(1.144%)
Burkina Faso	1990	2.335%(0.687%)	2.815%(1.138%)
Burkina Faso	1991	2.895%(0.774%)	2.524%(1.027%)
Burkina Faso	1992	3.261%(0.813%)	2.227%(0.911%)
Burkina Faso	1993	3.358%(0.789%)	2.085%(0.818%)
Burkina Faso	1994	3.229%(0.717%)	2.300%(0.801%)
Burkina Faso	1995	3.006%(0.633%)	2.739%(0.852%)
Burkina Faso	1996	2.746%(0.552%)	3.217%(0.920%)
Burkina Faso	1997	2.533%(0.490%)	3.719%(1.000%)
Burkina Faso	1998	2.372%(0.443%)	4.153%(1.052%)
Burkina Faso	1999	2.242%(0.399%)	4.250%(1.043%)
Cameroon	1981	0.320%(0.200%)	0.179%(0.212%)
Cameroon	1982	0.535%(0.351%)	0.352%(0.431%)
Cameroon	1983	0.535%(0.351%)	0.352%(0.431%)
Cameroon	1984	0.552%(0.349%)	0.469%(0.501%)
Cameroon	1985	0.528%(0.322%)	0.460%(0.469%)
Cameroon	1986	0.523%(0.307%)	0.460%(0.419%)
Cameroon	1987	0.611%(0.329%)	0.515%(0.342%)
Cameroon	1988	0.875%(0.383%)	0.757%(0.313%)
Cameroon	1989	1.328%(0.447%)	1.241%(0.366%)
Cameroon	1990	1.874%(0.510%)	1.917%(0.515%)
Cameroon	1991	2.529%(0.574%)	2.795%(0.744%)
Cameroon	1992	3.275%(0.650%)	3.769%(1.016%)
Cameroon	1993	4.083%(0.745%)	4.687%(1.269%)
Cameroon	1994	4.879%(0.846%)	5.459%(1.459%)
Cameroon	1995	5.433%(0.906%)	5.967%(1.546%)
Cameroon	1996	5.677%(0.911%)	6.168%(1.525%)
Cameroon	1997	5.672%(0.877%)	6.111%(1.432%)
Cameroon	1998	5.467%(0.819%)	5.877%(1.318%)
Cameroon	1999	5.092%(0.748%)	5.530%(1.235%)
Cameroon	2000	4.724%(0.697%)	5.248%(1.249%)
Ethiopia	1980	1.872%(0.242%)	2.016%(0.372%)
Ethiopia	1981	3.684%(0.491%)	3.899%(0.754%)
Ethiopia	1982	3.684%(0.491%)	3.899%(0.754%)
Ethiopia	1983	4.141%(0.571%)	4.326%(0.841%)
Ethiopia	1984	3.990%(0.545%)	4.241%(0.731%)
Ethiopia	1985	3.933%(0.504%)	4.381%(0.650%)
Ethiopia	1986	4.006%(0.482%)	4.738%(0.640%)

continued on the next page

Country	Year	Prevalence Age Cohort 25-35 in Final Year	Prevalence Age Cohort 35-45 in Final Year
Ethiopia	1987	4.022%(0.474%)	4.977%(0.671%)
Ethiopia	1988	4.144%(0.482%)	5.090%(0.712%)
Ethiopia	1989	4.380%(0.504%)	5.090%(0.750%)
Ethiopia	1990	4.692%(0.535%)	5.089%(0.798%)
Ethiopia	1991	4.928%(0.562%)	5.049%(0.849%)
Ethiopia	1992	5.027%(0.577%)	4.818%(0.869%)
Ethiopia	1993	5.013%(0.584%)	4.383%(0.839%)
Ethiopia	1994	4.856%(0.588%)	3.867%(0.751%)
Ethiopia	1995	4.856%(0.624%)	3.534%(0.731%)
Ethiopia	1996	4.819%(0.656%)	3.360%(0.787%)
Ethiopia	1997	4.619%(0.660%)	3.349%(0.878%)
Ethiopia	1998	4.224%(0.628%)	3.378%(0.938%)
Ethiopia	1999	3.730%(0.555%)	3.444%(0.959%)
Ethiopia	2000	3.288%(0.491%)	3.777%(0.992%)
Ethiopia	2001	2.887%(0.432%)	4.412%(1.056%)
Kenya	1978	0.115%(0.088%)	0.028%(0.044%)
Kenya	1979	0.250%(0.185%)	0.055%(0.081%)
Kenya	1980	0.250%(0.185%)	0.055%(0.081%)
Kenya	1981	0.334%(0.220%)	0.140%(0.100%)
Kenya	1982	0.335%(0.206%)	0.266%(0.156%)
Kenya	1983	0.325%(0.187%)	0.411%(0.267%)
Kenya	1984	0.316%(0.167%)	0.500%(0.332%)
Kenya	1985	0.315%(0.141%)	0.519%(0.341%)
Kenya	1986	0.420%(0.165%)	0.612%(0.371%)
Kenya	1987	0.675%(0.239%)	0.846%(0.440%)
Kenya	1988	1.089%(0.357%)	1.257%(0.533%)
Kenya	1989	1.580%(0.474%)	1.711%(0.606%)
Kenya	1990	2.086%(0.542%)	2.197%(0.644%)
Kenya	1991	2.503%(0.534%)	2.746%(0.669%)
Kenya	1992	2.810%(0.476%)	3.290%(0.687%)
Kenya	1993	3.149%(0.441%)	3.806%(0.730%)
Kenya	1994	3.665%(0.485%)	4.482%(0.878%)
Kenya	1995	4.331%(0.636%)	5.370%(1.186%)
Kenya	1996	5.026%(0.874%)	6.283%(1.590%)
Kenya	1997	5.587%(1.124%)	6.880%(1.897%)
Kenya	1998	5.827%(1.309%)	6.839%(1.949%)
Kenya	1999	5.793%(1.414%)	6.392%(1.781%)
Malawi	1978	0.752%(0.260%)	0.234%(0.231%)
Malawi	1979	1.549%(0.544%)	0.585%(0.492%)
Malawi	1980	1.549%(0.544%)	0.585%(0.492%)
Malawi	1981	2.340%(0.772%)	1.822%(0.745%)
Malawi	1982	2.358%(0.767%)	2.776%(0.842%)
Malawi	1983	2.248%(0.726%)	3.752%(0.964%)
Malawi	1984	2.092%(0.672%)	4.409%(1.059%)
Malawi	1985	1.953%(0.619%)	4.710%(1.094%)
Malawi	1986	2.144%(0.587%)	5.053%(1.075%)
Malawi	1987	2.596%(0.566%)	5.441%(1.009%)
Malawi	1988	3.273%(0.550%)	5.982%(0.920%)
Malawi	1989	4.171%(0.542%)	6.912%(0.855%)

continued on the next page

Country	Year	Prevalence Age Cohort 25-35 in Final Year	Prevalence Age Cohort 35-45 in Final Year
Malawi	1990	5.270%(0.561%)	8.287%(0.868%)
Malawi	1991	6.529%(0.616%)	10.174%(0.994%)
Malawi	1992	7.881%(0.698%)	12.483%(1.216%)
Malawi	1993	9.560%(0.825%)	15.463%(1.526%)
Malawi	1994	10.841%(0.932%)	17.445%(1.738%)
Malawi	1995	11.229%(0.980%)	17.485%(1.751%)
Malawi	1996	10.914%(0.992%)	16.155%(1.624%)
Malawi	1997	10.434%(0.988%)	14.740%(1.488%)
Malawi	1998	9.883%(0.964%)	13.788%(1.474%)
Malawi	1999	9.017%(0.895%)	13.358%(1.712%)
Malawi	2000	8.004%(0.796%)	14.228%(2.367%)
Mali	1980	0.536%(0.209%)	0.534%(0.260%)
Mali	1981	1.000%(0.413%)	1.093%(0.558%)
Mali	1982	1.000%(0.413%)	1.093%(0.558%)
Mali	1983	1.105%(0.461%)	1.500%(0.775%)
Mali	1984	1.081%(0.443%)	1.486%(0.757%)
Mali	1985	1.049%(0.418%)	1.452%(0.727%)
Mali	1986	1.029%(0.399%)	1.360%(0.669%)
Mali	1987	1.027%(0.383%)	1.254%(0.601%)
Mali	1988	1.073%(0.368%)	1.128%(0.532%)
Mali	1989	1.166%(0.358%)	1.056%(0.475%)
Mali	1990	1.309%(0.365%)	1.111%(0.464%)
Mali	1991	1.581%(0.398%)	1.371%(0.516%)
Mali	1992	1.997%(0.466%)	1.808%(0.626%)
Mali	1993	2.416%(0.540%)	2.338%(0.772%)
Mali	1994	2.733%(0.592%)	2.895%(0.926%)
Mali	1995	2.932%(0.623%)	3.438%(1.090%)
Mali	1996	3.048%(0.636%)	3.903%(1.240%)
Mali	1997	3.155%(0.648%)	4.156%(1.341%)
Mozambique	1979	0.555%(0.222%)	0.319%(0.244%)
Mozambique	1980	0.957%(0.413%)	0.579%(0.466%)
Mozambique	1981	0.957%(0.413%)	0.579%(0.466%)
Mozambique	1982	1.103%(0.495%)	0.741%(0.604%)
Mozambique	1983	1.092%(0.492%)	0.745%(0.585%)
Mozambique	1984	1.055%(0.468%)	0.738%(0.550%)
Mozambique	1985	1.002%(0.429%)	0.719%(0.506%)
Mozambique	1986	0.926%(0.386%)	0.693%(0.447%)
Mozambique	1987	0.864%(0.331%)	0.673%(0.366%)
Mozambique	1988	0.914%(0.288%)	0.764%(0.312%)
Mozambique	1989	1.125%(0.278%)	0.997%(0.300%)
Mozambique	1990	1.521%(0.302%)	1.453%(0.335%)
Mozambique	1991	2.079%(0.358%)	2.051%(0.415%)
Mozambique	1992	2.913%(0.468%)	2.969%(0.595%)
Mozambique	1993	3.809%(0.593%)	4.061%(0.848%)
Mozambique	1994	4.548%(0.697%)	5.004%(1.089%)
Mozambique	1995	5.162%(0.791%)	5.822%(1.306%)
Mozambique	1996	5.715%(0.884%)	6.493%(1.491%)
Mozambique	1997	6.113%(0.963%)	6.942%(1.632%)
Mozambique	1998	6.211%(0.996%)	7.103%(1.712%)

continued on the next page

Country	Year	Prevalence Age Cohort 25-35 in Final Year	Prevalence Age Cohort 35-45 in Final Year
Mozambique	1999	5.956%(0.967%)	6.981%(1.728%)
Namibia	1982	0.613%(0.260%)	1.162%(0.471%)
Namibia	1983	1.177%(0.520%)	2.105%(0.905%)
Namibia	1984	1.400%(0.637%)	2.450%(1.073%)
Namibia	1985	1.367%(0.629%)	2.393%(1.035%)
Namibia	1986	1.331%(0.617%)	2.283%(0.975%)
Namibia	1987	1.243%(0.553%)	2.108%(0.881%)
Namibia	1988	1.186%(0.491%)	1.911%(0.790%)
Namibia	1989	1.535%(0.524%)	2.008%(0.749%)
Namibia	1990	2.309%(0.614%)	2.507%(0.819%)
Namibia	1991	3.393%(0.734%)	3.287%(0.974%)
Namibia	1992	4.539%(0.852%)	4.249%(1.191%)
Namibia	1993	5.525%(0.942%)	5.211%(1.412%)
Namibia	1994	6.339%(1.005%)	6.111%(1.622%)
Namibia	1995	6.718%(1.008%)	6.682%(1.742%)
Namibia	1996	6.791%(0.970%)	6.749%(1.740%)
Tanzania	1978	0.224%(0.128%)	0.084%(0.109%)
Tanzania	1979	0.446%(0.261%)	0.165%(0.219%)
Tanzania	1980	0.446%(0.261%)	0.165%(0.219%)
Tanzania	1981	0.512%(0.305%)	0.208%(0.249%)
Tanzania	1982	0.493%(0.293%)	0.237%(0.219%)
Tanzania	1983	0.485%(0.284%)	0.368%(0.199%)
Tanzania	1984	0.539%(0.296%)	0.650%(0.242%)
Tanzania	1985	0.673%(0.326%)	1.137%(0.374%)
Tanzania	1986	0.856%(0.355%)	1.805%(0.584%)
Tanzania	1987	1.025%(0.377%)	2.450%(0.787%)
Tanzania	1988	1.214%(0.396%)	3.006%(0.916%)
Tanzania	1989	1.427%(0.417%)	3.379%(0.936%)
Tanzania	1990	1.642%(0.441%)	3.550%(0.865%)
Tanzania	1991	1.829%(0.469%)	3.614%(0.761%)
Tanzania	1992	2.024%(0.500%)	3.660%(0.704%)
Tanzania	1993	2.473%(0.557%)	4.001%(0.775%)
Tanzania	1994	3.101%(0.600%)	4.436%(0.926%)
Tanzania	1995	3.718%(0.617%)	4.708%(1.067%)
Tanzania	1996	4.210%(0.626%)	4.724%(1.119%)
Tanzania	1997	4.512%(0.631%)	4.548%(1.066%)
Tanzania	1998	4.592%(0.625%)	4.247%(0.973%)
Tanzania	1999	4.419%(0.588%)	3.980%(0.945%)
Tanzania	2000	4.167%(0.526%)	4.024%(1.108%)
Uganda	1976	1.004%(0.236%)	1.666%(0.429%)
Uganda	1977	1.986%(0.493%)	3.438%(0.907%)
Uganda	1978	1.986%(0.493%)	3.438%(0.907%)
Uganda	1979	2.386%(0.641%)	4.367%(1.231%)
Uganda	1980	2.298%(0.614%)	4.262%(1.168%)
Uganda	1981	2.462%(0.577%)	4.413%(1.083%)
Uganda	1982	3.070%(0.563%)	5.012%(1.034%)
Uganda	1983	4.030%(0.583%)	5.806%(1.031%)
Uganda	1984	5.227%(0.642%)	6.881%(1.083%)
Uganda	1985	6.581%(0.732%)	8.375%(1.194%)

continued on the next page

Country	Year	Prevalence Age Cohort 25-35 in Final Year	Prevalence Age Cohort 35-45 in Final Year
Uganda	1986	8.049%(0.847%)	10.123%(1.351%)
Uganda	1987	9.526%(0.978%)	11.782%(1.536%)
Uganda	1988	10.784%(1.091%)	13.351%(1.730%)
Uganda	1989	11.536%(1.158%)	14.548%(1.904%)
Uganda	1990	11.398%(1.142%)	14.711%(1.989%)
Uganda	1991	10.431%(1.042%)	13.674%(1.927%)
Uganda	1992	9.374%(0.932%)	12.311%(1.747%)
Uganda	1993	9.284%(0.895%)	11.738%(1.633%)
Uganda	1994	10.890%(0.983%)	12.965%(1.677%)
Uganda	1995	13.668%(1.177%)	15.928%(1.909%)
Uganda	1996	16.294%(1.418%)	19.622%(2.353%)
Uganda	1997	18.015%(1.676%)	23.675%(3.077%)
Zambia	1976	0.699%(0.205%)	0.531%(0.269%)
Zambia	1977	1.482%(0.438%)	1.166%(0.576%)
Zambia	1978	1.482%(0.438%)	1.166%(0.576%)
Zambia	1979	1.973%(0.592%)	2.037%(0.872%)
Zambia	1980	1.971%(0.560%)	2.425%(0.905%)
Zambia	1981	1.835%(0.503%)	2.608%(0.873%)
Zambia	1982	1.669%(0.456%)	2.779%(0.830%)
Zambia	1983	1.590%(0.446%)	3.116%(0.820%)
Zambia	1984	1.834%(0.485%)	3.997%(0.874%)
Zambia	1985	2.719%(0.568%)	5.872%(1.001%)
Zambia	1986	4.295%(0.666%)	8.466%(1.176%)
Zambia	1987	6.373%(0.772%)	11.166%(1.376%)
Zambia	1988	8.725%(0.879%)	13.495%(1.547%)
Zambia	1989	10.971%(0.979%)	15.117%(1.664%)
Zambia	1990	12.639%(1.056%)	15.965%(1.743%)
Zambia	1991	13.254%(1.093%)	15.964%(1.796%)
Zambia	1992	13.282%(1.134%)	15.778%(1.891%)
Zambia	1993	13.551%(1.211%)	15.997%(2.020%)
Zambia	1994	14.597%(1.311%)	16.266%(2.099%)
Zambia	1995	16.310%(1.388%)	15.803%(2.053%)
Zambia	1996	18.053%(1.406%)	14.260%(1.861%)
Zambia	1997	20.208%(1.568%)	12.567%(1.656%)
Zimbabwe	1979	0.172%(0.158%)	0.195%(0.133%)
Zimbabwe	1980	0.314%(0.302%)	0.315%(0.199%)
Zimbabwe	1981	0.314%(0.302%)	0.315%(0.199%)
Zimbabwe	1982	0.415%(0.302%)	0.457%(0.247%)
Zimbabwe	1983	0.473%(0.252%)	0.532%(0.256%)
Zimbabwe	1984	0.540%(0.276%)	0.604%(0.290%)
Zimbabwe	1985	0.674%(0.306%)	1.030%(0.336%)
Zimbabwe	1986	0.932%(0.297%)	1.767%(0.388%)
Zimbabwe	1987	1.461%(0.285%)	2.982%(0.453%)
Zimbabwe	1988	2.280%(0.287%)	4.747%(0.547%)
Zimbabwe	1989	3.389%(0.303%)	7.008%(0.690%)
Zimbabwe	1990	4.841%(0.356%)	9.636%(0.907%)
Zimbabwe	1991	6.573%(0.475%)	12.562%(1.213%)
Zimbabwe	1992	8.443%(0.657%)	15.509%(1.563%)
Zimbabwe	1993	10.330%(0.889%)	18.287%(1.927%)

continued on the next page

Country	Year	Prevalence Age Cohort 25-35 in Final Year	Prevalence Age Cohort 35-45 in Final Year
Zimbabwe	1994	12.120%(1.137%)	20.698%(2.253%)
Zimbabwe	1995	13.535%(1.358%)	22.355%(2.495%)
Zimbabwe	1996	14.686%(1.586%)	23.540%(2.707%)
Zimbabwe	1997	15.387%(1.801%)	23.888%(2.802%)
Zimbabwe	1998	15.639%(1.945%)	23.374%(2.713%)
Zimbabwe	1999	15.351%(1.922%)	22.182%(2.453%)
Zimbabwe	2000	14.335%(1.731%)	20.229%(2.125%)
Zimbabwe	2001	12.898%(1.529%)	17.901%(1.857%)

Notes: These are estimates of HIV rates in the sample countries. The rates are estimated for two age cohorts: the group aged 25-35 in the final year (which varies across countries) and the group aged 35-45 in that final year. Standard errors on the estimate are in parentheses.

Table 4. ***Mortality-Based Estimates and DHS Population Testing***

Dependent Variable: Mortality-Based Prevalence Estimate, by Group				
	(1)	(2)	(3)	(4)
	All	All	Women	Men
Explanatory Variables:				
DHS Testing Estimate	.7111***	.5972***	.5372***	.6562***
	(.069)	(.027)	(.043)	(.022)
Age Group 20-25	.0158	.0194	.0173	.0238*
	(.011)	(.012)	(.015)	(.012)
Age Group 25-30	−.009	.0002	−.0084	.0114
	(.013)	(.01)	(.016)	(.013)
Age Group 30-35	−.0167**	−.0044	−.0172	.009
	(.006)	(.01)	(.013)	(.015)
Age Group 35-40	−.0178*	−.0057	−.0196**	.0077
	(.01)	(.005)	(.008)	(.011)
Age Group 40-45	−.0238	−.014	−.0206	−.0098
	(.013)	(.009)	(.014)	(.007)
Female	−.0056	−.0024		
	(.004)	(.003)		
constant	.022***	.023***	.034***	.01
	(.006)	(.005)	(.006)	(.009)
Country FE	NO	YES	YES	YES
Number of Observations	72	72	36	36
R^2	.88	.93	.94	.95

standard errors in parentheses, clustered by country

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes: This table shows the relationship between the mortality-based estimates of HIV prevalence and the DHS testing data. An observation is an age group-gender (age groups 15-20, 20-25, 25-30, 30-35, 35-40).

Table 5. ***Prevalence Estimates, Control Countries***

Country	Year	Prevalence in Age Cohort 25-45 in Final Year
Brazil	1982	0.005% (0.011%)
Brazil	1983	0.009% (0.019%)
Brazil	1984	0.011% (0.021%)
Brazil	1985	0.012% (0.021%)
Brazil	1986	0.011% (0.020%)
Brazil	1987	0.012% (0.019%)
Brazil	1988	0.015% (0.019%)
Brazil	1989	0.017% (0.017%)
Brazil	1990	0.025% (0.021%)
Brazil	1991	0.039% (0.031%)
Brazil	1992	0.055% (0.044%)
Brazil	1993	0.073% (0.060%)
Brazil	1994	0.092% (0.078%)
Brazil	1995	0.108% (0.093%)
Brazil	1996	0.119% (0.105%)
Philippines	1984	0.047% (0.036%)
Philippines	1985	0.096% (0.075%)
Philippines	1986	0.132% (0.100%)
Philippines	1987	0.152% (0.111%)
Philippines	1988	0.156% (0.108%)
Philippines	1989	0.146% (0.101%)
Philippines	1990	0.134% (0.093%)
Philippines	1991	0.122% (0.083%)
Philippines	1992	0.108% (0.073%)
Philippines	1993	0.100% (0.072%)
Philippines	1994	0.100% (0.079%)
Philippines	1995	0.107% (0.093%)
Philippines	1996	0.117% (0.109%)
Philippines	1997	0.124% (0.118%)
Philippines	1998	0.129% (0.124%)

Notes: This table reports prevalence estimates in two "control" countries, for the cohort aged 25-45 in the final year of the sample (either 1998 in the Philippines or 1996 in Brazil). These estimates are generated using exactly the same procedure as the estimates for Africa, including the use of sibling histories. Standard errors are in parentheses.

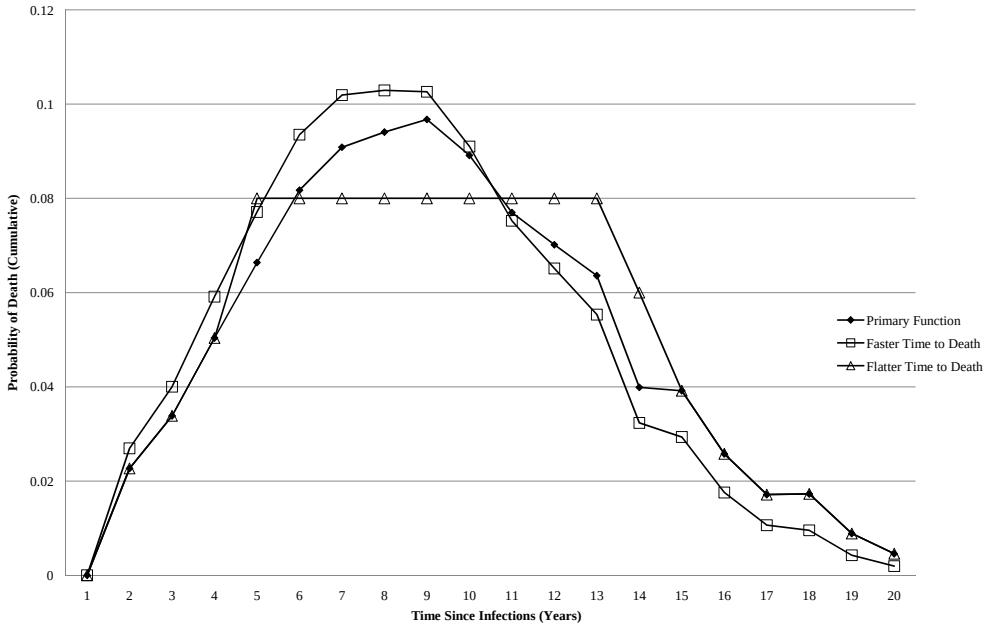
Table 6. ***Robustness of Estimates to Assumptions Generating Data***

Country	Primary Results	Flatter Death	Faster Death	Less Smooth	More Smooth
Burkina Faso	3.246%	3.250%	2.975%	3.315%	3.857%
Cameroon	4.986%	5.161%	4.522%	4.978%	4.943%
Ethiopia	3.650%	3.550%	3.140%	3.798%	4.105%
Kenya	6.092%	5.576%	5.956%	5.836%	7.070%
Malawi	11.116%	10.368%	10.838%	10.644%	10.450%
Mali	3.655%	3.780%	3.652%	3.720%	4.290%
Mozambique	6.469%	6.672%	5.961%	6.442%	6.512%
Namibia	6.770%	7.879%	6.066%	6.752%	8.258%
Tanzania	4.096%	4.307%	3.735%	4.114%	4.238%
Uganda	20.845%	17.686%	19.279%	19.191%	23.604%
Zambia	16.387%	14.843%	15.454%	15.735%	15.860%
Zimbabwe	15.400%	15.452%	14.168%	15.389%	15.320%

Notes: This table reports prevalence estimates in the final year of the sample, for the age cohort 25-45 in that year, for each country under varying assumptions generating incidence.

Appendix Figures and Tables

Appendix Figure 1:
Cumulative Probability of Death from HIV, Alternative Assumptions



Notes: This figure presents the yearly probability of death under the primary time-to-death function and the two variations explored in the robustness section.

Appendix A: Details of Smoothing in Estimation

This appendix provides more details on why, when estimating prevalence, it is necessary to impose a smoothing restriction.

Recall that we begin with the concept of solving a system of equations, as below (for simplicity, this appendix notation abstracts away from deaths from non-HIV causes):

$$\begin{aligned}
\mu_{i,t} &= d_4 b_{i-4,t-4} + d_5 b_{i-5,t-5} + ... + d_{20} b_{i-20,t-20} && (7)\\
\mu_{i-1,t-1} &= d_4 b_{i-5,t-5} + d_5 b_{i-6,t-6} + ... + d_{20} b_{i-21,t-21} && (8)\\
&\dots && (9)\\
\mu_{i-x,t-x} &= d_4 b_{i-x-4,t-x-4} && (10)
\end{aligned}$$

With noiseless data on mortality rates solving the system of equations for the **b** vector will work perfectly. However, noisy data on mortality will produce inappropriately high volatility in predicted infection rates. To see why, consider a simple example: a world with just two years of infections and two years of data on mortality.

$$\begin{aligned}
\mu_{i,t} &= d_4 b_{i-4,t-4} + d_5 b_{i-5,t-5}\\
\mu_{i-1,t-1} &= d_4 b_{i-5,t-5}
\end{aligned}$$

Assume that the actual values in the **b** vector are $b_{i-4,t-4} = b_{i-5,t-5} = \psi$ and that $d_4 = d_5 = 0.1$. If we observed the exact mortality, we would see $\mu_{i,t} = 0.2\psi$ and $\mu_{i-1,t-1} = 0.1\psi$ and would conclude that $b_{i-4,t-4} = b_{i-5,t-5} = \psi$. If the mortality rates are observed with noise, however, we could see $\mu_{i,t} = 0.2\psi + \varepsilon$ and $\mu_{i-1,t-1} = 0.1\psi + \eta$. Solving the system of equations will yield

$$\begin{aligned}
b_{i-5,t-5} &= \psi + \frac{\eta}{.1}\\
b_{i-4,t-4} &= \psi + \frac{\varepsilon - \eta}{.1}
\end{aligned}$$

If we assume that $\psi = 0.011$, $\eta = 0.002$, and $\varepsilon = -0.002$, then instead of estimating that $b_{i-4,t-4} = b_{i-5,t-5} = 0.011$, we will estimate that $b_{i-5,t-5} = 0.031$ and $b_{i-4,t-4} = -0.029$. Intuitively, a small amount of noise is translated into very large deviations between periods because of the very limited additional information in each of the simultaneous equations. The smoothing restriction will address this issue. By penalizing the maximization when any year is far from the two years surrounding it, this does not allow the system to diverge wildly across years (this is similar to an approach in Murphy and Welch, 1990).

Appendix B: Details on the Reliability of Sibling Mortality Histories

This appendix provides some additional defense of the use of sibling mortality histories by providing direct evidence on the comparison between these sibling histories and official mortality statistics, in a few cases where both are available. I also outline the theoretical issues associated with these data.

In general, there are two primary concerns with using sibling mortality histories. The first is a simple concern about underreporting, since individuals may forget siblings who have died or may not know that their siblings have died. In addition, even with perfect reporting sibling mortality data have the potential for either upward or downward bias. Sibling reports may overstate the overall death rate since they do not take into account the reporters themselves, who are alive. They may understate the overall death rate since families with low mortality profiles will be misrepresented, and families in which all siblings are dead will not be observed at all (Gadikou and King, 2006). Remarkably, Trussell and Rodriguez (1990) prove that with a random sample of the population *if mortality is uncorrelated with sibship size* these factors will cancel out and the report is unbiased. Although the assumption about independence sibship size is probably unrealistic when considering individual mortality in childhood, in terms of thinking about adult mortality from HIV it may be more appropriate (indeed, in the samples here we do not see a strong correlation).

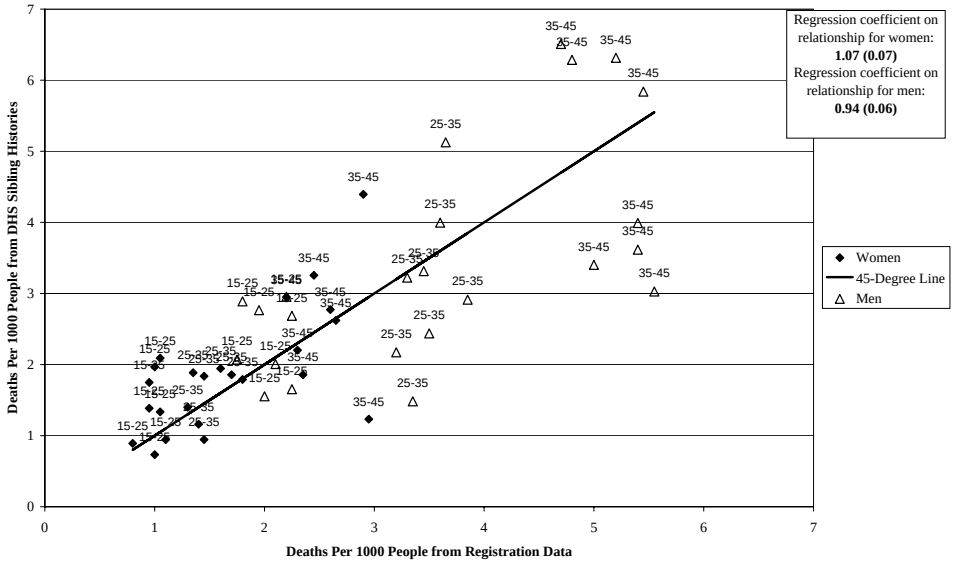
Rather than relying simply on the result in Trussell and Rodriguez (1990), however, I provide some direct evidence on the match between sibling mortality histories and official death data, using data from two

countries in which I have both sources: the Philippines and Zimbabwe. For the Philippines, I match death rates from sibling mortality histories collected by the DHS to official death rates reported in the United Nations Demographic Yearbook. The data come from 1980, 1983, 1984, 1986 and 1988-1991, and I generate mortality rates for each year by ten year age groups (15-25, 25-35 and 35-45) and gender. The data are shown in Figure B.1, along with the 45 degree line. On average, the death rates from sibling histories are very close to the official rates. A regression of the sibling history death rate on the official death rate for women yields a coefficient of 1.07 (standard error of 0.07) and, for men, a coefficient of 0.94 (standard error of 0.06). In neither case are these statistically distinguishable from a coefficient of one.

For Zimbabwe, data based on DHS sibling histories are compared with registration data from Feeney (2001), for 1982, 1986, 1990, 1991, 1992, and 1995 (although these are official registration data, it is worth noting that in no year do they cover more than 60% of the population). Figure B.2 replicates Figure B.1. Again, the match between the two datasets is quite strong. For women, the coefficient on the relationship is 1.02, not distinguishable from 1. For men, the coefficient is slightly smaller (at 0.81). This could reflect larger problems with under-reporting when women are asked about their brother, although the absolute magnitude of the errors is still very small. In general, Figures 4 and 5 show strong evidence (although for a limited set of countries) that sibling histories generate accurate mortality rates, similar to what we would see if we had access to official death registration data. This is true even in Zimbabwe, where mortality is heavily affected by HIV.

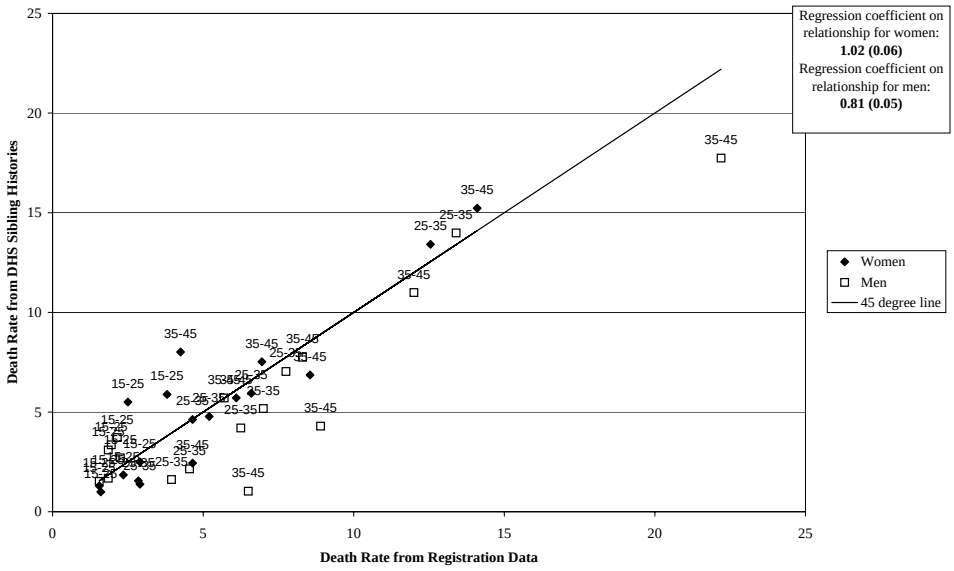
The evidence in Figures B.1 and B.2 suggests that the naive estimator of death rates from these data (sibling deaths over a period divided by siblings alive at the start) provides a good fit to registration data. It is also worth noting that similar results are obtained if I use the adjustments for bias worked out in Gakidou and King (2006).

Figure B.1:
Death Rates, DHS and Registration Data, Phlippines



Notes: This figure shows the death rate from sibling histories done in 1993 for age-gender groups (15-25, 25-35, 35-45) in the Philippines (years of data match: 1980, 1983,1984, 1986,1988,1989,1990,1991), graphed against matched mortality rates from the United Nations Demographic Yearbook Historical Supplement.

Figure B.2:
Death Rates, DHS and Registration Data, Zimbabwe



Notes: This figure shows the death rate from sibling histories for age-gender groups (15-25, 25-35, 35-45) in Zimbabwe in 1982, 1986, 1990, 1991, 1992 and 1995 graphed against matched mortality rates from registration data. Registration data is from Feeney (2001).