Estimates of Age-Specific Reductions in HIV Prevalence in Uganda: Bayesian Melding Estimation and Probabilistic Population Forecast with an HIV-enabled Cohort Component Projection Model

Samuel J. Clark, Jason Thomas and Le Bao*

Working Paper no. 106
Center for Statistics and the Social Sciences
University of Washington

August 30, 2010

*Department of Sociology, University of Washington, Seattle, WA, USA; Institute of Behavioral Science (IBS), University of Colorado at Boulder, CO, USA; MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa (Samuel J. Clark); Center for Demography and Ecology, University of Wisconsin-Madison, WI, USA (Jason Thomas); and Department of Statistics, University of Washington, WA, USA (Le Bao). Correspondence to Samuel J. Clark: **work@samclark.net**.

Abstract

We estimate age-specific HIV incidence and prevalence in Tanzania and Uganda in the late 1990s and forecast forward assuming no change in incidence. Comparisons between our forecasts of HIV prevalence and direct measures from the HIV/AIDS Indicator and Demographic and Health Surveys in the mid-2000s provide an age-specific measure of changes in HIV prevalence. In Tanzania our forecast accurately predicts age-specific HIV prevalence, suggesting little change in HIV incidence in Tanzania over the intervening decade. In Uganda our forecasts significantly overstate HIV prevalence. The age pattern of our forecast errors reflects the age-specific reductions in HIV prevalence and incidence in Uganda. Our estimates and forecasts are produced using an HIV-enabled cohort component model of population projection first proposed by Heuveline (2003). We refine that model (Thomas and Clark, 2008) and implement the Bayesian melding with IMIS estimation method (Raftery and Bao, 2010). This method allows us to estimate the parameters of the Heuveline model with robust measures of uncertainty and to quantify uncertainty in the model outputs, e.g. forecasts. We validate the model and estimation procedure by comparing maximum likelihood estimates with our Bayesian estimates and by confirming the calibration of the model outputs.

Keywords: HIV, Forecast, Africa, Uganda, Tanzania, Incidence, Prevalence, Estimation, Cohort Component Model of Population Projection, Bayesian melding, IMIS.

Contents

1 Introduction

This work makes two main contributions. *The first is an empirical investigation of the history of the HIV epidemic in East Africa.* We replicate the work of Heuveline (2003, published in this journal) to estimate sex-age-specific HIV incidence and prevalence in Tanzania and Uganda in the mid-to-late 1990s using our modified version of his HIV-enabled cohort component model of population projection. Then assuming no change in incidence, we make a probabilistic projection (i.e. forecast) of those HIV-infected populations and compare the projected HIV prevalence with the empirical estimates from the HIV/AIDS Indicator and Demographic and Health Surveys about ten years later. *The second contribution is adaptation and implementation of the Bayesian melding with IMIS estimation procedure* (Poole and Raftery, 2000; Raftery and Bao, 2010) *to work with the HIV-enabled cohort component model of population projection.* This Bayesian method has important advantages compared to the maximum likelihood approach used by both Heuveline and ourselves in previous work (Thomas and Clark, 2008), including the ability to produce probability distributions of the estimated parameters and model outputs which can be used for inference and projection.

HIV affects both mortality (increases) and fertility (decreases) and consequently has important effects on population growth and the sex and age composition of a population (UNAIDS, 2009a). The fundamental process in an HIV epidemic is transmission of the virus – how and between whom. The details of the transmission dynamics determine who is infected and at what age, and this then determines, with a delay, who is sick and dying at a given age. A fuller understanding of the biological and behavioral determinants of transmission would give us the ability to design effective *prevention* interventions that target specific mechanisms, situations and people.

It follows then that the most valuable indicator of an HIV epidemic is *incidence*, the ratio of new cases to those at risk of infection (Bongaarts et al., 2008; Hallett et al., 2007). Beyond an understanding of the dynamics of the epidemic as a whole, and in order to design and monitor well-targeted, effective and affordable interventions, it is necessary to refine measures of incidence by at least *sex* and *age*. The problem is that HIV incidence is extremely difficult and expensive to measure because it involves long-term followup of a large number[1] of HIV negative people (see for example: Mbulaiteye et al., 2002; Wambura et al., 2007). There is a promising test for recency of HIV infection being developed and tested, but so far it is difficult to calibrate the results accurately (Parekh et al., 2002; McDougal et al., 2006; McWalter and Welte, 2009, 2010). This leaves only one widely applicable option to learn about HIV incidence: *mathematical modeling*.

Mathematical and computational models of HIV epidemics (see for example: Cassels et al., 2008; Anderson, 1988; Hallett et al., 2006; Hallett, T.B. and Żaba, B. et al., 2008; Hallett, T.B. and Singh, K. et al., 2008; Granich et al., 2009) represent populations and the mechanisms that transmit the HI virus from one (type of) person to another. Essentially they perform either or both of two tasks: to *estimate* parameter values or to *project* the population forward in time in order to make predictions or investigate different scenarios. 'Parameters'

[1] 'Large' because HIV infection is a rare event which necessitates a large number of observations to accumulate enough infections to measure incidence rates with precision.

in this sense are variables whose values govern the behavior of the model; incidence (or something closely related) is often a variable in these models. Used in *estimation* mode, the objective is to find values of the parameters that produce model outputs that match a set of empirical values. Because HIV prevalence is comparatively easy and cheap to measure, models are often fit or estimated to match prevalence. In *projection* mode the model outputs themselves are the quantities of interest.

We use a mathematical model of a population with HIV to do both estimation and projection. First we estimate the model parameters necessary for the modeled population to closely match the HIV prevalence in a variety of study populations in Tanzania, Uganda and Burundi (East Africa) in the early-to-mid 1990s. This provides us with the trend and age-pattern of HIV incidence from the beginning of the epidemic up to then that are necessary to create the age-patterns of prevalence observed in each study population. We then move to projection mode and hold HIV incidence constant in each sex-age group and project the populations of Tanzania and Uganda forward in time until we have new representative measures of HIV prevalence from the HIV/AIDS Indicator and Demographic and Health Surveys, and at that time we compare our projected HIV prevalence to the estimates from the surveys. *Most of the differences we observe are attributable to changes in HIV incidence during the intervening years*[2].

Estimates and outputs from models like ours often appear as single numbers without corresponding measures of uncertainty or precision, or if they do have these, they are constructed in an *ad hoc* fashion. We address this problem by employing the Bayesian melding with IMIS method (Poole and Raftery, 2000; Raftery and Bao, 2010) that has the ability to properly quantify uncertainty in the estimated parameters and all of the model outputs, including HIV incidence, prevalence, age structures, etc. We use the estimated probability distributions of parameters and model outputs to confirm that significant, age-specific changes occurred in HIV prevalence and incidence in Uganda during the late 1990s.

This paper is organized in an unconventional way. Following the sensible suggestions of Ehrenberg (1982), the next section presents *Results & Recommendations*, followed by *Methods & Data*, *Background & Significance* and finally a *Discussion*. In this order the reader moves to increasingly detailed treatment of the work starting with a clear understanding of where it is going. The aim is to present the material in a convenient organization that reflects how we *actually* read papers, keeping in mind that the bulk of readers are most interested in the key results and an overview of the method. The most casual reader can stop after the *Abstract*, readers who are interested but do not need the real details can stop before Section 4 *Methods & Data*, and readers who need all the details can read through the remaining sections.

[2]Changes in prevalence may also arise from longer survival times resulting from increases in antiretroviral therapy (ART) coverage. The World Health Organization estimates that ART coverage in 2007 is 33% in Uganda and 31% in the United Republic of Tanzania (World Health Organization, 2008). Increasing coverage over time results in an upward pressure on the levels of HIV prevalence, and thus our estimates of changes in HIV incidence should be treated as lower bounds since we assume a constant survival schedule for the infected population.

2 Results & Recommendations

The HIV-enabled cohort component model of population projection (HCCMPP) is described in detail by Thomas and Clark (2008) and below in Appendices A and B (adapted from that paper). The standard cohort component model of population projection (see for example: Bowley, 1924; Cannan, 1895; Whelpton, 1936; Leslie, 1945; Pritchett, 1891; Pearl and Reed, 1920; Dorn, 1950) was enhanced by Heuveline (2003) to include five additional states to accommodate the duration-specific stages of HIV infection: (i) uninfected, (ii) infected 0–4 years, (iii) infected 5–9 years, (iv) infected 10-14 years and (v) infected 15+ years . A time-sex-age-specific incidence profile moves people from the uninfected to the first (0–4 year) infected group. Once people are in the infected groups they face diminished odds of surviving as they move to the next infected group. Infected women also experience slightly reduced fertility rates and consequently produce fewer births. As people move through the HIV infected duration groups, the effects of HIV become more pronounced to reflect the intensifying nature of their infections.

We have created a Leslie matrix representation of this model that allows us to run the model easily and allow some additional formal manipulation (see Appendix A). We start with a base population count by sex and age, a set of underlying mortality and fertility rates (all from the UN) and a set of parameters for the HIV incidence profile, and we multiply the column vector containing our population by the Leslie matrix. We divide both time and age into five-year periods, so one multiplication moves the population forward in time and age by five years. To go twenty years forward we multiply the population vector times the four Leslie matrices that represent the corresponding four five-year periods. The result is a new column vector for each sex containing the age- and HIV status-specific counts of the population twenty years in the future. From this we can calculate HIV prevalence by dividing the total HIV+ population count in a given sex-age group by the total population count in that same sex-age group.

If the starting population, vital rate schedules and HIV incidence parameters are fixed, the HCCMPP can be used in the traditional way to *project* an HIV-infected population forward in time. Alternatively, the HCCMPP can be used to estimate the values of unknown parameters. When used to estimate, the general idea is to vary the unknown parameters until a set of values are found that create a population that matches some set of criteria.

To estimate the HIV incidence parameters, we start with a reasonable base population and vital rates for the early 1980s (from the UN) and project the population forward ten to fifteen years until the mid 1990s when HIV prevalence measures began to become available (for small populations within Burundi, Tanzania, and Uganda). We then calculate the predicted HIV prevalence from the model and compare it to what was actually measured, and we adjust the incidence parameters until we have a close match to sex- and age-specific prevalence.

Table 1. Empirical Coverage of the Bayesian Prediction Intervals for HIV Prevalence

	50% Prediction Interval			80% Prediction Interval			95% Prediction Interval		
HIV Incidence Trend	<25%	[25%, 75%]	75%<	<10%	[10%, 90%]	90%<	<2.5%	[2.5%, 97.5%]	97.5%<
fixed gamma curve									
% of Observations	10.7%	42.9%	46.4%	3.6%	75.0%	21.4%	3.6%	92.9%	3.6%
		(2.6)			(4.9)			(7.5)	
estimated gamma curve									
% of Observations	10.7%	10.7%	78.6%	3.6%	25.0%	71.4%	3.6%	50.0%	46.4%
		(1.7)			(3.0)			(4.2)	
estimated sex-specific gamma curves									
% of Observations	10.7%	10.7%	78.6%	3.6%	28.6%	67.9%	3.6%	53.6%	42.9%
		(1.8)			(3.3)			(4.8)	
non-parametric trend									
% of Observations	10.7%	10.7%	78.6%	3.6%	39.3%	57.1%	3.6%	57.1%	39.3%
		(2.0)			(3.7)			(5.4)	
sex-specific non-parametric trends									
% of Observations	10.7%	25.0%	64.3%	3.6%	67.9%	28.6%	3.6%	89.3%	7.1%
		(3.1)			(5.7)			(8.4)	

Notes: Results for Tanzania for various specifications of the HIV incidence trend. Numbers in parentheses are mean interval widths measured in percentage points. There are 28 observations.

We have two methods to do this: a maximum likelihood approach analogous to Heuveline (2003) (described in Appendix B) and the new Bayesian melding with IMIS technique described below in Section 4.3. Using either method we identify the most likely *set* of HIV incidence parameter values that together with our assumptions about the base population and vital rates produce the sex-age-specific HIV prevalence observed in the mid 1990s (or something very similar), and additionally measures of uncertainty around those point estimates. The Bayesian estimates require fewer assumptions about the data and model and yield better behaved distributions of estimated parameter values; e.g. the values stay within the natural range for bounded parameters and can be directly interpreted in a probabilistic framework.

To project an HIV-infected population forward in time, the model is run with known values for all the parameters. This produces a predicted population corresponding to the base population and the sequence of parameter values used to govern the dynamics of the population over the projection interval (it is possible to change parameter values as time goes by to reflect changing vital rates or HIV incidence). We use the model in this way to produce probabilistic forecasts of the populations of Tanzania and Uganda (DHS data with information on HIV prevalence for Burundi are not available at the time of writing). The forecasts consist of a large number of projections that in effect create a distribution of populations at some time in the future, and from this distribution we can make probabilistic statements about how likely a given future population is. The projections constituting the forecast are created using the predictive (posterior) distribution of parameter values from the Bayesian estimation method, with the result that the most likely sets of parameter values translate into the most likely set of populations (net of assumptions about trends in parameter values over the projection period).

2.1 Main Findings

2.1.1 Model Validation: Calibration and Predictive Performance for Tanzania

We validate our model by examining its predictive performance in Tanzania. We estimate the model using high quality data collected by small community-based studies in Tanzania, Uganda and Burundi in the mid 1990s (see the entries for Tanzania, Uganda and Burundi in Table B-1 in Appendix B). There was little change in the prevalence of HIV in Tanzania from the mid 1990s until the mid 2000s (UNAIDS, 2009a; Asamoah-Odei et al., 2004), and consequently if we use our estimated parameter values from the mid 1990s we should be able to forecast the sex-age distribution of HIV prevalence in the mid 2000s accurately and with reasonable confidence.

Sex-age-specific HIV prevalence measured by the HIV/AIDS Indicator and Demographic and Health Surveys in Tanzania in 2004 and 2007 serve as the targets for our forecast. To produce the forecast we use the best-fitting fixed gamma trend in overall HIV incidence and hold it constant from year fifteen of each projection (roughly the year 2000 in calendar time). We use our estimated distributions of HIV incidence by sex and age with no modifications for the duration of the forecast. The distribution of forecast HIV prevalence values is generated

by making multiple draws from the estimated joint parameter distribution and projecting the population forward for each of those with a constant overall incidence trend, see Figure 1. The result is a sex-age-specific distribution of HIV prevalence values at various times in the 2000s that we can use to compare with the empirical values measured by the surveys.

To assess the accuracy and calibration of the forecast we take the predicted distributions for sex-age-specific HIV prevalence for Tanzania[3] and calculate the quantiles of the 50%, 80%, and 95% credible intervals and compare these to the corresponding observations from the HIV/AIDS Indicator and Demographic and Health Surveys. Table 1 displays these 'coverage' results. There is one row for each of the overall HIV incidence trends that we tried and three sets of columns for the 50%, 80% and 95% credible intervals. Each of these contains the percent of the empirical observations that fall below the lower limit, within the central interval and above the upper limit. Reading the first row of the table we find that 11% of the observations fall below the 25^{th} percentile, 43% fall between the 25^{th} and 75^{th} percentiles and 46% above the 75^{th} percentile, etc.

The forecast using the fixed gamma for the overall trend in HIV incidence clearly produces the best calibrated results (the observed coverage comes closest to what we expect), and the calibration is acceptable. 92.9% of observations fall within the 95% credible interval with an even 3.6% below and above. Calibration deteriorates slightly as the credible intervals shrink, and there is a slight tendency to understate prevalence as indicated by the fact that more observations fall above the prediction intervals than expected. Altogether the calibration results for Tanzania indicate that the model is reasonably accurate and representing uncertainty in a way that corresponds to empirical observation.

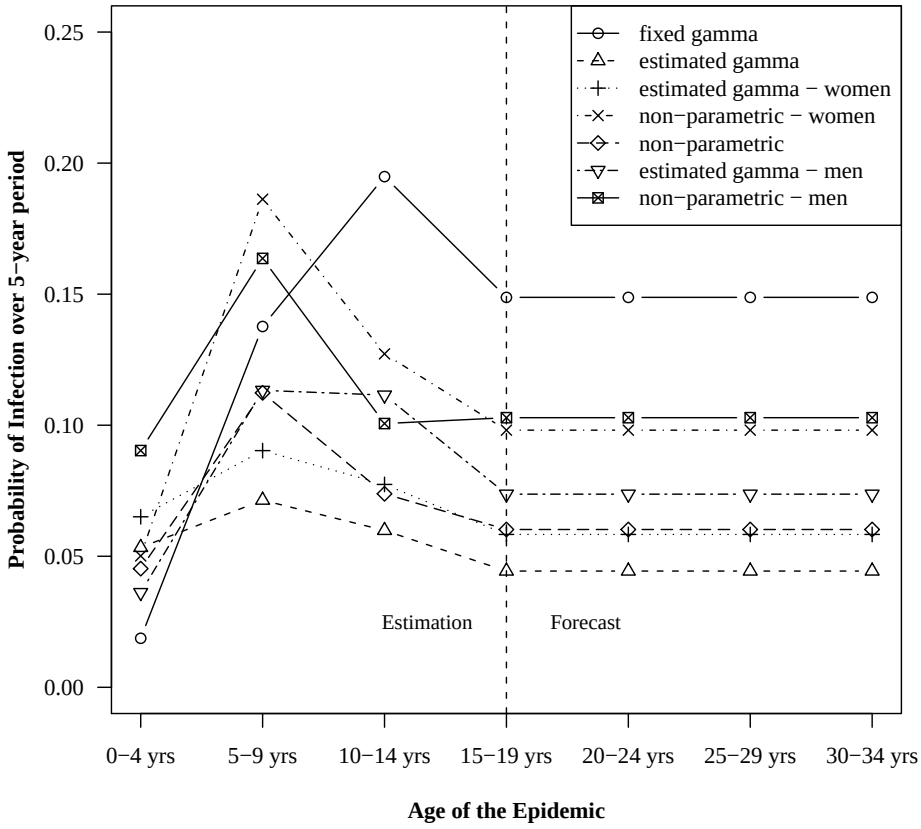
2.1.2 Parameter Estimates: Maximum Likelihood and Bayesian Melding

Estimates of the HIV incidence parameters are shown in Figures 1 and 2. Figure 1 displays a number of HIV incidence trends that we tried, including several non-parametric estimated curves. We use Bayes factors to identify the trend that simultaneously fits the data well and is parsimonious (see below in Section 3.1). Ironically the best fitting trend (see below in Section 3.2) is the fixed (i.e. *not* estimated) gamma curve originally proposed by Heuveline (2003), displayed with open circles. This trend controls the level of incidence with time for each population included in the estimation procedure, and because different populations have different overall levels of incidence, there is another parameter (not shown) that multiplies this trend to allow each population to have its own baseline level of incidence over time – the figure shows the specific curve for Rakai in Uganda, and there are curves with the same shape but at different levels for each of the other populations. Figure 1 reveals that incidence rises sharply in the first fifteen years of the epidemic and then falls slightly in order to produce the prevalence levels observed in the mid 1990s.

Figure 2 displays the estimated sex-age-specific relative HIV incidence rate ratios. These describe how the risk of acquiring HIV changes with sex and age. This set of incidence profiles is estimated jointly using all of the populations in the estimation procedure. Results

[3]These predicted distributions are specific to the years when the AIDS Indicator and Demographic and Health Survey data were collected.

Figure 1. Example HIV Incidence Trends: Rakai, Uganda



Notes: The level of incidence corresponds to the sex-specific age group 25-29 years when separate trends are used for women and men, and for women aged 25-29 years when a single trend is used for both groups. Values to the left of the vertical line are used to estimate the HCCMPP parameters, and values to the right are used to make forecasts to compare with data from the DHS surveys.

from the maximum likelihood (ML) and Bayesian melding (BM) estimation methods are very similar to each other and to Heuveline's 2003 estimates. Both sets of estimates for women, top panel of Figure 2, show that the risk of infection increases significantly from the 15–19 age group to the 20–24 age group, with the latter experiencing the highest level of incidence. The risk of infection then declines until reaching a fairly stable level after age 35. There are also very few differences between the ML and BM estimates for men, shown in the bottom panel of Figure 2. With either approach, the estimated risk of infection for men is relatively low among those aged 15-19 and clearly increases among the next two older age groups. Uncertainty makes it difficult to identify differences in the risk of infection among men between the ages of 25 and 49, but the ML and BM results seem to suggest that men

in their fifties experience a lower risk of infection than men aged 25-34. For both women and men, the ML and BM intervals around the point estimates increase with age, which is expected given the increasingly smaller number of observations at older ages.

While there is general agreement between the ML and BM estimates, it is worth noting a difference in their measures of uncertainty. Some of the ML confidence intervals include *negative* values, suggesting negative HIV incidence rates are possible. More specifically, the confidence intervals for women aged 50-54 and for men in their fifties stretch below zero. This problem also occurs when estimating other HCCMPP parameters with bounded parameter ranges, namely the fertility impairment parameters (results not shown). In contrast, all of the BM credible intervals are within the appropriate range of each parameter, a significant advantage of this approach, especially when draws from the parameter distribution are used to generate forecasts.

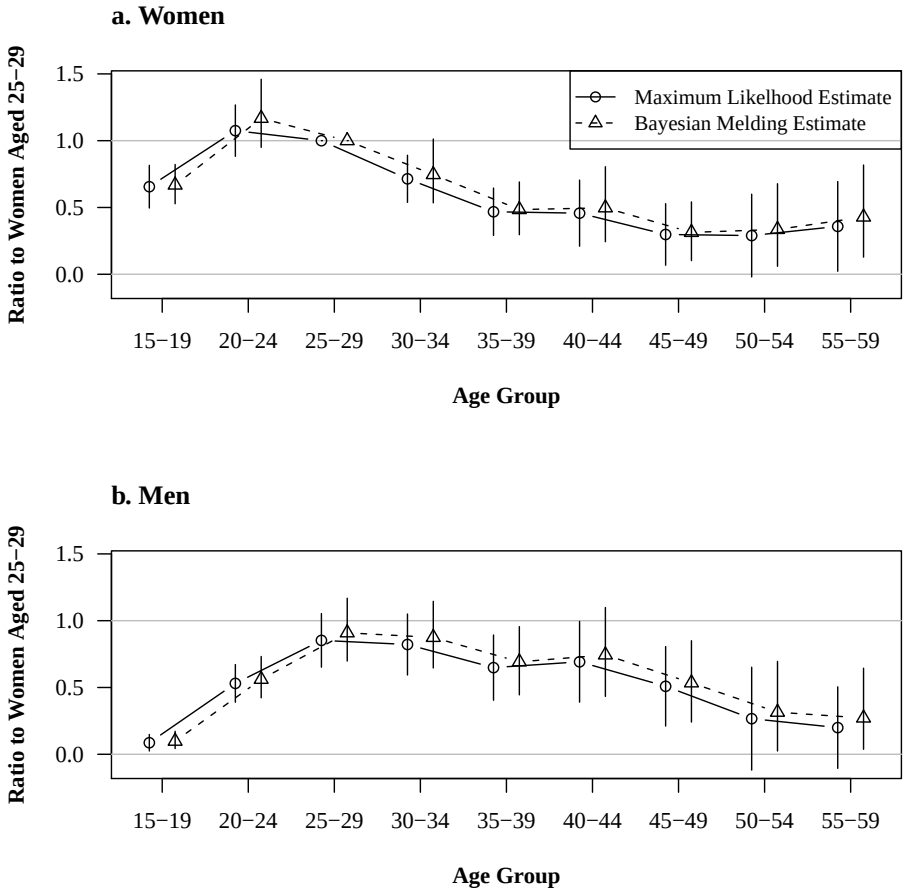
2.1.3 Forecasts: Probabilistic Projection of HIV Prevalence

Figure 3 displays the the forecast errors for Tanzania 2004 and 2007 and Uganda 2004, again using the best-fit fixed gamma trend in overall HIV prevalence with a constant value for years after 2000. Each plot contains the distribution of forecast errors by age group summarized with a boxplot. If the forecasts were perfect these boxplots would describe compact distributions centered at zero. Each forecast error is the residual between the observed and forecast values (observed – forecast). The distributions arise because there is a distribution of predicted values for each sex-age category. Our forecasts take into account uncertainty in HIV incidence but not underlying vital rates, and consequently we expect that uncertainty will be slightly underestimated.

Tanzania. From the previous section we already know that the forecast for Tanzania is reasonable. This is reflected in the left two columns of plots in Figure 3 in which the boxplot for every age group is centered near zero with comparatively short tails. The only systematic deviations from zero are in the age range 30–44 for women and ages 40–44 (2004) and 35–39 (2007) for men. For those ages the forecast appears to be slightly too low. Overall the forecast errors for Tanzania are small – a few percentage points – and the error distributions contain small variation and are centered close to zero.

Uganda. The forecast errors for Uganda clearly reveal the extent and age-pattern of the decline in HIV incidence over the intervening decade, rightmost column of plots in Figure 3. Compared to Tanzania, the forecast error distributions are much more variable and deviate from zero in a systematic age-dependent way. The greater variability likely corresponds to the greater geographic variability in the Ugandan data sources used to estimate the model parameters, including four different sites (Fort Portal, Gulu, Masaka, and Rakai) in both rural and urban areas. For women the 99^{th} percentile of the error distributions does not include zero until age 30. For younger ages the forecast errors are very significantly negative (the forecast overstates prevalence) with a clear trough in the 25–29 age group at a median error of $\sim$ –22%. This trough relaxes slowly through the older age groups toward a constant median error of about –5% in the age range 45–59. *This striking age-pattern indicates that HIV incidence fell most dramatically in the 15–24 age group (reflected in the later and older*

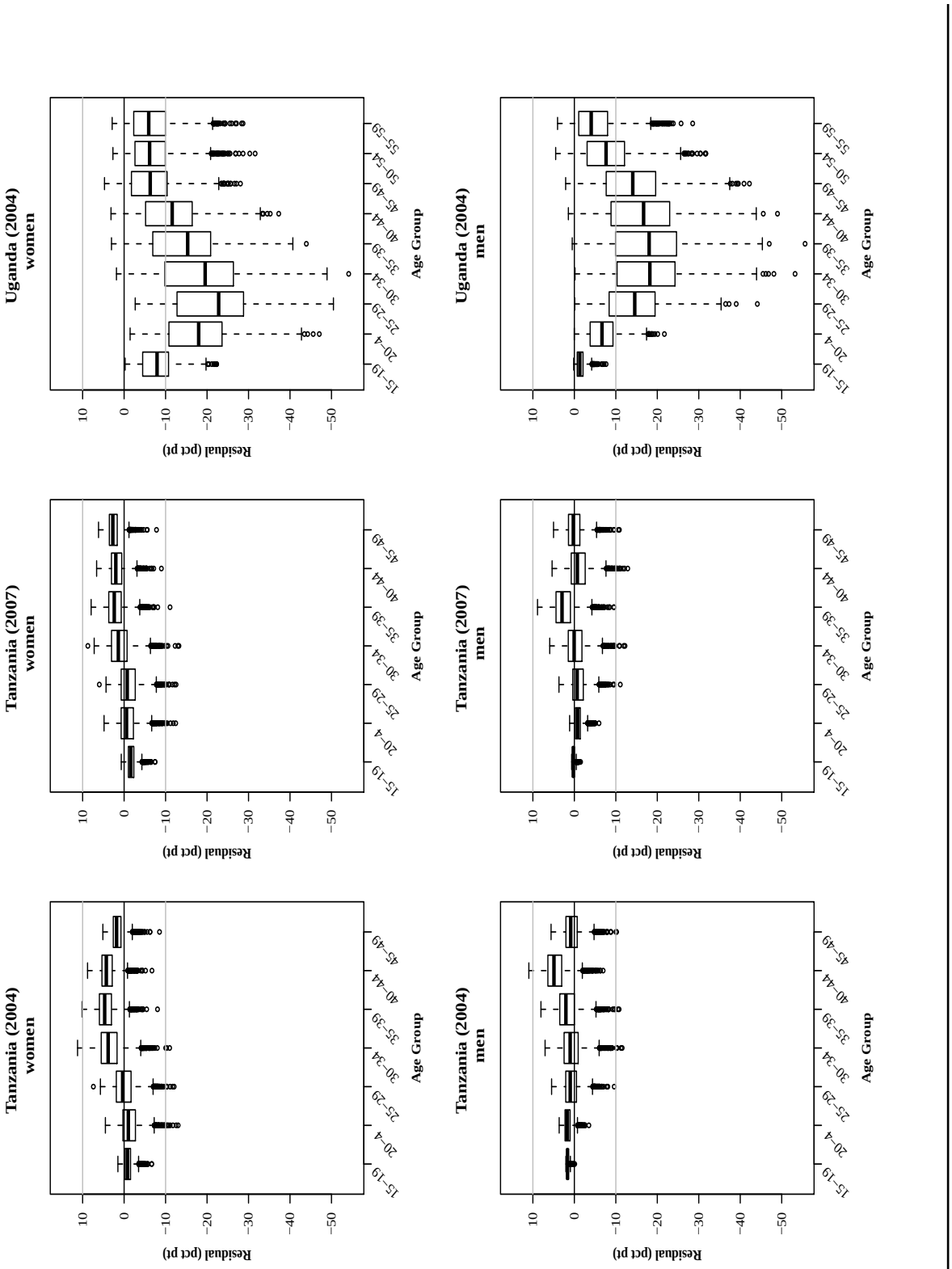
Figure 2. Estimated Age Schedules of HIV Incidence



Notes: These estimates are obtained using the fixed gamma curve incidence trend. The incidence of women aged 25-29 is given by the value of the incidence trend multiplied by a population-specific scale factor. All other sex-age categories in this figure are referenced to women aged 25-29 in a straightforward multiplicative sense, i.e. men aged 20-24 experience HIV incidence a little over 50% as great as women aged 25-29. Consequently this figure displays 'relative' HIV incidence by sex and age. The vertical lines running through the point estimates cover the 95% confidence intervals for the maximum likelihood results and the 95% credible intervals for the Bayesian results.

drop in prevalence in the 25–34 age group) by perhaps more than 20%. Further, there must have been significant reductions in incidence at all other ages, particularly $\sim$ 25–34.

Figure 3. HIV Prevalence Forecast Error Distributions



For Ugandan men the situation is similar, but the magnitude of the errors is slightly less and the age-pattern is different. The trough for men is wider covering about ages 30–44, but not quite as low, reaching a minimum of $\sim$ –19%. The trough also begins to develop at older ages, only showing strong deviation from zero in the 20–24 age group. Similar to women, the 99^{th} percentile of the error distributions does not include zero until age 35. The age-pattern of deviations in the male errors indicate a reduction in incidence over a broad range of ages from roughly 20–29 to 40–49, with the largest reductions over roughly ages 20–44.

2.2 Conclusions

Using some of the early measures of HIV prevalence from community-based studies in Tanzania, Uganda and Burundi during the early to mid 1990s, we estimate the age profile of HIV incidence that is consistent with underlying vital rates and the observed age pattern of HIV prevalence. We apply the new Bayesian melding with IMIS estimation procedure to 'fit' the HIV enabled cohort component model of population projection created by Heuveline (2003). Our results corroborate both his and our own earlier work using a maximum likelihood estimation procedure. The age profile of incidence is younger and more focused for women with peaks in the 20-24 year age group for women and 25-29 year age group for men.

The Bayesian estimation framework provides three useful advantages compared to maximum likelihood: (i) the ability to constrain parameter values in sensible ways that ensure parameters never take nonsensical values, and the ability to quantify uncertainty in (ii) estimated parameter values and (iii) model outputs in a statistically valid way that can be interpreted and manipulated in a fully probabilistic framework. In comparison to our maximum likelihood results, the predictive distributions of parameter values do not contain invalid (i.e. negative) values when their range is bounded below by zero. The predictive distributions for the estimated parameters are also well behaved and communicate levels of uncertainty that are reasonable and change as one would expect with respect to sample size, etc.

Most important to us, however, is the opportunity to produce probabilistic projections – true forecasts – of the HIV-affected populations. This allows us to validate our model in one more way by comparing (probabilistic) forecasts of HIV prevalence with empirical measures of prevalence in Tanzania. There was little change in HIV incidence in Tanzania between the mid 1990s and early 2000s, and we are able to predict with reasonably calibrated accuracy age-specific HIV prevalence in Tanzania in the early 2000s by forecasting forward with no change in our estimated HIV incidence pattern.

We use one further advantage of the Bayesian framework to compare models with different specifications of the trend in HIV incidence. The Bayesian framework allows us to use Bayes factors to compare the models and determine that the simple fixed gamma curve originally specified by Heuveline produces better forecasts compared to a variety of more flexible specifications with more parameters – the Bayes factor takes into account the number of parameters, effectively penalizing models with larger numbers of parameters (degrees of

freedom).

Finally, we use the probabilistic forecast of HIV prevalence for Uganda to characterize the age-pattern of reductions in prevalence resulting from the well-documented declines in HIV incidence that took place in Uganda between the early 1990s and mid 2000s. This age pattern of change in HIV prevalence reflects the earlier and younger changes in HIV incidence that were required to reduce the HIV+ fraction of the population.

2.3 Recommendations

1. Because HIV incidence is such an important indicator of an HIV epidemic, and because it is so difficult to measure HIV incidence empirically; epidemiologists, demographers and statisticians should prioritize further development of mathematical models and statistical procedures that allow us to estimate HIV incidence with uncertainty. To be of practical use to decision makers in the small areas where interventions are implemented and evaluated, these techniques should attempt to provide estimates of incidence by time, sex and age.

2. Given the success of Bayesian melding with IMIS applied to the 30+ parameter HCCMPP and the inherent advantages of the Bayesian framework, epidemiologists and demographers should consider applying this and similar procedures to other models and estimation procedures common to their disciplines.

3. We have successfully produced probabilistic forecasts of HIV epidemics taking into account uncertainty in HIV incidence. Using a similar Bayesian framework, future work on both HCCMPP and regular non-HIV CCMPP should incorporate uncertainty in vital rates and migration to produce probabilistic forecasts that take into account all major sources of uncertainty.

4. The Bayesian framework gives us the ability to conduct Bayesian model comparison using Bayes factors. This ability should be used to investigate the effects of interventions by comparing models that do and do not model the intervention. Bayesian model comparison will tell us if the intervention model fits the data better than the non-intervention model, and if so, the parameter estimates and model outputs will tell us what the effects are and *how* effective the intervention is.

3 Detailed Findings

3.1 Bayesian Melding with IMIS

Bayesian melding (BM) has been used in previous analyses involving deterministic models of population dynamics and HIV/AIDS that include less than five parameter inputs (Poole and Raftery, 2000; Alkema et al., 2007). A key finding in this paper is the successful implementation of BM in a relatively high dimension parameter space using the incremental mixture

importance sampling (IMIS) algorithm introduced by Raftery and Bao (2010). We are able to perform statistical inference and to make probabilistic projections using models that range from the simplest with 29 parameters up to the most complicated with 36 parameters. The IMIS algorithm proved to be much more efficient than the sampling importance resampling (SIR) technique (Rubin, 1987, 1988) that has been used in previous work to implement BM (Poole and Raftery, 2000; Alkema et al., 2007, e.g)[4]. It is also worth noting that the BM parameter estimates obtained via the IMIS algorithm are very similar to the maximum likelihood estimates, with the exception that the BM estimates of uncertainty are all within the natural bounds of the constrained parameters, whereas the maximum likelihood estimates of uncertainty include negative values for HIV incidence and other nonsensical results.

Having been able to successfully implement BM with various specifications of the model, we are left with the task of choosing among the different models that are distinguished by the trend in HIV incidence. Palloni (1996) pointed out that in a demographic model with HIV/AIDS the force of infection that produces the current level of prevalence should depend on the past level of prevalence, and thus the trend in HIV incidence is endogenous. To make this problem tractable with HCCMPP, Heuveline (2003) assumes an incidence trend based on a gamma curve, a strategy also used in previous models of HIV/AIDS epidemics (e.g., Chin and Lwanga, 1991; Salomon and Murray, 2001), and treats it as a fixed model input. While the gamma curve may yield a plausible trend, there is at least some uncertainty around this part of the model. In our analysis, we relax the assumption of a fixed gamma curve by estimating the trend in HIV incidence and allowing it to vary by sex and in functional form (see Section 4 for more details about the implementation and estimation). The estimated trends are discussed in the next section, and here we focus on the comparison of the following five models included in the analysis: (i) fixed gamma curve, (ii) estimated gamma curve, (iii) sex-specific estimated gamma curves, (iv) non-parametric curve, and (v) sex-specific non-parametric curves. Given the seemingly large differences between the trends shown in Figure 1, it is natural to be concerned with the relative merit of each model. One standard criterion is how closely each model fits the data. A simple metric for assessing model fit is the sum of squared residuals. According to this measure, there are only slight differences across all of the models with the values ranging from a high of 0.225 for the HCCMPP with the fixed gamma curve to a low of 0.201 for the model with the sex-specific non-parametric trends. An alternative measure for comparing models is Bayes factor (Jeffreys, 1939; Kass and Raftery, 1995), which is easily calculated from the IMIS approach taken here (Raftery and Bao, 2010). The model comparisons based on Bayes factor favor the HCCMPP with the fixed gamma curve over all of the other models, but the evidence is fairly weak since all of the Bayes factors are less than 1.1 – generally values greater than 3 indicate important differences between the models being compared (Raftery, 1995). Although the evidence is weak, it is interesting to note that the Bayesian model comparison favors the simplest model with the fixed gamma curve, which is also the model that is the best at predicting future observations – another important criterion for evaluating the relative merit of different models.

[4]Posterior samples of size 3,000 typically include less than 100 unique points when SIR is used to implement BM, as opposed to posterior samples obtained from IMIS that generally include around 1,500 or more unique points. For more details about the relative efficiency of these two approaches see Raftery and Bao (2010).

3.2 Trends in HIV Incidence

The different trends in HIV incidence used with the HCCMPP to forecast HIV prevalence are shown in Figure 1. Each trend in the plot shows the estimated probabilities of infection over a five-year period for the reference group in Rakai, Uganda (only the posterior means are shown for the estimated trends). For the models where the incidence trend is the same for women and men, only the estimated probabilities for women aged 25-29 are shown[5]. In our specifications where the trend is sex-specific, separate curves are shown in the plot corresponding to the sex-specific reference group aged 25-29. The vertical line indicates the last period for which we use data to estimate the model parameters. HIV incidence is assumed to stabilize during this time period, 15–19 years into the epidemic, and the corresponding level of incidence is used to forecast subsequent levels of HIV prevalence.

The most striking feature in Figure 1 is how the trend from the fixed gamma curve, which was used in the original analysis (Heuveline, 2003), reaches a level that is much higher than the estimated trends during the period 15–19 years into the epidemic. Conversely, when the trend in HIV incidence is estimated, the level of incidence is higher during the initial period of the epidemic and the peak occurs earlier than the trend from the fixed gamma curve. A second finding is that when separate curves are estimated for men and women, the trends appear to be different, for both the gamma and non-parametric specifications. Among those aged 25–29, estimated incidence based on the non-parametric trend is higher for men during the first five years of the epidemic, with a cross-over in the subsequent projection period, and convergence during the period 15–19 years into the epidemic. This cross-over of the incidence trends for men and women is consistent with the HIV-1 hypothesis described by Gregson et al. (1997), which posits a pattern of HIV transmission in rural areas where men are typically infected first, perhaps while working in an urban center or town, and then infect their female partners at a later point in time. The sex-specific trends estimated using gamma curves, however, suggest that the trend for men and women follow the same pattern and that only the levels are different. A final note is that these differences and cross-overs are only suggestive since there is uncertainty around the point estimates shown in Figure 1.

4 Methods & Data

4.1 Data

Three compilations of data are used in this analysis, the first of which is taken from Heuveline (2003) who reviewed the epidemiological literature and compiled data on HIV-related outcomes from populations located in East Africa[6]. For the current analysis, we use only those data from Burundi, Tanzania, and Uganda to limit the geographic heterogeneity across

[5]In the models where a single HIV incidence trend is used, the shape of the curve will be the exact same for men and women, but the levels may be different.

[6]The specific locations are Fort Portal, Uganda; Gulu, Uganda; Masaka, Uganda; Mara, Tanzania; Mwanza, Tanzania; Bujumbura, Burundi; Mangochi, Malawi; Lusaka, Zambia; Mposhi, Zambia; and Mutasa, Zimbabwe.

the local epidemics. The types of outcomes include: HIV test results in a general-population sample; HIV test results in an ANC-patient sample; HIV test results in all or a sample of births from HIV+ mothers; HIV test results during a follow-up of an HIV- sample; and survival during a follow-up of HIV+ individuals. These data, which are used to estimate the HCCMPP parameters, were all collected before 1998 with the majority collected during the 1990s and a few from the late 1980s. The outcomes are differentiated by age and sex, and were collected in rural, semi-urban or urban locations. After calibrating the model with the data collected before 1998, we then use the HCCMPP to make forecasts of sex-age-specific HIV prevalence in Tanzania and Uganda, and compare the forecasts to the levels observed in the HIV/AIDS Indicator Surveys and Demographic and Health Surveys collected in 2004 and 2007 for Tanzania, and in 2004 for Uganda (neither source of data is available for Burundi). The third compilation of data is taken from the United Nations global demographic estimates United Nations (2007), which provides the basic model inputs needed to make the forecasts. The HCCMPP requires an initial age distribution for women and men as well as sex-age-specific rates of fertility and mortality for the uninfected populations in each country over time. All of these model inputs are treated as fixed (i.e. not estimated) in our analysis.

4.2 Model

In this section we give a brief overview of the model as originally posed by Heuveline (2003) and extended by us; a thorough description of the original model can be found in Appendix A. The basic idea is to project an initial age distribution of women classified by the standard five-year age groups through time by applying survival rates to each age group. In order to replenish the population, age-specific fertility rates are applied to the fecund groups and the resulting female births are then projected by exposing them to the appropriate survival rate. Population projections are made for men by applying separate survival rates to an initial age distribution of males using the number of male births from the female population to create each new birth cohort.

HIV is added by creating new sex-age-specific HIV-infected 'duration groups' to contain HIV+ people who have been infected for 0–4, 5–9, 10–14 and 15+ years. A number of new parameters mediate the effects of HIV related to infection, survival and fertility. The duration groups allow the the effects of HIV to depend on the time since infection. For example, the model includes duration-specific parameters that capture the reduction in fecundity associated with progression toward AIDS (Lee et al., 2000; du Loû et al., 1999; Sedgh et al., 2005; Nguyen et al., 2006; Hunter et al., 2003). Similarly, the increased force of mortality associated with HIV infection is modeled using fixed age-duration-specific parameters.

Our extensions and assessment of the model are based primarily on HIV infection, and thus we focus our attention on the model parameters related to incidence. Consider HIV-negative women in the five-year age group a at time t_1 in a population located in region r. In the HCCMPP, the proportion of these women who are alive and HIV-positive five years later at time t_2 is denoted by $i_{f,a,t_1,r}$ and can be decomposed as

$$i_{f,a,t_1,r} \;=\; 1 - \exp\left\{-\Gamma_{t_2-t_1}\, H_r\, j_{f,a}\right\}, \tag{1}$$

where subscript f refers to females, $\Gamma_{t_2-t_1}$ captures the trend in HIV incidence between times t_1 and t_2, H_r is a population-specific parameter that determines the size of the epidemic in region r, and $j_{f,a}$ is the age-specific parameter for females that measures incidence relative to women aged 25-29 years for whom the value is fixed at one (hereafter age-specific relative incidence ratio). The corresponding model input for males, $i_{m,a,t_1,r}$ has the same decomposition, with *women* aged 25-9 years again serving as the reference group. This decomposition allows the level of HIV incidence to vary by age and sex, as well as across populations in different locations, but the general shape of the trend through time will be the same. The value for the trend in HIV incidence between times t_1 and t_2 is calculated from the gamma distribution as follows

$$\Gamma_{t_2-t_1} \;=\; \int_{t_1}^{t_2} \frac{x^{\alpha-1}\, e^{-x/\beta}}{(\alpha-1)!\;\beta^{\alpha}}\, dx, \tag{2}$$

where α and β are parameters taking only positive values. The parameters $j_{f,a}$, $j_{m,a}$, and H_r are estimated using data compiled by Heuveline (2003), but the time trend is fixed and determined by Equation 2 with $\alpha = 5$ and $\beta = 3$. It should also be noted that an initial year t_0 needs to be chosen for when the country-specific epidemic began. This year is assumed to be the date when HIV prevalence reached 1% in the general population, and the corresponding values for the countries in our analysis are taken from the United Nations (1998).

We extend the work of Heuveline (2003) by exploring several different specifications for the trend in HIV incidence $\Gamma_{t_2-t_1}$ using two basic approaches. The first approach simply involves estimating the parameters of the gamma curve, α and β, along with the other HCCMPP parameters. This approach is then extended by estimating separate gamma curves for men and women, with sex-specific reference groups aged 25-29 (i.e. $j_{f,25-29} = 1$ and $j_{m,25-29} = 1$). The second strategy is to include an additional parameter for each of the first four projection periods and to estimate them in conjunction with the other HCCMPP parameters. Again, we explore two specifications that include a single curve shared by women and men as well as sex-specific curves.

Forecasts of HIV prevalence are made using each of these specifications for incidence, and the corresponding predictive performance is assessed by comparing the forecasts to the observed levels of HIV prevalence in the HIV/AIDS Indicator and Demographic and Health Surveys for Tanzania and Uganda.

4.3 Estimation

Maximum Likelihood. We implement a standard maximum likelihood estimation procedure described in full elsewhere (Thomas and Clark, 2008) and in Appendix B. This procedure produces point estimates for each of the model parameters and standard 95% confidence intervals, but does not provide a statistically sound method for making probabilistic projections.

Bayesian Melding. In the Bayesian framework, parameters are treated as random variables. Prior beliefs about the parameters are quantified in the form of a joint probability density $p(\theta)$, where θ is a vector of parameters for which we will make inference. The data

$\mathbf{y}$ are brought in by specifying a likelihood $\mathcal{L}(\mathbf{y}|\theta)$, which is the probability of the observed data for a given value of the parameters. Using Bayes' Theorem and the marginal density of the data $p(\mathbf{y})$, we can update our prior beliefs to obtain the posterior distribution[7]

$$\begin{aligned} p(\theta|\mathbf{y}) &= \frac{\mathcal{L}(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \\ &\propto \mathcal{L}(\mathbf{y}|\theta)p(\theta), \end{aligned} \tag{3}$$

which is used to make inference for θ.

Bayesian melding (Poole and Raftery, 2000) was designed for problems in which a deterministic model, such as HCCMPP, is used in the likelihood function. Let M represent the model which transforms a set of parameter inputs θ into a set of model outputs $\phi = M(\theta)$. As described above, the Bayesian approach requires a prior density for the model inputs $p(\theta)$ and a likelihood for the outputs and the data $\mathcal{L}(M(\theta))$. These two sources of information are combined to produce the following posterior distribution for the model inputs

$$p(\theta|\mathbf{y}) \propto \mathcal{L}(\mathbf{y}|M(\theta))p(\theta).$$

Inference is performed by sampling from $p(\theta|\mathbf{y})$ and summarizing the resulting posterior sample. Furthermore, we can run HCCMPP for each set of inputs in the posterior sample to obtain a posterior sample of the model outputs $p(\phi|\mathbf{y})$. Sex-age-specific HIV prevalence is the model output that interests us because we can use it to assess forecasts. Note that the posterior sample reflects the distribution of model outputs, and thus the quantiles of the posterior sample can be used to make probabilistic statements about the values of the model outputs. This feature of the Bayesian framework is used to make probabilistic forecasts of HIV prevalence which can be assessed by comparing these predictive intervals to observed data.

Bayesian Melding Estimation. In our implementation of Bayesian melding with the HCCMPP we specify independent uniform priors that are relatively uninformative and thereby place most of the influence with the observed data. We use a beta-binomial likelihood to allow for heterogeneity across the different types of data and geographic regions from which they are collected. The beta-binomial distribution is a mixture of binomial distributions $n \sim binomial(N, \pi)$, with the mean of the binomials following a beta distribution $\pi \sim beta(a, b)$. We adopt the re-parameterization $\pi \sim beta(\mu, M)$ of the beta distribution used by Grassly et al. (2004), where $\mu = a/(a+b)$ and $M = a+b$. The extra variation in the beta-binomial distribution (relative to the binomial) is determined by M, and the mean and variance of n are $N\pi$ and $\{1+(1+N)/(M+1)\}\pi(1-\pi)/N$, respectively. In our application the M parameter is estimated along with the other HCCMPP parameters. The likelihoods for each age, sex, and location are treated as independent and multiplied together to produce a total likelihood.

With the HCCMPP it effectively impossible to derive the analytic form of the posterior distribution because of the complexity of the model. We address this in the standard way by

[7]Equation 3 arises from the fact that $p(\mathbf{y})$ does not depend on θ, so the posterior distribution only needs to be known up to a constant and is thus proportional to the product of the likelihood and the prior.

drawing a sample from the posterior distribution and carrying out inference for the model parameters by summarizing the posterior sample. The posterior distribution is estimated by resampling from an initial sample drawn from the importance sampling distribution using weights that identify sample members that have relatively high posterior probabilities. A transparent way to implement this approach is the sampling importance resampling (SIR) algorithm suggested by Rubin (1987, 1988) which uses the likelihood function to form the resampling weights. In this case the prior distribution serves as the importance sampling distribution. Bayesian melding has been successfully implemented with the SIR algorithm in the past (see for example Poole and Raftery, 2000; Alkema et al., 2007), but with HCCMPP the SIR approach did not work. Because the HCCMPP has so many parameters, samples from the prior distribution failed to cover important regions of the posterior distribution resulting in a poor approximation. A similar problem often occurs if the posterior distribution is multimodal or concentrated in curved manifolds (Raftery and Bao, 2010).

A more efficient approach is incremental mixture importance sampling (IMIS) which was originally introduced by Steele et al. (2003, 2006) and further developed for posterior distributions of continuous parameters by Raftery and Bao (2010). IMIS is an iterative technique that builds up an importance sampling distribution by adding new points in areas of high posterior probability at each step, based on the idea of defensive mixture distributions developed by Hesterberg (1995). This feature of IMIS ensures that the target distribution (the posterior in our case) is adequately covered by the importance sampling distribution, resulting in much greater efficiency than SIR. The following steps outline the IMIS algorithm we use to implement Bayesian melding. We refer to this version of the algorithm as IMIS-*opt* because it includes steps that require the use of a function optimizer[8].

1. Begin by drawing $B0 = d * 1,000$ inputs $\theta_1, \ldots, \theta_{B0}$ from the prior distribution $p(\theta)$, where d is the dimension of θ. Calculate the importance weights

$$w_i^{(0)} \propto \frac{\mathcal{L}_i}{\sum_{i=1}^{B0} \mathcal{L}_i}$$

   where $\mathcal{L}_i$ is the likelihood for the i^{th} input.

2. Use the input with the maximum weight as the starting value for an optimization routine that maximizes the log likelihood using 100 function evaluations. If the local optimum has a likelihood larger than any other input from the prior, then save the local optimum, θ_1^{opt}, and calculate the inverse of the Hessian matrix, Σ_1^{opt}. If the Hessian does not yield a positive definite covariance matrix, then use the matrix of first derivatives of the likelihood times the prior (evaluated at the local optimum) to create a new information matrix by adding it to the precision matrix of the prior distribution, and using the inverse this new matrix as the covariance matrix.

3. For i = 2:D (D = 10) exclude the starting points and the fraction of inputs $\frac{1}{D}$ that have the smallest Mahalonobis distance to $\theta_{(i-1)}^{opt}$. Of the remaining inputs, choose the one with the largest weight as the new starting point for obtaining θ_i^{opt}.

[8] In our work with the HCCMPP, we use the `optim` routine in the `R` programing language (R Foundation for Statistical Computing, 2010).

4. For each saved local optimum, indexed by s, sample $B = 400$ new inputs, H_s, from a multivariate Gaussian distribution with center θ_s^{opt} and covariance matrix Σ_s^{opt}.

5. For $k = 1, 2, \ldots$ repeat the following steps until stopping criterion is satisfied.

   (a) Form the posterior sampling weights

   $$w_i^{(k)} \propto \frac{\mathcal{L}_i \, p(\theta_i)}{q^{(k)}(\theta_i)}$$

   where $q^{(k)}(\theta) = \frac{B0}{N_k} p + \frac{B}{N_k} \sum_{s=1}^{D^*+k-1} H_s$, N_k is the total number of inputs at stage k, and D^* is the number of saved local optima.

   (b) Take the input with the maximum weight, θ^k, as the center of a multivariate Gaussian distribution, H_{D^*+k}. Use the $d * 100$ inputs with the smallest distance, with respect to the covariance of the prior distribution, from the mean to calculate the weighted covariance matrix, $\Sigma^{(k)}$, with weights that are proportional to the average of the importance weights and $\frac{1}{N_k}$. Sample B new inputs from H_{D^*+k}.

   (c) If the expected number of unique points

   $$\hat{Q}(w) = \sum_{i=1}^{N_k} (1 - (1 - w_i)^M)$$

   is greater than $M * (1 - \frac{1}{e})$, then stop iterating and re-sample M inputs with replacement from $\theta_1, \ldots, \theta_{N_k}$ with weights $w_1, \ldots, w_{N_k}$. In our application to HCCMPP, we set $M = 3,000$ which requires the expected number of unique points to be 1,896.

The first step in the IMIS-opt algorithm is essentially the same as the SIR algorithm, except that with the latter approach the resampling is done with weights proportional to $w_i^{(0)}$. The additional optimization steps (2 - 4) in the IMIS-opt algorithm seek to cover areas in the posterior sampling space that have high posterior probability (relative to the prior). Given several local optima, the algorithm then proceeds by adding new components to the sampling function that are centered around the inputs with the largest weights, with the local neighborhoods providing the covariance information.

4.4 Calibration & Validation

Our assessment of the HCCMPP is based on an attempt to accurately forecast sex-age-specific HIV prevalence in Tanzania, as measured in the 2004 and 2007 DHS. The first step in this assessment involves calibrating the model to adequately reproduce the first twenty years of the HIV epidemic in this country. We use data collected before 2000 from urban and rural areas located in Burundi, Tanzania, and Uganda to estimate the HCCMPP parameters. Observations from these other countries were included to increase the stability and precision of our parameter estimates, and seemed reasonable given the close proximity of these countries. After calibrating the model to the local epidemic in East Africa, these

estimates are then used to make HCCMPP forecasts of the levels of HIV prevalence observed in the 2004 and 2007 DHS in Tanzania. The Bayesian melding framework which we employ for parameter estimation and forecasting yields a posterior distribution of HIV prevalence by age and sex. These distributions allow us to assess the accuracy of the HCCMPP forecasts by comparing the observed coverage of the predictions with the nominal coverage. For example, we expect half of the observed levels of HIV prevalence to fall within the 50% prediction intervals, and similarly for the 80% and 95% prediction intervals reported here. Because our forecasts take into account uncertainty in HIV incidence but not the vital rates, our forecasts understate the amount of uncertainty around future levels of HIV prevalence – i.e. the coverage of our forecasts will be slightly too low.

5 Background & Significance

The most important indicator of an HIV epidemic is incidence, the rate at which uninfected members of the population become infected. The logistic and economic difficulties with measuring and tracking HIV incidence motivates the use of a modeling approach to study changes and variation in HIV incidence. Thus, we focus our attention here on the implementation of new infections in the HCCMPP and how this relates to the more recent empirical record. Palloni (1996) points out that in a demographic model with HIV/AIDS the force of infection that produces the current level of prevalence should depend on the past level of prevalence. This endogeneity of HIV incidence is not modeled directly in the HCCMPP; instead, a simple approximation is used to provide a plausible trend in HIV incidence. Heuveline (2003) adopted a gamma curve to determine the incidence trend, a strategy also used in previous models of HIV/AIDS epidemics (e.g., Chin and Lwanga, 1991; Salomon and Murray, 2001). Additional parameters are included in the HCCMPP to allow the risk of infection to vary by age, sex, and location, but the underlying trend is the same. In other words, the levels are estimated and the pattern over time is fixed.

While the fixed gamma curve used by Heuveline (2003) may provide a plausible course of development for an HIV epidemic, there has been little (if any) validation of this assumption. In the face of such model uncertainty the usual practice is to turn to the empirical record for guidance. In the present case, however, the available evidence is fairly limited as there are few studies that provide information on the trends of HIV incidence in sub-Saharan Africa. Among the exceptions is Mbulaiteye et al. (2002) who tracked HIV incidence in a cohort study carried out in Masaka, Uganda from 1989 to 1999 and found evidence of a decline over that period. However, a study by Kamali et al. (2000) in the same area produced estimates of incidence by sex and age that illustrated how difficult it is to identify trends and differences given the large amount of uncertainty around the point estimates when disaggregating the population by sex and age. In an open-cohort study carried out at a demographic surveillance system in rural Tanzania, Wambura et al. (2007) collected data on HIV incidence by village type, sex and age with serosurveys conducted in 1994-1995, 1996-1997, 1999-2000, and 2003-2004[9]. The three point estimates of HIV incidence for the

[9]There were around 2,700 men and 3,300 women tested in each serosurvey. Also, Wambura et al. (2007) combined the central trading center with roadside villages in their analysis, and compared these to the

intervals between the serosurveys suggested differences in the trends for men and women in roadside villages, as well as between women living in remote rural villages and those living in roadside villages. Among men living in roadside villages, there was a significant increase in the crude incidence rate from the first to the second estimate, while the third point estimate did not differ from the second. A similar trend was found for both men and women living in remote villages, but the level of incidence was significantly lower than that of men in roadside villages for each time period. The trend among women in the roadside villages differed significantly in each time period with an increase followed by a decline in crude incidence. Wambura et al. (2007) also explored HIV incidence trends by broad age groups for men and women in different locations. Although the uncertainty around the point estimates makes it difficult to draw conclusions, the results do motivate the hypothesis that the incidence trends may differ by age, sex, and location.

In addition to direct measures of HIV incidence, there are other sources that provide information about the trends over time. One suggested option is to use HIV prevalence among women aged 15-24 years who attend antenatal clinics (UNAIDS, 2009b). The underlying logic is that these young women have only been exposed to the risk of infection via heterosexual transmission for a short period of time, and thus those who are HIV positive are likely to have been infected fairly recently, with a very small percentage dying from AIDS. While this metric may serve as a useful means for monitoring the epidemic, it provides little information about HIV incidence at older ages (see also Wawer et al., 1997; Zaba et al., 2000; Ghys et al., 2006). Another potentially helpful source of information includes the hypotheses explored with other epidemiological and demographic models of HIV/AIDS. For example, Gregson et al. (1997) describe the predictions of a model that fit the so-called HIV-1 hypothesis, which posits a pattern of sexual behavior in rural areas where men are typically infected first, perhaps while working in an urban center or town, and then infect their female partners. This pattern of sexual mixing may result in different trends in HIV incidence between men and women, or differences between urban and rural areas. Results from other models suggest that the peak in HIV incidence occurs earlier than what is produced from the gamma curve used in the HCCMPP (Stoneburner et al., 1996; Salomon and Murray, 2001).

These various sources of information concerning the trends in HIV incidence in sub-Saharan Africa motivate our attempt to explore new specifications of the HCCMPP. Efforts to this end will be useful in helping to identify useful sources of variation in the risk of infection, which will in turn help to formulate successful plans for interventions and treatment.

6 Discussion

6.1 Summary

This paper makes two main contributions, the first is to validate the HCCMPP developed by Heuveline (2003). We find that the model can produce accurate forecasts of age- and

remote rural villages.

sex-specific HIV prevalence in Tanzania, and that an assumption of a stabilized trend in HIV incidence provides estimates of the extent of the decline in the risk of HIV infection in Uganda. In order to produce accurate forecasts with the HCCMPP for other countries in sub-Saharan Africa, the model may require new modifications to capture the geographic heterogeneity in the HIV epidemics across this region. Other potential sources of variability not captured by the model are the uncertainty around the start date of the epidemic and around vital rates. Including these as estimated model inputs may be useful in terms of improving the predictive performance of HCCMPP and applying it to other countries. Despite these issues, it is impressive how well the model does when considering the differences between the data used to calibrate it (i.e. estimate the HCCMPP parameters) and the data used to validate the model forecasts. The former were collected from relatively small community-based studies while the latter were collected from nationally representative samples.

The second contribution is to use a new Bayesian estimation technique designed for deterministic models. We have shown that the IMIS algorithm (Raftery and Bao, 2010) can be used successfully to implement the Bayesian melding estimation approach with the 30+ parameter HCCMPP model. This suggests that the approach could be used more generally to enable demographers to quantify in a statistically rigorous way the uncertainty around both parameter estimates and model outputs in many of the deterministic models they use. With respect to the CCMPP (HIV-enabled or not), the ability to quantify uncertainty around demographic projections is useful in a fundamental sense. This allows decision-makers to define the probability of both the best and worst case scenarios and make appropriate cost-benefit and risk tolerance decisions in a valid probabilistic framework. And further as we demonstrate above, probabilistic projections produced using a CCMPP-type model and Bayesian melding with IMIS can be used to validate a the model and compare competing model specifications using Bayes factors.

Finally, we attempt to shed some light on the likely pathways that epidemics follow as they mature by using data to estimate the trends in HIV incidence. Our results suggest that the gamma curve used by Heuveline (2003) in the original work with HCCMPP provides the best predictive performance of sex-age-specific HIV prevalence in Tanzania.

6.2 Ideas for Future Work

While the HCCMPP appears to be a useful tool for helping to monitor and study HIV epidemics, there are several areas in which the model could be improved. For example, while the fixed gamma curve used to represent the trend in HIV incidence is simple and easy to work with, it does not acknowledge the fundamental endogeneity of HIV incidence identified by Palloni (1996). Future work with the HCCMPP should modify the model so that HIV incidence depends on the current and past levels of prevalence. Another obvious area of improvement is the need to account for the increases in the coverage of antiretroviral therapies (ART). Given an estimate of ART coverage for a particular population, this share of the infected population should experience improved survival and fertility prospects, and a diminished likelihood of infecting others, relative to those who are HIV+ but not receiving ART. Additional modifications could include the ability to model potential interventions

related to male circumcision and microbicides (McNeil Jr., 2010). Building in these features would make the HCCMPP more realistic and potentially improve the model's prospects for successfully monitoring and forecasting HIV epidemics. However, in order to add these new features the overall model would have to be made simpler to require fewer parameters in order to ensure that the whole thing remains identifiable and tractable enough to estimate. This could be done by modeling existing parameters and defining (a smaller number of) new hyper parameters to govern those models and/or by collapsing across age groups in which there is little meaningful variation.

7 Attributions & Acknowledgements

Jason Thomas conducted the bulk of the analysis for this project, wrote and ran all of the R code and wrote the first draft of the manuscript. Samuel Clark conceived and supervised this project and wrote the final manuscript. With Adrian Raftery Le Bao refined the IMIS method and wrote the original IMIS code that was adapted for use in this project, and Le Bao contributed significantly to the adaptation of his IMIS code for use in this project. All authors have read and approved this manuscript.

This project was supported by a seed grant from the Center for Statistics and the Social Sciences (CSSS), University of Washington and by grants K01 HD057246 and R01 HD054511 from the National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health (NIH). The second author gratefully acknowledges the institutional support of the Center for Demography and Ecology at the University of Wisconsin-Madison, funded by P30 AG17266. The content of the work presented here is solely the responsibility of the authors and does not necessarily represent the official views of the the NIH. The authors are very grateful to Patrick Heuveline for his correspondence and for making his work available. We also thank Adrian Raftery, Leontine Alkema, Jennifer Chunn, Greg Mathews, David Sharrow, and Mark Wheldon for invaluable discussion during the preparation of this work. The second author would also like to thank colleagues at the Africa Centre for Health and Population Studies, University of KwaZulu-Natal, South Africa for their hospitality and useful comments on an early version of this work.

This project was entirely conducted using open source software. The methods described above are implemented in the R programming language and run in the R statistical software package (R Foundation for Statistical Computing, 2010). This manuscript was prepared using the LaTeX(LaTeX Project, 2010) typesetting language.

References

Alkema, L., A. E. Raftery, and S. J. Clark (2007). Probabilistic projections of HIV prevalence using Bayesian melding. *The Annals of Applied Statistics 1*(1), 229–248.

Anderson, R. (1988). The role of mathematical models in the study of HIV transmission

and the epidemiology of AIDS. *JAIDS Journal of Acquired Immune Deficiency Syndromes 1*(3), 241.

Asamoah-Odei, E., J. Calleja, and J. Boerma (2004). HIV prevalence and trends in sub-Saharan Africa: no decline and large subregional differences. *The Lancet 364*(9428), 35–40.

Bongaarts, J., T. Buettner, G. Heilig, and F. Pelletier (2008). Has the HIV epidemic peaked? *Population and Development Review 34*(2), 199–244.

Bowley, A. (1924). Births and population in Great Britain. *Economic Journal 34*(134), 188–92.

Cannan, E. (1895). The probability of a cessation of the growth of population in England and Wales during the next century. *Economic Journal 5*(20), 505–15.

Carpenter, L. M., J. S. Nakiyingi, A. Ruberantwari, S. S. Malamba, A. Kamali, and J. A. Whitworth (1997). Estimates of the impact of HIV infection on fertility in a rural Ugandan population cohort. *Health Transition Review 7*(Supplement 2), 113–126.

Cassels, S., S. Clark, and M. Morris (2008). Mathematical models for HIV transmission dynamics: tools for social and behavioral science research. *JAIDS Journal of Acquired Immune Deficiency Syndromes 47*, S34.

Chin, J. and S. Lwanga (1991). Estimation and projection of adult AIDS cases: A simple epidemiological model. *Bulletin of the World Health Organization 69*(4), 399–406.

Dorn, H. F. (1950). Pitfalls in population forecasts and projections. *Journal of the American Statistical Association 45*(251), 311–334.

du Loû, A. D., P. Msellati, A. Yao, V. Noba, I. Viho, R. Ramon, C. Welffens-Ekra, and F. Dabis (1999). Impaired fertility in HIV-1-Infected pregnant women: a clinic-based survey in Abidjan, Côte d'Ivoire, 1997. *AIDS 13*(4), 517–521.

Ehrenberg, A. (1982). Writing technical papers or reports. *The American Statistician 36*(4), 326–329.

Fabiani, M., S. Accorsi, M. Lukwiya, T. Rosolen, E. Ayella, P. Onek, and S. Declich (2001). Trend in HIV-1 prevalence in an antenatal clinic in north Uganda and adjusted rates for the general population. *AIDS 15*, 97–103.

Fylkesnes, K., Z. Ndhlovu, D. Dasumba, R. Musonda, and M. Sichone (1998). Studying the dynamics of the HIV epidemic: Population-based data compared with sentinel surveillance in Zambia. *AIDS 12*, 1227–34.

Ghys, P., E. Kufa, and M. G. for the UNAIDS Reference Group on Estimates, Modelling and Projections' ad hoc Working Group on interpreting trends in prevalence and incidence of HIV infection in countreis with generalised epidemics (2006). Measuring trends in prevalence and incidence of HIV infection in countries with generalised epidemics. *Sexually Transmitted Infections 82*(Supplement 1), i52–i56.

Ghys, P. D., T. Brown, N. C. Grassly, G. Garnett, K. A. Stanecki, J. Stover, and N. Walker (2004). The UNAIDS estimation and projection package: a software package to estimate and project national hiv epidemics. *Sexually Transmitted Infections 80*, i5–i9.

Granich, R., C. Gilks, C. Dye, K. De Cock, and B. Williams (2009). Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. *The Lancet 373*(9657), 48–57.

Grassly, N., M. Morgan, N. Walker, G. Garnett, K. Stanecki, J. Stover, T. Brown, and P. Ghys (2004). Uncertainty in estimates of HIV/AIDS: the estimation and application of plausibility bounds. *Sexually Transmitted Infections 80*, i31–i38.

Gray, R. H., M. J. Wawer, D. Serwadda, N. Sewankambo, C. Li, F. Wabwire-Mangen, L. Paxton, N. Kiwanuka, G. Kigozi, J. Konde-Lule, T. C. Quinn, and C. A. Gaydos (1998). Population-based study of fertility in women with HIV-1 infection in Uganda. *Lancet 351*(9096), 98–103.

Gregson, S., R. M. Anderson, J. Ndlovu, T. Zhuwau, and S. K. Chandiwana (1997). Recent upturn in mortality in rural Zimbabwe: evidence for an early demographic impact of HIV-1 infection? *AIDS 11*, 1269–1280.

Gregson, S. and G. Garnett (2000). Contrasting gender differentials in HIV-1 prevalence and associated mortality increase in eastern and southern Africa: Artefact of data or natural course of epidemics? *AIDS 14*(Suppl.3), S85–99.

Grosskurth, H., F. Mosha, J. Todd, K. Senkoro, J. Newell, A. Lkokke, J. Changalucha, B. West, P. Mayaud, A. Gavyole, R. Gabone, D. Mabey, and R. Hayes (1995). A community trial of the impact of improved sexually transmitted disease treatment on the HIV epidemic in rural Tanzania. *AIDS 9*, 927–34.

Hallett, T., J. Aberle-Grasse, G. Bello, L. Boulos, M. Cayemittes, B. Cheluget, J. Chipeta, R. Dorrington, S. Dube, A. Ekra, J. Garcia-Calleja, G. P. Garnett, S. Greby, S. Gregson, J. Grove, S. Hader, J. Hanson, W. Hladik, S. Ismail, S. Kassim, W. Kirungi, L. Kouassi, A. Mahomva, L. Marum, C. Maurice, M. Nolan, T. Rehle, J. Stover, and N. Walker (2006). Declines in HIV prevalence can be associated with changing sexual behaviour in Uganda, urban Kenya, Zimbabwe, and urban Haiti. *Sexually Transmitted Infections 82*(suppl 1), i1.

Hallett, T., P. White, and G. Garnett (2007). Appropriate evaluation of HIV prevention interventions: from experiment to full-scale implementation. *Sexually transmitted infections 83*(suppl 1), i55.

Hallett, T.B. and Singh, K., J. Smith, R. White, L. Abu-Raddad, and G. Garnett (2008). Understanding the impact of male circumcision interventions on the spread of HIV in southern Africa. *PLoS One 3*(5).

Hallett, T.B. and Żaba, B., J. Todd, B. Lopman, W. Mwita, S. Biraro, S. Gregson, J. Boerma, and on behalf of the ALPHA Network (2008). Estimating incidence from prevalence in generalised HIV epidemics: methods and validation. *PLoS Med 5*(4), e80.

Hesterberg, T. (1995). Weighted average importance sampling and defensive mixture distributions. *Technometrics 37*, 185–194.

Heuveline, P. (2003, May). HIV and population dynamics: A general model and maximum-likelihood standards for east Africa. *Demography 40*(2), 217–245.

Hira, S., J. Kamanga, G. Bhat, C. Mwale, G. Tembo, N. Luo, and P. Perine (1989). Perinatal transmission of HIV-1 in Zambia. *British Medical Journal 299*(6718), 1250–52.

Hunter, S.-C., R. Isingo, J. T. Boerma, M. Urassa, G. M. Mwaluko, and B. Żaba (2003). The association between HIV and fertility in a cohort study in rural Tanzania. *Journal of Biosocial Science 35*, 189–199.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.

Kamali, A., L. M. Carpenter, J. A. G. Whitworth, R. Pool, A. Ruberantwari, and A. Ojwiya (2000). Seven-year trends in HIV-1 infection rates, and changes in sexual beaviour, among adults in rural Uganda. *AIDS 14*, 427–434.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kengeya-Kayondo, J., A. Kamali, A. Nunn, A. Ruberantwari, H. Wagner, and D. Mulder (1996). Incidence of hiv-1 nfection in adults and socio-demographic characteristics of seroconverters in a rural population in Uganda: 1990-1994. *International Journal of Epidemiology 25*, 1077–82.

Kilian, A., S. Gregson, B. Ndyanabangi, K. Walusaga, W. Kipp, G. Sahlmüller, G. Garnett, G. Asiimwe-Okiror, G. Kabagambe, P. Weis, and F. von Sonnenburg (1999). Reductions in risk behaviour provide the most consistent explanation of declining HIV prevalence in Uganda. *AIDS 13*, 391–98.

LaTeX Project (2010). *LaTeX - A Document Preparation System*. http://www.latex-project.org.

Lee, L. M., P. M. Wortley, P. L. Fleming, L. J. Eldred, and R. H. Gray (2000). Duration of human immunodeficiency virus infection and likelihood of giving birth in a medicaid population in Maryland. *American Journal of Epidemiology 151*(10), 1020–1028.

Leslie, P. (1945). On the use of matrices in certain population mathematics. *Biometrika 33*(3), 183–212.

Mbulaiteye, S., C. Mahe, J. Whitworth, A. Ruberantwari, J. Nakiyingi, A. Ojwiya, and A. Kamali (2002). Declining HIV-1 incidence and associated prevalence over 10 years in a rural population in south-west Uganda: a cohort study. *Lancet 360*.

McDougal, J., B. Parekh, M. Peterson, B. Branson, T. Dobbs, M. Ackers, and M. Gurwith (2006). Comparison of HIV type 1 incidence observed during longitudinal follow-up with incidence estimated by cross-sectional analysis using the BED capture enzyme immunoassay. *AIDS Research & Human Retroviruses 22*(10), 945–952.

McNeil Jr., D. G. (2010). African studies give women hope in HIV fight. *New York Times* (July 19).

McWalter, T. and A. Welte (2009). A comparison of biomarker based incidence estimators.

McWalter, T. and A. Welte (2010). Relating recent infection prevalence to incidence with a sub-population of assay non-progressors. *Journal of mathematical biology 60*(5), 687–710.

Morgan, D., C. Mahe, B. Mayanja, J. Okongo, R. Lubega, and J. Whitworth (2002). HIV-1 infection in rural Africa: Is there a difference in median time to AIDS and survival compared with that in industrialized countries? *AIDS 16*, 597–603.

Nguyen, R. H. N., S. J. Gange, F. Wabwire-Mangen, N. K. Sewankambo, D. Serwadda, M. J. Wawer, T. C. Quinn, and R. H. Gray (2006). Reduced fertility among HIV-infected women associated with viral load in Rakai district, Uganda. *International Journal of STD and AIDS 17*, 842–846.

Nunn, A., D. Mulder, A. Kamali, A. Ruberantwari, J. Kengeya-Kayondo, and J. Whitworth (1997). Mortality associated with HIV-1 infection over five years in a rural Ugandan population: Cohort study. *British Medical Journal 315*(7111), 767–71.

Nunn, A. J., J. F. Kengeya-Kayondo, S. S. Malamba, J. A. Seeley, and D. W. Mulder (1994). Risk factors for HIV-1 infection in adults in a rural Ugandan community: a population study. *AIDS 8*, 81–6.

Palloni, A. (1996). The demography of HIV/AIDS. *Population Index 62*(4), 601–652.

Parekh, B., M. Kennedy, T. Dobbs, C. Pau, R. Byers, T. Green, D. Hu, S. Vanichseni, N. Young, K. Choopanya, et al. (2002). Quantitative detection of increasing HIV type 1 antibodies after seroconversion: a simple assay for detecting recent HIV infection and estimating incidence. *AIDS Research and Human Retroviruses 18*(4), 295–307.

Pearl, R. and L. J. Reed (1920). On the rate of growth of the population of the United States since 1790 and its mathematical representation. *Proceedings of the National Academy of Science 6*(6), 275–288.

Poole, D. and A. E. Raftery (2000). Inference for deterministic simulation models: The bayesian melding approach. *Journal of the American Statistical Association 95*(452), 1244–1255.

Pritchett, H. (1891). A formula for predicting the population of the United States. *Publications of the American Statistical Association 2*(14), 278–286.

R Foundation for Statistical Computing (2010). *The R Project for Statistical Computing*. http://www.r-project.org.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology 25*, 111–163.

Raftery, A. E. and L. Bao (2010). Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics 66*(DOI: 10.1111/j.1541-0420.2010.01399.x), in press.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputation when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association 82*(398), 543–546.

Rubin, D. B. (1988). *Bayesian Statistics 3*, Chapter Using the SIR Algorithm to Simulate Posterior Distributions, pp. 395–402. Oxford, U.K.: Clarnedon Press.

Saidel, T., D. Sokal, J. Rice, T. Buzingo, and S. Hassig. (1996). Validation of a method to estimate age-specific human immunodeficiency virus (HIV) incidence in developing countries using population-based seroprevalence data. *American Journal of Epidemiology 144*, 214–23.

Salomon, J. A. and C. J. Murray (2001). Modelling HIV/AIDS epidemics in sub-Saharan Africa using seroprevalence data from antenatal clinics. *Bulletin of the World Health Organization 79*(7), 596–607.

Sedgh, G., U. Larsen, D. Spiegelman, G. Msamanga, and W. W. Fawzi (2005). HIV-1 disease progression and fertility in Dar es Salaam, Tanzania. *Journal of Acquired Immune Deficiency Syndrome 39*(4), 439–445.

Serwadda, D., M. Wawer, S. Musgrave., N. Sewankambo, J. Kaplan, and R. Gray (1992). HIV risk factors in three geographic strata of rural Rakai district, Uganda. *AIDS 6*, 983–89.

Sewankambo, N., M. Wawer, R. Gray, D. Serwadda, C. Li, R. Stallings, S. Musgrave, and J. Konde-Lule (1994). Demographic impact of HIV infection in rural Rakai district, Uganda: Results of a population-based cohort study. *AIDS 8*, 1707–13.

Shao, J., G. Burbaker, A. Levin, A. Kibauri, E. Massesa, Z. Siso, E. Konings, Y. Clayton, D. Kumby, S. Alexander, D. Waters, J. Drummond, R. Biggar, G. Scott, G. Miller, J. Goedert, and W. Blattner. (1994). Population-based study of HIV-1 infection in 4,086 subjectgs in northwest Tanzania. *Journal of Acquired Immune Deficiency Syndrome 7*, 397–402.

Slutsker, L., J. Cabeza, J. Wirima, and R. Steketee (1994). HIV-1 infection among women of reproductive age in a rural district in Malawi. *AIDS 8*, 1337–40.

Sokal, D., T. Buzingo, N. Nitunga, P. Kadende, and B. Standaert (1993). Geographic and temporal stability of HIV seroprevalence among pregnant women in Bujumbura, Burundi. *AIDS 7*, 1481–84.

Steele, R. J., A. E. Raftery, and M. J. Edmonds (2006). Computing normalizing constants for finite mixture models via incremental importance sampling (IMIS). *Journal of Computational and Graphical Statistics 15*(3), 712–734.

Steele, R. J., A. E. Raftery, and M. J. Emond (2003). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (IMIS). Technical Report 436, Department of Statistics, University of Washington, Seattle, Washington.

Stoneburner, R. L., D. Low-Beer, G. S. Tembo, T. E. Mertens, and G. Asiimwe-Okiror (1996). Human immunodeficiency virus infection dynamics in east Africa deduced from surveillance data. *American Journal of Epidemiology 144*(7), 682–695.

Thomas, J. and S. J. Clark (2008). More on the cohort-component model of population projection in the context of HIV/AIDS: A leslie matrix representation and new estimation methods. *CSSS Working Paper no. 88*.

Todd, J., R. Balira, H. Grosskurth, P. Mayaud, F. Mosha, G. ka-Gina, A. Klokke, R. Gabbone, A. Gavyole, D. Mabey, and R. Hayes (1997). HIV-associated adult mortality in a rural Tanzanian population. *AIDS 11*, 801–07.

UNAIDS (2009a). *AIDS Epidemic Update*. Joint United Nations Programme on HIV/AIDS (UNAIDS) and the World Health Organization (WHO).

UNAIDS (2009b). *Monitoring the Declaration of Commitment on HIV/AIDS: Guidelines on Construction of Core Indicators: 2010 Reporting*. Geneva: Joint United Nations Programme on HIV/AIDS (UNAIDS).

United Nations (1982). *Model Life Tables for Developing Countries*. New York: United Nations.

United Nations (1998). AIDS, Mortality, and Population Change – Technical Meeting on the Demographic Impact of HIV/AIDS, Tuesday, 10 November 1998. Available online at http://www.un.org/esa/population/pubsarchive/hivmtg/aidsrep.pdf (downloaded on February 13, 2008).

United Nations (1999). *World Population Prospects: the 1998 Revision*. New York: United Nations.

United Nations (2007). *World Population Prospects: the 2006 Revision*. New York: United Nations.

Wambura, M., M. Urassa, R. Isingo, M. Ndege, M. Marston, E. Slaymaker, J. Mngara, J. Changalucha, T. J. Boerma, and B. Żaba (2007). HIV prevalence and incidence in rural Tanzania: Results from 10 years of follow-up in an open-cohort study. *Journal of Acquired Immune Deficiency Syndrome 46*(5), 616–623.

Wawer, M., D. Serwadda, S. Musgrave, J. Konde-Lule, M. Musagara, and N. Sewankambo. (1991). Dynamics of spread of HIV-1 infection in a rural district of Uganda. *British Medical Journal 303*(6813), 1303–06.

Wawer, M., N. Sewankambo, S. Berkley, D. Serwadda, S. Musgrave, R. Gray, M. Musagara, R. Stallings, and J. Konde-Lule (1994). Incidence of HIV-1 infection in a rural region of Uganda. *British Medical Journal 303*(6813), 1303–06.

Wawer, M. J., D. Serwadda, R. H. Gray, N. K. Sewankambo, C. Li, F. Nalugoda, T. Lutalo, and J. K. Konde-Lule (1997). Trends ni HIV-1 prevalence may not reflect trends in incidence in mature epidemics: data from the Rakai population-based cohort, Uganda. *AIDS 11*, 1023–1030.

Whelpton, P. (1936). An empirical method of calculating future population. *Journal of the American Statistical Association 31*(195), 457–473.

World Health Organization (2008). *Towards Universal Access: Scaling up the priority of HIV/AIDS interventions in the health sector*. Geneva: WHO Press.

Żaba, B., T. Boerma, and R. White (2000). Monitoring the AIDS epidemic using HIV prevalence data among young women attending antenatal clinics: prospects and problems. *AIDS 14*, 1633–1645.

Appendices

The following appendices adapted in whole from Thomas and Clark (2008).

A HIV-enabled CCMPP

A.1 CCMPP

Heuveline (2003) extends the standard CCMPP to accommodate a population categorized by duration of infection with HIV using five 'HIV duration' groups. There are four HIV+ duration groups (0-4 years, 5-9 years, 10-14 years, and 15+ years) as well as an HIV– group. In this section we present Heuveline's multi-state CCMPP for a population with 17 age groups (0-4, 5-9, ..., 80+) in each of the five HIV duration groups. The model is introduced with a series of equations representing the transition from one group/time period to the next. While the model can be applied to both men and women, the description presented here only includes the details for women.

Begin by dividing the population into age groups where $a = 1, 2, \ldots, 17$ correspond to age groups $0-4, 5-9, \ldots, 80+$. Denote membership in the HIV duration groups by d, with $d = 1, 2, \ldots, 5$ corresponding to HIV–, HIV+ for 0-4 years, ..., HIV+ for more than 15 years. Time is indexed by t noting that the duration between t and $t+1$ is equal to the width of a standard age interval, i.e. 5 years. Let $n_{a,d,t}$ be the number of women in age group a and duration group d at time t. For $1 < a < 17$, we have:

$$n_{a+1,1,t+1} = n_{a,1,t}\ s_{a,1,t}\ (1 - i_{a,t}) \quad \text{(A-1)}$$

$$n_{a+1,2,t+1} = n_{a,1,t}\ s_{a,1,t}\ i_{a,t}\ s_{a,2} \quad \text{(A-2)}$$

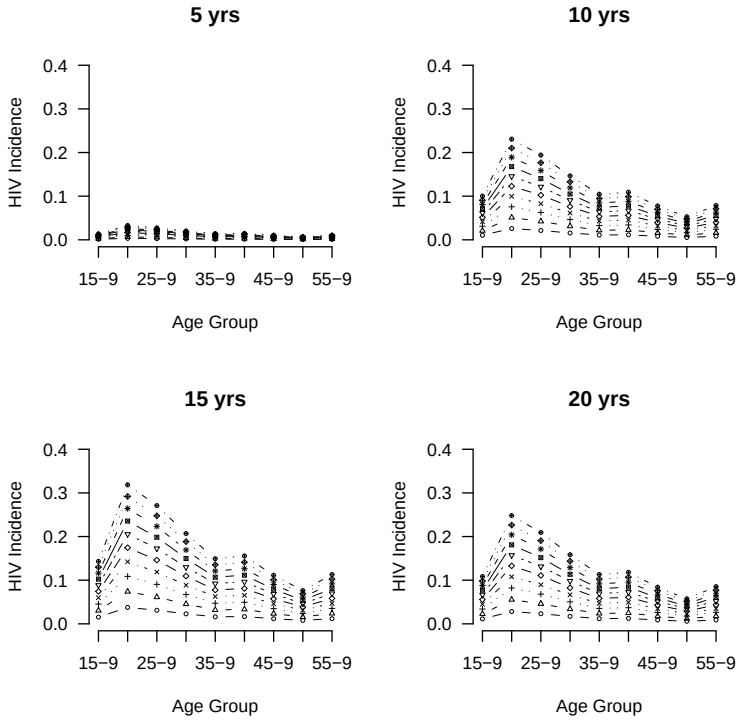
$$n_{a+1,d,t+1} = n_{a,d-1,t}\ s_{a,1,t}\ s_{a,d} \quad \text{for } d > 2 \quad \text{(A-3)}$$

where $s_{a,d}$ is the survivorship ratio for age group a and duration group d. Note that for $2 < d < 5$ this survivorship ratio determines the transition from one age group to the next, as well as from one duration group to the next. Each HIV+ group is exposed to the same underlying survivorship ratio as the HIV– group, $s_{a,1,t}$, in addition to this extra survivorship ratio that accounts for the increased mortality associated with different durations of infection. The parameter $i_{a,t}$ is the fraction of women in age group a who become infected with HIV over the projection interval. To allow for the heterogeneity of HIV epidemics across populations, this parameter is decomposed as:

$$i_{a,t} = 1 - \exp\left\{-\Gamma_{t-t_0}\ H\ j_a\right\} \quad \text{(A-4)}$$

where Γ_{t-t_0} is a parametric curve used to model the time trend in the HIV epidemic from the start time t_0. The actual values for Γ_{t-t_0} are presented in Table A-1 (see the next section for more details). The parameter H is a population-specific scale parameter that captures the overall magnitude of the epidemic. The parameter j_a is an sex-age-specific scaling factor for incidence that represents the multiplicative difference in HIV incidence between age group a

Figure A-1. Age-specific HIV incidence rates, $i_{a,t}$, for different values of the population-specific scale parameter, H, over time.



and a reference age group which is held constant at a value of 1.0 in order to make the model identifiable. Following Heuveline we set the reference age group to 25-29, i.e. $j_5 = 1.0$.

For a given age profile of incidence (a specific set of values for j_a), Figure A-1 demonstrates how the different values for H scale the incidence profile. Each panel in this figure corresponds to a different time in the epidemic, with the overall scale of HIV incidence determined by the values of Γ. Within each panel, each line corresponds to a value of the population-specific scale parameter H ranging from 0.1 to 1.0.

The projection equations are slightly different for the youngest and oldest age groups. The oldest (open-ended) age group is incremented by two sources, those 75-79 and 80+ in the previous time period. Thus for $a = 17$ we have:

$$\begin{aligned} n_{17,1,t+1} &= n_{16,1,t}\, s_{16,1,t}\, (1 - i_{16,t}) \\ &\quad +\ n_{17,1,t}\, s_{17,1,t}\, (1 - i_{17,t}) \qquad \text{(A-5)} \\ n_{17,2,t+1} &= n_{16,1,t}\, s_{16,1,t}\, i_{16,t}\, s_{16,2} \\ &\quad +\ n_{17,1,t}\, s_{17,1,t}\, i_{17,t}\, s_{17,2} \qquad \text{(A-6)} \\ n_{17,d,t+1} &= n_{16,d-1,t}\, s_{16,1,t}\, s_{16,d} \\ &\quad +\ n_{17,d-1,t}\, s_{17,1,t}\, s_{17,d} \quad \text{for } 2 < d < 5 \qquad \text{(A-7)} \end{aligned}$$

$$
\begin{aligned}
n_{17,5,t+1} &= n_{16,4,t}\; s_{16,1,t}\; s_{16,5} + n_{17,4,t}\; s_{17,1,t}\; s_{17,5} \\
&+ \; n_{16,5,t}\; s_{16,1,t}\; s_{16,5} + n_{17,5,t}\; s_{17,1,t}\; s_{17,5}
\end{aligned} \tag{A-8}
$$

As with the single-state CCMPP, the number of children in first age group at the end of the projection interval is determined by surviving forward the births that occur during the projection interval. The number of births that occur is calculated by applying age-specific fertility rates to the average number of women in each age group during the projection interval, taking into account the fact that HIV+ women who have been infected for different durations will, to varying degrees, be less likely to have children. To capture the relationship between fertility and HIV status, Heuveline defined three additional parameters. First, consider the number of HIV– births:

$$
\begin{aligned}
n_{1,1,t+1} &= s_{0,1,t}\; \frac{1}{1+SRB} \times \\
&\left(\sum_{a=\alpha}^{\beta} f_{a,1,t}\; \frac{n_{a,1,t} + p^{-}_{a-1,1,t} n_{a-1,1,t}}{2} \;+\; \sum_{d=2}^{5} \sum_{a=\alpha}^{\beta} f^{-}_{a,d,t} \frac{n_{a,d,t} + p_{a-1,d-1,t} n_{a-1,d-1,t}}{2} \right)
\end{aligned} \tag{A-9}
$$

In Equation A-9 above, the $f_{a,1,t}$'s are simply the age-specific fertility rates for HIV– women, and the lower and upper bounds of the childbearing age range are α and β, respectively. Fertility among HIV+ women introduces the following parameters

$$
f^{-}_{a,d,t} \;=\; f_{a,1,t}\; e_{a,d}\; g_d\; (1 - v_d) \tag{A-10}
$$

for $1 < d$. The superscript in $f^{-}_{a,d,t}$ designates HIV– births (i.e. $d = 1$) to women who are HIV+. The parameter v_d is the probability that an HIV+ woman in duration group d will give birth to an HIV+ child, the *vertical transmission* rate. The parameter $e_{a,d}$ captures the higher level of sexual activity and resulting fertility among HIV+ women age 15-19 who have been infected for 0-4 years ($d = 2$). In other words we expect $e_{a=4,d=2} > 1$ while $e_{a\neq 4,d}$ are constrained to be 1.0. The parameter g_d represents the *fertility impairment* experienced by women in duration group d, a number that becomes smaller as the time since infection increases, reflecting increasing fertility impairment with time since infection. The corresponding equations for HIV+ births are:

$$
n_{1,2,t+1} \;=\; s_{0,1,t}\; \frac{1}{1+SRB} \sum_{d=2}^{5} \sum_{a=\alpha}^{\beta} f^{+}_{a,d,t}\; \frac{n_{a,d,t} + p_{a-1,d-1,t}\; n_{a-1,d-1,t}}{2} \tag{A-11}
$$

$$
f^{+}_{a,d,t} \;=\; f_{a,1,t}\; e_a\; g_d\; v_d \tag{A-12}
$$

Finally, we define the factors used to approximate the average number of women at the beginning and end of the period, $p^{-}_{a,1,t}$ and $p_{a,d,t}$:

$$
p^{-}_{a,1,t} \;=\; s_{a,1,t}\; (1 - i_{a,t}) \tag{A-13}
$$

$$
p_{a,1,t} \;=\; s_{a,1,t}\; i_{a,t}\; s_{a,2} \tag{A-14}
$$

$$
p_{a,d,t} \;=\; s_{a,1,t}\; s_{a,d} \quad \text{for } \; d > 1. \tag{A-15}
$$

A.2 The HIV Incidence Trend

Age-specific incidence in CCMPP is modeled as follows:

$$i_{a,k,t} = 1 - \exp\{-\Gamma_{t,\,t_0}\ H\ j_{a,k}\} \tag{A-16}$$

where $i_{a,k,t}$ is the fraction of individuals age a who will become infected over the projection interval. $\Gamma_{t,\,t_0}$ represents the shape of the incidence trend from the start of the epidemic in year t_0 to the projection period t. This incidence trend is shifted up or down by H, an overall scale parameter for the epidemic. Finally, $j_{a,k}$ is the incidence ratio comparing those of age a and sex k to women age 25-29. The details of the incidence trend, $\Gamma_{t,\,t_0}$, are described in this section, along with several other possible specifications.

The incidence trend $\Gamma_{t,\,t_0}$ used by Heuveline (2003) is borrowed from `EpiModel`, a computer program developed by the World Health Organization to make short-term projections of adult AIDS cases (Chin and Lwanga, 1991) (the precursor of EPP, UNAIDS' estimation and projection package software used to estimate the prevalence of HIV (Ghys et al., 2004)), and is based on the gamma family of distributions:

$$g(t) = \frac{t^{\alpha-1}e^{-x/\beta}}{(\alpha-1)!\beta^{\alpha}}, \quad \text{for } t \geq 0,\ \alpha > 0\ \beta > 0 \tag{A-17}$$

The α parameter is typically referred to as the shape parameter since it affects how peaked the density is. The scale parameter β is associated with how diffuse or spread out the density is.

Before discussing the calculations made by Heuveline (2003) it is helpful to discuss a particular property of the gamma distribution. The mode, or the the value for which the the function reaches its maximum, is equal to $(\alpha-1)\beta$. This quantity has a nice interpretation in that it is the number of years after the start of the epidemic t_0 when the epidemic peaks. For example, if the shape parameters is 5 and the scale parameter is 3, then the epidemic peaks 12 years after it began[10].

This is precisely the density used by Heuveline (2003) to calculate the five-year incidence rates (i.e. $\alpha = 5, \beta = 4$). The actual rates are calculated by integrating the gamma density over the appropriate five-year span, i.e:

$$\Gamma_{5-t_0} = \int_0^5 \frac{t^{5-1}e^{-x/4}}{(5-1)!4^5}dt \tag{A-18}$$

$$\Gamma_{10-t_0} = \int_5^{10} \frac{t^{5-1}e^{-x/4}}{(5-1)!4^5}dt \tag{A-19}$$

$$\Gamma_{15-t_0} = \int_{10}^{15} \frac{t^{5-1}e^{-x/4}}{(5-1)!4^5}dt \tag{A-20}$$

$$\Gamma_{20-t_0} = \int_{15}^{20} \frac{t^{5-1}e^{-x/4}}{(5-1)!4^5}dt \tag{A-21}$$

[10] Chin and Lwanga (1991) report that setting $\alpha = 5$ results in "the best empirical 'fit' to the reported AIDS-case curves in countries with reliable case-reporting systems." They set $\beta = 1$.

Table A-1. Five-Year Incidence Rates Calculated from the Gamma Density and an Exponential Curve.

Time Period	$\Gamma_{t,\,t_0}$	$E_{t,\,t_0}$
0 - 5 years	0.028	0.063
6 - 10 years	0.216	0.191
11 - 15 years	0.316	0.323
16 - 20 years	0.235	0.457
20+ years[a]	0.163	0.540

[a]For the gamma model it is assumed that the HIV incidence rate will level off at the rate equal to the integral of the gamma density from 20 and 21 multiplied by five. For the exponential model it is assumed that the HIV incidence rate will level off at a rate equal to $5 * (h(t = 21) - h(t = 20))$. See the text for the definition of $h(t)$.

$$\Gamma_{20+-t_0} \quad = \quad 5 \times \int_{20}^{21} \frac{t^{5-1} e^{-x/4}}{(5-1)! 4^5} dt \tag{A-22}$$

The five-year incidence rate for twenty years after the start date is different because the decline in the gamma density for values greater than 20 is too rapid to represent an actual decline in incidence. The values used to estimate the CCMPP parameters are presented in Table A-1.

We will now discuss two other possible specifications for the trend in HIV incidence. The first is an exponential curve that models a continual increase in HIV over time. While this may not be realistic in the long run, it does provide an upper bound for the trend. A reasonable lower bound is a constant rate of new infections (i.e. no change) over time. Since the second specification is simply a constant[11] we will focus our attention on the exponential model.

The exponential curve used to model the trend in HIV incidence takes the following functional form:

$$h(t) = \frac{e^{\beta t}}{\beta} - t \quad \text{for } \; t = 1, 2, 3, \ldots; \;\; \text{and } \; \beta > 0 \tag{A-23}$$

The five-year HIV incidence rates are calculated by differencing $h(t)$ (at lag one) and summing over the five year period of interest:

$$E_{t,\,t_0} = \sum_{j=1}^{5} h(t) - h(t-1) \tag{A-24}$$

Heuveline (2003) chooses a value of $\beta = 0.005$ to obtain the five-year HIV incidence rates based on the exponential model, shown in Table A-1. Finally, similar to the gamma model described above, the incidence rate after twenty years takes a different form. For the exponential curve, $E_{20+,t_0} = 5 \times (h(t = 21) - h(t = 20))$.

[11]Heuveline (2003) uses 0.2 as the five-year incidence rate.

A.3 Additional HIV-related Force of Mortality

In the HIV-enabled CCMPP individuals infected with HIV ($d > 1$) experience an additional force of mortality that is not experienced by those in the HIV– state ($d = 1$). This mortality differential can be seen in the following projection equations

$$n_{a+1,d=2,t+1} = n_{a,d=1,t}\ s_{a,d=1,t}\ i_{a,t}\ s_{a,d=2} \quad \text{(A-25)}$$

$$n_{a+1,d>2,t+1} = n_{a,d-1,t}\ s_{a,d=1,t}\ s_{a,d>2} \quad \text{(A-26)}$$

where $s_{a,d>1} < 1$; survival in the HIV+ states is reduced compared to the HIV– state. Recall that the vital rates in CCMPP are treated as fixed parameters and need to be set by the user.

Heuveline turns to the epidemiological literature for guidance on choosing values for the survival rates of individuals infected with HIV. One of the more important findings (concerning the HIV-enabled CCMPP) is that that the progression from HIV to death is faster for older individuals. Morgan et al. (2002) report this finding in a study of a cohort from rural Uganda for whom the time of HIV infection is reasonably well known. Their data include 10, 18, and 19 deaths among 65, 68, and 35 participants in age groups 15-24, 25-39, and 40+ years. The cumulative probability of survival for each group is 79% (95%CI: 63-88%), 72% (95%CI: 56-83%), and 20% (95%CI: 6-40%), respectively. Morgan et al. (2002) also report a faster progression from seroconversion to AIDS for the oldest age group (40+ years)[12].

Given the evidence from the epidemiological literature, Heuveline (2003) specifies the survival rates for HIV+ individuals as a function of age at infection. This dependence comes through in the choice of a particular survival schedule, defined by the median number of years lived after infection. The schedules include median survival times of 3, 8, 10, and 12 years. Children who are infected perinatally follow the 3 year schedule, while the oldest age groups follow the 8-year schedule. Before describing the age dependence in greater detail, it is helpful to take a slight digression and define some more notation.

Heuveline (2003) defines these survival schedules with reference to the projection interval (i.e. five years). Let $y_{d,m}$ be the expected number of years lived by an individual in duration group d following survival schedule m, where $m = 3, 8, 11, 12$. For example, the average number of years lived by a person infected 5-9 years ago who is following the survival schedule with a median survival time of 11 years is $y_{d=3,m=11} = 3.375$. The values for $y_{d,m}$ are listed in Table A-2.

Now we are in a position to define $s_{a,d>1}$. Let us begin with those who have been infected for 0-4 years ($d = 2$). Children born and infected (perinatally) during the projection interval are exposed to:

$$s_{a=1,d=2} = \frac{y_{d=2,m=3}}{5} \quad \text{(A-27)}$$

The HIV-related survival ratios for the next two age groups are defined to be 1.0 because persons between the of ages 5 and 9 are not able to be infected given our current assumptions about incidence. Recall that the age-specific incidence rates are zero for the first three age

[12] In this same study, the median time from seroconversion to AIDS is 9.4 years, IQR: 5.5 – 10.1 years. The median time from seroconversion to death is 9.8 years, IQR: 6.1 – > 10.3 years (Morgan et al., 2002).

Table A-2. Expected Number of Person-Years Lived Over a Five-Year Interval by Survival Schedule and Duration Group.

		Survival Schedule			
Duration Group		**3**	**8**	**11**	**12**
0 to 0-4	(d=2)	2.7750	4.7100	4.8000	4.8310
0-4 to 5-9	(d=3)	0.4250	2.4300	3.3750	3.6000
5-9 to 10-4	(d=4)	0.0000	0.8600	2.0000	2.4125
10-4 to 15-9	(d=5)	0.0000	0.3150	1.0000	1.5375

groups. For those who are ages 15-19, 25-34, or above age 45, the additional force of mortality caused by HIV takes a form similar to the equation just above:

$$s_{a=4,d=2} = \frac{y_{d=2,m=12}}{5} = 0.9662 \quad \text{(A-28)}$$

$$s_{8\geq a\geq 6,d=2} = \frac{y_{d=2,m=11}}{5} = 0.9600 \quad \text{(A-29)}$$

$$s_{a\geq 10,d=2} = \frac{y_{d=2,m=8}}{5} = 0.9420 \quad \text{(A-30)}$$

Note how survival declines as age increases before leveling off at age 45 and above. For the age groups not already mentioned, the survival parameters are calculated by taking the average over two adjacent survival schedules:

$$s_{a=5,d=2} = \frac{y_{d=2,m=11}+y_{d=2,m=12}}{2} = 0.9631 \quad \text{(A-31)}$$

$$s_{a=9,d=2} = \frac{y_{d=2,m=8}+y_{d=2,m=11}}{2} = 0.9651 \quad \text{(A-32)}$$

We now turn our attention to the third duration group, individuals who have been infected for 5-9 years. For this group, we start with those ages 5-9:

$$s_{a=2,d=3} = \frac{y_{d=3,m=3}}{y_{d=2,m=3}}. \quad \text{(A-33)}$$

The corresponding parameters for the older age groups take a similar form. The expected number of years lived for the third duration group is divided by the expected number of years lived by the second duration group (for a given survival schedule). The dependence on age for $d = 3$ takes the same form as for the previous duration group, only that the age groups are incremented by one. The actual equations are:

$$s_{a=5,d=3} = \frac{y_{d=3,m=3}}{y_{d=2,m=3}} \quad \text{(A-34)}$$

$$s_{a=6,d=3} = \frac{y_{d=3,m=12}+y_{d=3,m=11}}{y_{d=2,m=12}+y_{d=2,m=11}} \quad \text{(A-35)}$$

$$s_{9\geq a\geq 7,d=3} = \frac{y_{d=3,m=11}}{y_{d=2,m=11}} \quad \text{(A-36)}$$

$$s_{a=10,d=3} = \frac{y_{d=3,m=11}+y_{d=3,m=8}}{y_{d=2,m=11}+y_{d=2,m=8}} \quad \text{(A-37)}$$

$$s_{a\geq 11,d=3} = \frac{y_{d=3,m=8}}{y_{d=2,m=8}} \quad \text{(A-38)}$$

Table A-3. Survival Probabilities Applied to HIV+ ($s_{a,d>1}$).

	HIV Duration Group			
Age Group	**0-4 yrs** ($d = 2$)	**5-9 yrs** ($d = 3$)	**10-4 yrs** ($d = 4$)	**15+ yrs** ($d = 5$)
0-4	0.5550	–	–	–
5-9	–	0.1532	–	–
10-4	–	–	0.0000	–
15-9	0.9662	–	–	0.0000
20-4	0.9631	0.7452	–	–
25-9	0.9600	0.7242	0.6701	–
30-4	0.9600	0.7031	0.6326	0.6373
35-9	0.9600	0.7031	0.5926	0.5751
40-4	0.9510	0.7031	0.5926	0.5000
45-9	0.9420	0.6104	0.5926	0.5000
50-4	0.9420	0.5159	0.4927	0.5000
55-9	0.9420	0.5159	0.3539	0.4598
60-4	0.9420	0.5159	0.3539	0.3663
65-9	0.9420	0.5159	0.3539	0.3663
70-5	0.9420	0.5159	0.3539	0.3663
75-9	0.9420	0.5159	0.3539	0.3663
80+	0.9420	0.5159	0.3539	0.3663

The pattern continues for the fourth and fifth duration groups. All of the parameters for the additional force of mortality due to HIV are listed in Table A-3.

Life tables can also be constructed using the survival rates presented in Table A-3, where cohorts defined by age at infection are exposed to the survival rates aligned along the diagonal cells of the table. For example children infected by their mothers will never reach age 15 years because the survival probability is zero between the ages of 10 and 14. The mortality experienced by the cohort of women infected at age 15 is summarized in the life table presented in Table A-4.

Table A-4. Life Table for HIV+ Women Infected at Age 15.

Age Group	${}_np_x$	${}_nq_x$	l_x	${}_nd_x$	${}_nL_x$	T_x	$\overset{0}{e}$
15	0.9566	0.0434	100,000	4,345	489,138	1,612,020	16.1202
20	0.7347	0.2653	95,655	25,381	414,822	1,122,882	11.7389
25	0.6587	0.3413	70,274	23,983	291,412	708,060	10.0757
30	0.6255	0.3745	46,291	17,338	188,110	416,648	9.0006
35	0.5634	0.4366	28,953	12,642	113,160	228,538	7.8934
40	0.4883	0.5117	16,311	8,346	60,690	115,378	7.0736
45	0.4863	0.5137	7,965	4,092	29,595	546,88	6.8660
50	0.4823	0.5177	3,873	2,005	14,352	25,093	6.4789
55	0.4370	0.5630	1,868	1,052	6,710	10,741	5.7498
60	0.3388	0.6612	816	540	2,730	4,031	4.9395
65	0.3220	0.6780	276	187	912	1,301	4.7125
70	0.2940	0.7060	89	63	288	389	4.3669
75	0.2500	0.7500	26	20	80	101	3.8713
80	0.1586	0.8414	6	6	21	21	3.4425

The size of the cohort exposed to the mortality risks in Table A-4 is reduced to half after fifteen years. After thirty years there is just under ten percent of the cohort still living.

A.4 Matrix Notation for HIV-enabled CCMPP

These equations for the multi-state, HIV-enabled CCMPP can be conveniently expressed in matrix notation. For a population with 17 age groups and five HIV duration groups, the

population at time t is represented by an 85×1 column vector

$$\mathbf{n}_t = \left[\begin{array}{c} n_{1,1,t} \\ n_{2,1,t} \\ \vdots \\ n_{17,1,t} \\ \hline \vdots \\ \hline n_{1,4,t} \\ n_{2,4,t} \\ \vdots \\ n_{17,4,t} \end{array}\right] \tag{A-39}$$

The corresponding *Leslie* matrix is:

$$\mathbf{A}_t = \begin{bmatrix} \mathbf{B}_{1,1} & \mathbf{B}_{1,2} & \mathbf{B}_{1,3} & \mathbf{B}_{1,4} & \mathbf{B}_{1,5} \\ \mathbf{B}_{2,1} & \mathbf{B}_{2,2} & \mathbf{B}_{2,3} & \mathbf{B}_{2,4} & \mathbf{B}_{2,5} \\ \mathbf{0} & \mathbf{B}_{3,2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_{4,3} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{B}_{5,4} & \mathbf{B}_{5,5} \end{bmatrix} \tag{A-40}$$

where $\mathbf{B}_{i,j}$ is a 17×17 sub-matrix that models how group j at time t contributes to group i at time $t+1$. Note that $\mathbf{B}_{3,1}$ is a zero matrix since women who are HIV– at time t cannot give birth to children who have been HIV positive for ten years by $t+1$ (i.e. five years into the future). Similar reasoning applies for the other zero matrices.

The calculations involving $\mathbf{B}_{1,j}$ produce the projection for the number of HIV– births (i.e. $n_{1,1,t+1}$) contributed by duration group j. Similarly, $\mathbf{B}_{2,j>1}$ projects the number of HIV positive births contributed by duration group $j > 1$. $\mathbf{B}_{1,1}$ and $\mathbf{B}_{2,1}$ are a little different in that they project each age group to the next oldest age group *and* from one HIV duration group to the next. Let us first consider $\mathbf{B}_{1,1}$:

$$\mathbf{B}_{1,1} = \begin{bmatrix} b^-_{1,1,t} & b^-_{2,1,t} & \cdots & & & b^-_{17,1,t} \\ p^-_{1,1,t} & 0 & \cdots & & & 0 \\ 0 & p^-_{2,1,t} & \ddots & & & \vdots \\ 0 & 0 & \ddots & & & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & p^-_{16,1,t} & p^-_{17,1,t} \end{bmatrix}. \tag{A-41}$$

Recall that the number in the first age group at time $t+1$ is equal to the number of births summed across the fecund age groups. Let $b^-_{a,d,t}$ be the factor needed to calculate the number

of HIV– births to mothers in age group a at time t and in duration group d:

$$b^{-}_{a,1,t} \;=\; s_{0,1,t}\,\frac{1}{1+SRB}\,f^{-}_{a,1,t}\,\frac{1+p^{-}_{a-1,1,t}\frac{n_{a-1,1,t}}{n_{a,1,t}}}{2}. \tag{A-42}$$

In our application of the multi-state, HIV-enabled CCMPP fertility only occurs among women aged 15-49 (i.e. $\alpha = 4,\ \beta = 10$). Consequently $b^{-}_{a<4,1,t} = b^{-}_{a>10,1,t} = 0$. In the equation above the factor $\frac{n_{a-1,1,t}}{n_{a,1,t}}$ is used to approximate the number of women at risk of giving birth. If the count in the denominator $n_{a,1,t}$ is ever zero the entire ratio is simply replaced by zero. This issue arises when dealing with fertility of the HIV+ groups. The same procedure is used in the analogous HIV+ equations if they involve dividing by zero.

$\mathbf{B}_{1,d}$ for $d > 1$ projects forward HIV– births contributed by duration group d and can be written as:

$$\mathbf{B}_{1,d} = \begin{bmatrix} b^{-}_{1,d,t} & b^{-}_{2,d,t} & b^{-}_{3,d,t} & \cdots & b^{-}_{17,d,t} \\ 0 & \cdots & & & 0 \\ \vdots & \ddots & & & \\ & & & & \\ 0 & & & & 0 \end{bmatrix} \tag{A-43}$$

where

$$b^{-}_{a,d,t} \;=\; s_{0,1,t}\,\frac{1}{1+SRB}\,f^{-}_{a,d,t}\,\frac{1+p_{a-1,d-1,t}(\frac{n_{a-1,d-1,t}}{n_{a,d,t}})}{2} \tag{A-44}$$

for $d > 1$. The $\mathbf{B}_{2,d}$'s determine the number of people infected with HIV for less than five years at time $t+1$, contributed by those in duration group d at time t. For the first two duration groups we have:

$$\mathbf{B}_{2,1} = \begin{bmatrix} b^{+}_{1,1,t} & b^{+}_{2,1,t} & \cdots & & & b^{+}_{17,1,t} \\ p_{1,1,t} & 0 & \cdots & & & 0 \\ 0 & p_{2,1,t} & \ddots & & & \vdots \\ 0 & 0 & \ddots & & & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & p_{16,1,t} & p_{17,1,t} \end{bmatrix}. \tag{A-45}$$

The model does now allow HIV– women to become infected and give birth to HIV+ children in the same projection interval, zeros in the first rows.

$\mathbf{B}_{2,d}$ for $d > 1$ projects forward the number of HIV+ births contributed by duration group d. It can be written as:

$$\mathbf{B}_{2,d} = \begin{bmatrix} b^{+}_{1,d,t} & b^{+}_{2,d,t} & b^{+}_{3,d,t} & \cdots & b^{+}_{17,d,t} \\ 0 & \cdots & & & 0 \\ \vdots & \ddots & & & \\ & & & & \\ 0 & & & & 0 \end{bmatrix} \tag{A-46}$$

where

$$b^{+}_{a,d,t} \;=\; s_{0,1,t}\; s_{0,1,t}\; \frac{1}{1+SRB}\; f^{+}_{a,d,t}\; \frac{1+p_{a-1,d-1,t}(\frac{n_{a-1,d-1,t}}{n_{a,d,t}})}{2}. \tag{A-47}$$

The remaining non-zero sub-matrices – $\mathbf{B}_{3,2}, \mathbf{B}_{4,3}, \mathbf{B}_{5,4}$ and $\mathbf{B}_{5,5}$ – project people forward in both age and time, and consequently the only non-zero elements occur along the sub-diagonal:

$$\mathbf{B}_{i,j} = \begin{bmatrix} 0 & 0 & \cdots & & & 0 \\ p_{1,d=j,t} & 0 & \cdots & & & \vdots \\ 0 & p_{2,d=j,t} & \ddots & & & \\ 0 & 0 & \ddots & & & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 0 & p_{16,d=j,t} & p_{17,d=j,t} \end{bmatrix} \tag{A-48}$$

The Leslie matrix representation of the model greatly facilitates the implementation and use of the multi-state, HIV-enabled CCMPP using the `R` programming language (R Foundation for Statistical Computing, 2010). Repeated matrix multiplication produces the projected population at five-year intervals making it possible to explore the long term behavior of the population and the epidemic.

B PARAMETER ESTIMATION

The CCMPP projections are used to estimate thirty-two of the model parameters. As mentioned earlier the vital rates, the initial population counts, and the HIV survival schedules are all fixed. The parameters we are interested in estimating are:

- v: vertical transmission parameter that is constrained to be between 0 and 1; although the model is described as having a vertical transmission rate for each duration group, there are not enough data to estimate separate parameters – (1)
- e_a: fertility selection parameter that is constrained to be equal to 1 for all groups except women aged 15-19 in the first HIV duration group, for whom we expect the value to be greater than 1 – (1)
- g_d: fertility impairment parameter for women in duration group d, for $d = 2, 3, 4$; the fertility impairment parameter for $d = 5$ is constrained such that $g_{d=5} = g_{d=4}$, and the values for all duration groups are constrained to be between 0 and 1 – (3)
- $j_{a,k}$: relative incidence rate ratio parameter that is constrained to be equal to 1 for women age 25-29 and non-negative for all other groups; values are estimated for women ($k = 1$) age $15-19, 20-24, 30-34, 35-39, \ldots, 55-59$ and for men in the age groups between 15-59 (i.e. 8 and 9 age-specific parameters for women and men, respectively) – (17)

- H_h: scale parameters for the trend in HIV incidence for population h – (11)

For a given set of parameter values we obtain a set of not necessarily unique sex-age-specific counts. These model outputs are used to calculate predicted values for the observed data. For example the ratio of the projected number of HIV+ women age 20-25 over the total number of women projected in that age group is used to predict the observed HIV prevalence for women in that age group. Several types of observed data, such as HIV prevalence, are used to estimate the values of the parameters that are most likely.

In this section we discuss this topic in greater detail. The first focus is on the types of data used in the analysis. We then shift to the likelihoods specific to each type of data. Finally we turn to the techniques used to estimate the parameters, namely maximum likelihood (ML) estimation.

B.1 Data Types

Heuveline (2003) uses data published in the literature to estimate the model parameters. These data consist of observations from eleven different East African populations collected from antenatal clinics (ANCs), demographic surveillance sites, hospitals and general surveys. Both rural and urban areas are included, and the years of data collection range from 1989 to 1998. The data are classified into the following five categories (see Table 1, Heuveline, 2003):

1. HIV test results in a general-population sample (10 data sets)
2. HIV test results in an ANC-patient sample (3 data sets)
3. HIV test results in all or a sample of births from HIV+ mothers (3 data sets)
4. HIV test results during a follow-up of an HIV– sample (3 data sets)
5. Survival during a follow-up of HIV+ individuals (3 data sets)

All of the data sources are listed in Table B-1, with information about the corresponding location and year of data collection.

It should also be mentioned that the CCMPP also requires vital rates for the uninfected population and an initial age distribution. These model inputs are taken from the United Nations global population estimates (United Nations, 1999) and model life tables (United Nations, 1982).

B.2 Likelihoods

Each category of data provides the pieces needed for a proportion which leads to the use of the binomial distribution in the likelihood specification. The binomial likelihood can be

Table B-1. Descriptions for each source that contributes data used in the estimation of the CCMPP parameters. Data types: (1) HIV test results in a general-population sample; (2) HIV test results in an ANC-patient sample; (3) HIV test results in all or a sample of births from HIV+ mothers; (4) HIV test results during a follow-up of an HIV– sample; (5) Survival during a follow-up of HIV+ individuals.

Country	City/ Town	Data Type	Citation	Year of Data Colllection	Start Year of Epidemic	Urban/ Rural
Burundi	Bujumbura	1	Saidel et al. (1996)	1990	1973	Urban
	Bujumbura	2	Sokal et al. (1993)	1991-2	1973	Urban
	Bujumbura	4	Saidel et al. (1996)	1991	1973	Urban
Tanzania	Mara	1	Shao et al. (1994)	1990	1975	Urban
	Mwanza	1	Grosskurth et al. (1995)	1992	1975	Rural
	Mwanza	5	Todd et al. (1997)	1994	1975	Rural
Uganda	Fort Portal	1	Kilian et al. (1999)	1995	1975	Urban
	Gulu	2	Fabiani et al. (2001)	1993, 1997	1975	Rural
	Masaka	1	Nunn et al. (1994)	1989	1975	Rural
	Masaka	3	Carpenter et al. (1997)	1990	1975	Rural
	Masaka	4	Kengeya-Kayondo et al. (1996)	1990-4	1975	Rural
	Masaka	5	Nunn et al. (1997)	1990	1975	Rural
	Rakai	1	Wawer et al. (1991)	1989-91	1980	Rural
	Rakai	4	Wawer et al. (1994)	1990	1980	Rural
	Rakai	1	Serwadda et al. (1992)	1989-91	1980	Rural
	Rakai	3	Gray et al. (1998)	1995	1980	Rural
	Rakai	5	Sewankambo et al. (1994)	1980	1990	Rural
	1995	1975	Urban			
Zambia	Chelston	1	Fylkesnes et al. (1998)	1995	1975	Urban
	Kapiri Mposhi	1	Fylkesnes et al. (1998)	1996	1980	Rural
	Lusaka	3	Hira et al. (1989)	1987	1975	Urban
Zimbabwe	Mutasa	1	Gregson and Garnett (2000)	1998	1977	Rural

written as:

$$\mathcal{L} = \prod \binom{N}{x} \pi^x (1-\pi)^{N-x} \tag{B-1}$$

where N is the total number of events, x is the number of "successes", and π is the probability of success. The first two quantities N and x are taken from the data and we use CCMPP to calculate π given values for the parameters.

Before discussing the finer details of how the CCMPP outputs are used in the likelihood, it is important to cover two points. First, we need to temporally match up the CCMPP projections with the observed data. The start year for the model is the year when widespread transmission of HIV began, t_0. The population is then projected forward to the year when the data were collected. For example if widespread transmission in a country began in 1980 and the observed data are from 2000, then we can take the projected counts 20 years from the start time and compare these to the observed data. Given that the projections are in five year increments it is sometimes necessary to take the average across two projection periods to match the year of data collection. Estimates of when widespread HIV transmission began are taken from a report by the United Nations (1998, Table 1).

Second, the data come from populations at eleven different locations[13]. At a given location there can also be several different types of data. For example data from the population living in Mwanza, Tanzania include both HIV prevalence from a general population and survival information for those who are HIV+. As a result there is a separate likelihood for the 23 combinations of location and data type retrieved from the literature. These are indicated using h for the population (and location) and c for the type or category of data. Finally, the data consist of sex- and age-specific information so the likelihoods may also be indexed by these characteristics as well.

B.2.1 HIV Test Results in a General-Population Sample

Various studies have collected data on sex- and age-specific HIV prevalence in a sample from the general population (Kilian et al., 1999; Nunn et al., 1997; Wawer et al., 1991; Serwadda et al., 1992; Shao et al., 1994; Grosskurth et al., 1995; Saidel et al., 1996; Fylkesnes et al., 1998; Gregson and Garnett, 2000). The age groups range from 15-19 to 55-59. This type of data, labeled '1' ($c = 1$), usually include the number of people tested and the percent who tested positive for HIV by sex and age. The likelihood, however, requires a count of individuals who are HIV+, so we calculate this quantity from the data and round it to the nearest integer. Let $N_{a,k,t,h,c=1}$ denote the total number of individuals in age group a of sex k at time t at location h and let $x_{a,k,t,h,c=1}$ be the number in this group who tested positive.

$n_{a,d,t,h}$ the projected counts from CCMPP are used to predict sex- and age-specific prevalence for a given location as follows:

$$\pi_{a,k,t,h,c=1} = \frac{\sum_{d=2}^{5} n_{a,d,t,h}}{\sum_{d=1}^{5} n_{a,d,t,h}}, \tag{B-2}$$

[13]Bujumbura, Burundi; Mangochi, Malawi; Mara, Tanzania; Mwanza, Tanzania; Fort Portal, Uganda; Gulu, Uganda; Masaka, Uganda; Rakai, Uganda; Lusaka, Zambia; Mposhi, Zambia and Mutasa, Zimbabwe.

where the sum is taken across HIV duration groups. Having chosen the projection period that matches up with the year the data were observed, we can use $\pi_{a,k,t,h,c=1}$ in the binomial likelihood. A final note is that the observed data may be reported by age groups that do not match those of our projections. In this case weighted sums of the projected counts can be used to estimate HIV prevalence. For example if observed prevalence is reported for individuals age 22-30, then predicted prevalence can be calculated as:

$$\pi_{age=(22-30),k,t,h,c=1} = \frac{\sum_{d=2}^{5}(n_{a=4,d,t,h} \times 0.6 + n_{a=5,d,t,h})}{\sum_{d=1}^{5}(n_{a=4,d,t,h} \times 0.6 + n_{a=5,d,t,h})} \tag{B-3}$$

This issue arises with the other data types as well.

B.2.2 HIV Test Results in an ANC-Patient Sample

Four of the data sets used in this analysis provide age-specific information on HIV prevalence for female attendees of ANCs, typically from age 15 to 49 (Kilian et al., 1999; Fabiani et al., 2001; Slutsker et al., 1994; Sokal et al., 1993). This type of data, indexed by $c = 2$, takes a form similar to the observed prevalence from a general population except that they only refer to women. Both the total number of women tested $N_{a,k=1,t,h,c=2}$ and the age-specific prevalence are reported. The data are included in the binomial likelihood as counts, so we calculate the number of women who tested positive $x_{a,k=t,h}$ rounded to the nearest integer.

The predicted prevalence for the ANC attendees is calculated differently than for the general population. Recall that there are two primary assumptions of CCMPP concerning the fertility of HIV+ women. The first is that HIV+ women age 15-19 have higher fertility which is captured by the fertility selection parameter $e_{a=4}$. Second, fertility is expected to decline as time since infection increases; modeled by the fertility impairment parameters $g_{d>1}$. Since the HIV+ women observed in the data are pregnant, these parameters are included in the calculation of predicted prevalence. The formula is:

$$\pi_{a,k=1,t,h,c=2} = \frac{\sum_{d=2}^{5} n_{a,d,k=1,t,h} \times e_a \times g_d}{n_{a,d,k=1,t,h} + \sum_{d=2}^{5} n_{a,d,k=1,t,h} \times e_a \times g_d}, \tag{B-4}$$

where the sum is taken over the duration groups. Having chosen the projection period that matches up with the year the data were observed, we can use $\pi_{a,k=1,t,h,c=2}$ in the binomial likelihood.

B.2.3 HIV Test Results in all or a Sample of Births from HIV-Positive Mothers

Heuveline (2003) found three data sets consisting of information on the fertility of HIV+ women. However one of these sources, Hira et al. (1989), differs from the others in that it provides information on whether or not an HIV+ mother infected her child. Data from the other two sources, Carpenter et al. (1997) and Gray et al. (1998), consist of the number of children born to both HIV+ and HIV– women, by age group. The likelihoods for the latter two sources are nearly identical to those in the data category $c = 2$, HIV test results in an

ANC sample. The only difference is that the observed counts (i.e. the data) refer to the total number of children born to female ANC attendees in a specific age group, and the number of children born to HIV+ attendees. The probability that a child is born to an infected mother is calculated exactly the same as $\pi_{a,k=1,t,h,c=2}$. Given this similarity the data reported by Carpenter et al. (1997) and Gray et al. (1998) are classified here as $c = 2$.

Data that take the form of Hira et al. (1989) will be indicated by $c = 3$. The corresponding counts used in the likelihood refer to the total number of children born to infected mothers in age group o, $N_{o,t,h,c=3}$, and the number of these children who are infected by their mothers $x_{o,d=2,t,h,c=3}$. The predicted rate of vertical transmission using the model outputs is calculated as:

$$\pi_{o,t,h,c=3} = \frac{\sum_{d=2}^{5} n_{a,k=1,d,t,h} \times f_{a,d=1,t,h} \times e_a \times g_d \times v_d}{\sum_{d=2}^{5} n_{a,k=1,d,t,h} \times f_{a,d=1,t,h} \times e_a \times g_d}, \tag{B-5}$$

where the sum is taken over the duration groups.

Unfortunately, this leaves only one data source to inform the estimate of the vertical transmission parameter. This issue is especially problematic when the parameter is constrained to be equal across duration groups (i.e. $v_d = v$, for all d). Note that all the like terms in the numerator and denominator cancel out, so the projections have no influence on the likelihood. Thus including more data of this type would be beneficial for future analysis using this model.

B.2.4 HIV Test Results During a Follow-up of an HIV-Negative Sample

Data on sex- and age-specific HIV incidence are also used to estimate CCMPP parameters. These data indexed by $c = 4$ are typically reported in terms of the number of people who become infected and the total number of person-years lived while uninfected (Kengeya-Kayondo et al., 1996; Wawer et al., 1994; Saidel et al., 1996). For the binomial likelihood however, we need the counts of the initial population observed $N_{a,t,d=1,t,h,c=4}$ and the number of who become infected $X_{a,t,d=2,t,h,c=4}$. Thus the initial population size needs to be calculated from the observed data, and this calculation can be done as follows:

$$\text{Initial Population} = \frac{\text{\# Converted}}{1 - \exp\left\{-T \times \frac{\text{\# Converted}}{\text{Person-Years}}\right\}} \tag{B-6}$$

where T is the total number of years observed[14].

The model outputs are then used to calculate the probability of becoming infected for men or women in a certain age group at a given location and time. That is:

$$\pi_{a,k,d=1,t,h,c=4} = \frac{n_{a,k,d=1,t,h} \times s_{a,k,d=1,t,h} - n_{a+1,k,d=1,t+1,h}}{n_{a,k,d=1,t,h} \times s_{a,k,d=1,t,h}} \tag{B-7}$$

[14]In deriving this equation it is helpful to note that the number of person-years lived by a population of initial size N_0 and of size N_T T years later is equal to $\frac{(N_T - N_0) \times T}{\log(N_T/N_0)}$.

where $\pi_{a,k,d=1,t,h,c=4}$ is the proportion of HIV– women/men who become infected after five years. As discussed earlier, the period of observation for the data may not be equal to the projection interval of five years. If the observation period is only four years, then the quantity of interest is calculated as:

$$1 - \exp\left\{\frac{4}{5} \times \log\left(1 - \frac{\#\text{ Converted}}{\text{Initial Population}}\right)\right\} \tag{B-8}$$

B.2.5 Survival During a Follow-up of HIV+ Individuals

The final category of data describes the survival (mortality) of HIV+ individuals Nunn et al. (1997); Sewankambo et al. (1994); Todd et al. (1997). This category indexed by $c = 5$ is similar to the previous one in that the data reported include the number of deaths observed among a cohort of HIV+ individuals of a particular age and sex $X_{a,k,t,h,c=5}$ and the number of person-years observed for each group. As before, the likelihoods require the initial population size for each group $N_{a,k,t,h,c=5}$ (see Equation B-6). These two inputs $X_{a,k,t,h,c=5}$ and $N_{a,k,t,h,c=5}$ are the counts needed for the binomial likelihood with the corresponding proportion referring to the probability of death over a given period of time.

The procedure for calculating the probability of death from the model outputs is best described in two steps. First, we calculate the probabilities by age, sex, and duration group. This can be written as:

$$q_{a,k,d,t,h,c=5} = 1 - \left(\frac{n_{a+1,k,d+1,t+1,h}}{n_{a,k,d,t,h}}\right)^{\frac{T}{5}} \quad \text{for} \quad d \geq 2 \tag{B-9}$$

where T is the number of years over which the data were observed. Since the observed data do not contain information on duration group, we must calculate the weighted average where the weights are the counts in each duration group. This second step is performed as follows:

$$\pi_{a,k,t,h,c=5} = \frac{\sum_{d=2}^{5} q_{a,k,d,t,h,c=5} \times n_{a,k,d,t,h}}{n_{a,k,d,t,h}} \tag{B-10}$$

This is the probability used in the binomial likelihood.

B.3 Parameter Estimation

A maximum likelihood (ML) approach is used to estimate the most likely parameter values (given the data and the mode) and the uncertainty around those point estimates. Given the data from an individual site, the likelihood of a specific set of CCMPP parameter values can be calculated using the binomial expressions described above. There are twenty-three likelihoods in the original data compilation, one for each of the twenty-three locations from which data come. We follow Heuveline (2003) and combine these likelihoods by taking the product across all twenty-three locations and data types, assuming independence. The set of parameter values that maximizes the combined likelihood is the ML point estimate.

The ML estimation as well as the implementation of CCMPP is performed using the `R` programming language (R Foundation for Statistical Computing, 2010). This languages provides an optimization routine `optim` that is used to find the parameter values that maximize the combined likelihood described above. In addition `optim` calculates the Hessian matrix of the likelihood function at the maximum. Standard errors are obtained by inverting the Hessian matrix (after multiplying it by negative one) and taking the square root of the diagonal elements.