HIV Epidemic Reconstruction from Demographic Surveillance and General Population HIV Cohort Data using Bayesian Penalized B-spline Models

Jeffrey W Eaton\(^1\), Anne Cori\(^1\), Georges Reniers\(^2\), Emma Slaymaker\(^2\), Mike Pickles\(^1\), Christophe Fraser\(^1\), Basia Zaba\(^2\) on behalf of the ALPHA Network

\(^1\)Department of Infectious Disease Epidemiology, Imperial College London, United Kingdom
\(^2\)Department of Population Health, London School of Hygiene and Tropical Medicine, United Kingdom

Extended abstract for submission to PAA 2016 Annual Meeting.
25 September 2015

Short Abstract: Six general-population HIV cohorts established during the 1990s have been primary data sources for the epidemiologic and demographic consequences of hyper-endemic HIV in eastern and southern Africa. We fit a statistical model simultaneously to data about HIV prevalence, incidence, and mortality to reconstruct the history of HIV epidemic in each of these sites. We used penalized B-splines to flexibly model HIV incidence, non-HIV mortality, and the effects of ART by time and age. The likelihood for individual-level longitudinal data was calculated from these functions, using an established distribution for survival by age at HIV infection. Model parameters were estimated through Bayesian inference, with posterior estimation via Hamiltonian Monte Carlo. Joint inference from prevalence and mortality data robustly identified peaks and declines in HIV incidence and resolved trends to around seven years before the establishment of surveillance. Before this, estimates relied on informative prior information about zero-prevalence before the epidemic start.
Introduction

HIV epidemics emerged in eastern and southern Africa during the 1970s and 1980s and swept rapidly through populations to reach hyper-endemic levels. However, epidemics were only detected and systematically monitored after they had reached high levels and in many instances estimated to have nearly or already peaked. Most HIV epidemic estimates are generated by fitting mathematical models to surveillance data and relying on myriad demographic and epidemiologic assumptions (Brown et al. 2014; Murray et al. 2014; Stover et al. 2014).

HIV epidemics have also been a major demographic event. Because of its disproportionate burden in otherwise healthy young and middle-aged adults, temporal and age patterns of adult mortality in sub-Saharan Africa have been strongly determined by HIV. Consequently, model-based estimates of HIV-related deaths are a key input to estimates and projections of adult mortality in sub-Saharan Africa (Gerland et al. 2014; Wang et al. 2012). By the same token, data about trends in adult mortality potentially contain valuable information about the growth of HIV epidemics before the availability of direct HIV surveillance.

A primary source of data about the population-level epidemiology and demographic impacts of HIV are general population cohort studies in six specific sites in eastern and southern Africa (Maher et al. 2010). These studies consisted of routine demographic surveillance (measurement of all births, deaths, and migrations) in a geographically-defined area and routine HIV testing of all adults in the population (typically every one to three years). These data are heavily relied on for assumptions and parameter values used by national HIV epidemic estimates and projections, including natural survival from HIV infection to death (Todd et al. 2007), age patterns of HIV incidence (Stover et al. 2014), effects of HIV on fertility, and the population effects of antiretroviral treatment (ART) on mortality (Reniers et al. 2014). Analyses have typically treated each outcomes independently—HIV prevalence, incidence, and mortality—not leveraging the intrinsic demographic relationships between these.

We aimed to fit a statistical model simultaneously to all data about HIV prevalence, incidence, and mortality in order to more finely resolve estimates of the the epidemiologic and demographic trends as HIV epidemics spread and since the scale-up of life-saving ART. The model leverages the well-characterized distribution of survival with HIV as a function of age-at-infection to relate longitudinal measurement of individual HIV infection status and survival. Moreover since HIV prevalence and mortality represent the accumulation of previous HIV incidence, we hypothesized that it may be possible to make inference about age-specific HIV epidemic trends, with robust uncertainty, in the years before the establishment of HIV surveillance.
Methods

Data

Data were from general-population HIV cohort studies in six districts in eastern and southern Africa: Karonga, Malawi; Masaka, Uganda; Rakai, Uganda; Kisesa, Tanzania; Manicaland, Zimbabwe; and uMkhanyakude, South Africa. In each site, births, deaths, and migrations in each household were reported by proxy-respondents through demographic surveillance fieldwork visits typically every three months to one year (less frequently in Manicaland). Adults were tested for HIV through regular rounds of HIV serosurveillance, typically every one to three years, resulting in direct observation of prevalent and incident HIV infection.

Table 1 describes the year in which surveillance was established in each site, population size, and year in which ART became available. In one site, Karonga, routine demographic surveillance only began in 2005, but age-specific HIV prevalence data were available from a number of previous studies as early as 1981 (Glynn et al. 2001; White et al. 2007).

<table>
<thead>
<tr>
<th>Site</th>
<th>Country</th>
<th>Cohort established</th>
<th>HIV testing established</th>
<th>Total person-years (x1000)</th>
<th>ART available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Masaka</td>
<td>Uganda</td>
<td>1990</td>
<td>1990</td>
<td>170.9</td>
<td>2004</td>
</tr>
<tr>
<td>Rakai</td>
<td>Uganda</td>
<td>1994</td>
<td>1994</td>
<td>319.7</td>
<td>mid-2004</td>
</tr>
<tr>
<td>Kisesa</td>
<td>Tanzania</td>
<td>1994</td>
<td>1994</td>
<td>276.2</td>
<td>2005</td>
</tr>
<tr>
<td>Karonga</td>
<td>Malawi</td>
<td>2002</td>
<td>1981/2005</td>
<td>191.8</td>
<td>mid-2005</td>
</tr>
<tr>
<td>Manicaland</td>
<td>Zimbabwe</td>
<td>1998</td>
<td>1998</td>
<td>121.3</td>
<td>mid-2005</td>
</tr>
<tr>
<td>uMkhanyakude</td>
<td>South Africa</td>
<td>2000</td>
<td>2002</td>
<td>552.2</td>
<td>2004</td>
</tr>
</tbody>
</table>

Data used for this analysis consist of individual-level data about survival and HIV status for all adults (age 15+ years) resident within each study site. Entry into the population is observed via initial enumeration, ageing into the cohort, or in-migration. Exit from the population is via death, out-migration, or censoring of the cohort at most recent follow-up. Dates of entry and exit are assumed to be exactly observed. Thus survival data are left truncated (individuals are observed conditional on surviving until the time they enter the cohort) and right censored (death could occur after individuals exit the cohort).

HIV status information was collected through regular community-based HIV testing every six months to three years (depending on study site). Some individuals may not appear in each round. Thus data about HIV seroconversion may be left censored (already HIV-positive at the first observation), right censored (HIV-negative at last observation), or interval censored (unknown exact seroconversion date between observed negative and positive tests).
Analyses are conducted separately for each study site and sex. It is assumed that all seroconversion occurs after age 15. Thus, taken together the data available for an individual $i$ of a given sex in a given study site can be summarized as

$$Y_i = \{t_i^B, t_i^S, t_i^E, \delta_i, t_i^N, t_i^P, y_i\}$$

where

- $t_i^B$ = date of birth;
- $t_i^S$ = date of entry into cohort (first observed alive);
- $t_i^E$ = date of exit from cohort (death or censoring);
- $\delta_i$ = indicator variable indicating death (= 1) or censoring (= 0) at $t_i^E$;
- $t_i^N$ = start of potential seroconversion interval (either last HIV-test or age 15);
- $t_i^P$ = end of potential seroconversion interval (either first HIV+ test or $t_i^E$);
- $y_i$ = indicator of known HIV+ (=1) or censoring (=0) at $t_i^P$.

It is possible that some individuals may not have any observed HIV serostatus data, in which case $t_i^N = 15, t_i^P = t_i^E, y_i = 0$.

**Model**

The model is specified by three hazard functions over time and age: $\lambda(t, a)$ defining the HIV incidence rate at time $t$ and age $a$; $\mu(t, a)$ defining the background non-HIV mortality rate at time $t$ for age $a$; and $\rho(t, a, u)$ defining the HIV-mortality rate at time $t$, for a person of age $a$, and duration of infection $u$.

The log HIV incidence rate was modelled as a bivariate B-spline surface with a evenly spaced knots on a 5 year by 5 year lattice with coefficients $\beta_{ij}, i = 1, ..., I, j = 1, ..., J$, spanning the range of the data. Letting $\beta = \{\beta_{ij}\}$ be the matrix of spline coefficients and $B_3(x)$ define the appropriate B-spline basis function, then

$$\log \lambda(t, a) = B_3(t)\beta B_3(a)'$$

Differences between neighboring spline coefficients are penalized using the conditional densities

$$\beta_{ij}|\beta_{-ij} \sim N \left( \frac{\beta_{i-1,j} + \beta_{i,j-1} + \beta_{i+1,j} + \beta_{i,j+1}}{4}, \frac{\sigma^2_\lambda}{4} \right)$$

and adjusted appropriately at the boundaries, resulting in the improper multivariate normal prior

$$p(\beta) \propto \sigma^{-l-1}_\beta \exp \left\{ -\frac{1}{2\sigma^2_\beta} \beta' \Sigma_\beta^{-1} \beta \right\}$$

for singular precision matrix $\Sigma_\beta$ of dimension $I \times J \times I \times J$ and rank $I \times J - 1$.

Natural (non-HIV) mortality was modelled as an additive model

$$\log \mu(t, a) = g(t) + g(a)$$

where each of the functions $g(t)$ and $g(a)$ were represented by one-dimensional B-spline functions again with evenly spaced knots every 5-years, with coefficients $\gamma_i^t$ and $\gamma_j^a$. The
function $g(t)$ was defined such that non-HIV mortality is assumed to be constant before the start of mortality surveillance in each cohort, that is if demographic surveillance began at time $t_0$, then

$$g(t) = \begin{cases} g(t_0) & \text{if } t < t_0 \\ g(t) & \text{if } t \geq t_0 \end{cases}.$$ 

Thus $g(t)$ is only defined on a subset of knots $\gamma_i^t, i \in \{1, \ldots, l\}$. For identifiability of $\mu(t,a)$, $\gamma_j^a$ were constrained such that $\sum_i \gamma_i^a = 0$. First order knot differences were penalized

$$\gamma_i^t - \gamma_{i-1}^t \sim N(0, \sigma_{\gamma_t}^2),$$
$$\gamma_j^a - \gamma_{j-1}^a \sim N(0, \sigma_{\gamma_a}^2).$$

In the absence of ART, survival after HIV seroconversion was modelled using a Weibull distribution with shape parameter 2.3 and scale parameter $72.2 \cdot a_0^{-0.53}$ for age at seroconversion $a_0$, as parameterized by Bellan et al based on data from HIV seroconverter cohorts in Europe (Bellan et al. 2013; Collaborative Group on AIDS Incubation and HIV Survival 2000). After the introduction of ART at time $t_{\text{ART}}$, the effect of ART is incorporated by reducing the HIV-related mortality hazard as function of time $t > t_{\text{ART}}$. Thus $\rho(t,a,u)$ is defined as a product of a Weibull hazard function and a piecewise ART effect:

$$\rho(t,a,u) = \frac{2.3}{72.2 \cdot (a - u)^{-0.53}} \left( \frac{u}{72.2 \cdot (a - u)^{-0.53}} \right)^{2.3 - 1} \cdot h(t)$$

where

$$h(t) = \begin{cases} 1 & \text{if } t < t_{\text{ART}} \\ h_{\text{ART}}(t) & \text{if } t \geq t_{\text{ART}} \end{cases}.$$ 

The function $h_{\text{ART}}(t)$ is defined with three properties in mind: (1) continuity of $h(t)$, that is $h_{\text{ART}}(t_{\text{ART}}) = 1$, (2) monotonically decreasing over time, and (3) positivity, to ensure $\rho(t,a,u) > 0$. This is accomplished by defining $\log h_{\text{ART}}(t)$ as an integral function

$$\log h_{\text{ART}}(t) = \int_{t_{\text{ART}}}^{t} dh(t) \, dt.$$ 

The function $dh(t)$ is defined to be piecewise constant and strictly negative, such that

$$dh(t) = dh_x, \, t \in [x, x + 0.2)$$

where the constants $dh_x$ are constrained to be negative and first-order differences are penalised

$$dh_x - dh_{x-0.2} \sim N(0, \sigma_{\text{ART}}^2).$$

An improper flat prior was implicitly placed on the overall mean level of the curves $f(t,a)$, $g(t)$, $g(a)$, and $dh(t)$. Half-cauchy prior distributions are placed on each of the penalty variance components

$$\sigma_x \sim \text{half-Cauchy}(0, 2.5),$$
$$\sigma_{\gamma_t} \sim \text{half-Cauchy}(0, 2.5),$$
$$\sigma_{\gamma_a} \sim \text{half-Cauchy}(0, 2.5),$$
\[ \sigma_{\text{ART}} \sim \text{half-Cauchy}(0, 2.5). \]

\[ \text{Likelihood} \]

The likelihood is defined for each individual based on the vector \( Y_i \) defined above in terms of the three hazard functions \( \lambda(t, a), \mu(t, a), \) and \( \rho(t, a, u) \). For each individual, the likelihood calculation proceeds by calculating the probability of surviving to the exit time \( t_i^F \), conditional on surviving to entry into the cohort at time \( t_i^E \). Probability of survival to \( t_i^F \) depends on the observed HIV status information \( t_i^N, t_i^P, Y_i \). Interval censoring of HIV serostatus observations is accounted by integrating over all possible seroconversion times \( s \in [t_i^N, t_i^P] \). The total likelihood is the product of the likelihood for each individual. Further mathematical specification of the likelihood will be provided.

For computing the likelihood, time is discretised into 0.2-year time-steps and all event dates and ages are rounded to the nearest two-tenths of the year. Thus integrals of smoothed hazards become discrete sums, and computation of the likelihood is accelerated by aggregating individuals into 0.2-year birth cohorts and calculating the likelihood sequentially along each birth cohorts. Comparison showed that results were similar using 0.2-year time-steps as when using 0.1-year time-steps.

\[ \text{Prior on HIV prevalence early in epidemic period} \]

Informative prior information about zero HIV prevalence before the start of the HIV epidemic was inserting an HIV-person at regular intervals between age 15 and 60 in a specific year. Intervals were 1.0, 0.5, 0.2, and 0.1 years, resulting in sample sizes of \( n=46, 91, 226, \) and \( 451 \) HIV-adults, respectively. The year in which the zero-prevalence prior was applied was 1975 for Masaka, Rakai, and Kisesa (all around Lake Victoria), in 1980 for Karonga and Manicaland (further south, and based on early prevalence surveys in Karonga), and in 1985 for uMkhanyakude in South Africa. The incidence model starts 5 years before this zero-prevalence prior date. In the case of Karonga, the earlier prevalence survey data was not used during analysis of model sensitivity to the weight of the prevalence prior.

\[ \text{Model fitting} \]

The model is implemented in the Stan language and the posterior distribution for model parameters was estimated using Hamiltonian Monte Carlo (HMC) through the No-U-Turn Sampler implemented in the Stan software (Hoffman and Gelman 2014). Six Markov chains with 500 samples each were simulated to estimate parameters for each sex and study site. Convergence was assessed using the split-chain potential scale reduction statistic \( R \) and effective sample size. Further HMC diagnostics were checked including max tree depth and divergent integration steps. Stan and R code for model implementation are available from https://github.com/jeffeaton/alpha-le-stan.
Results (preliminary)

Across all sites, observations consist of 1.6 million person-years, 17,659 deaths, and 510,709 HIV tests among 239,490 adults.

Model fit to demographic and HIV data

Figure 1 illustrates model estimates for adult mortality via the probability of dying between ages 15 and 60 ($45q_{15}$). These are compared to annual non-parametric Kaplan-Meier estimates for $45q_{15}$. The figure illustrates the dramatic effect of ART on decreasing population-wide adult mortality in these populations.

Figure 2 illustrates cumulative HIV incidence by sex, that is the probability of becoming infected between ages 15 and 60 based on synthetic age-specific incidence at a point in time. These estimates are compared with Kaplan-Meier estimates for cumulative incidence using only HIV incidence cohort data (i.e. individuals observed to have at least one negative HIV test). Non-parametric estimates were generated by randomly imputing a seroconversion date during the unknown seroconversion interval and repeating 1000 times to estimate bootstrap 95% CIs.

For both mortality and incidence, model-based estimates jointly synthesizing cohort data produced more precise estimates compared to non-parametric estimates of each outcome independently. Posterior predictive distributions for HIV prevalence will be compared to observed HIV prevalence by age and time.

Sensitivity to prevalence prior

Figure 3 illustrates cumulative HIV incidence over time estimated assuming different amount of weight placed on the zero-prevalence prior near the start of the epidemic. Even in the absence of zero-prevalence prior early in the epidemic, the model estimates that a peak and decline in HIV incidence occurred, often before the start of cohort data (think vertical dashed line).

For around 7-9 years before the start of cohort data, HIV incidence is well identified by the data. This is illustrated by the substantial similarity of the HIV incidence estimate and 95% CI irrespective of the amount of prior information imposed. Before this period, the incidence trend is sensitive to the weight placed on the zero prevalence prior. In the absence of any prior information, mean incidence estimates tend towards a constant level in backward projection, consistent with the first-order difference penalties on spline coefficients.
Discussion and further work

We have been able to fit a flexible model for HIV incidence by age and time to longitudinal general population cohort data, fully utilizing available data about HIV prevalence, incidence, and mortality. Estimates for key epidemiologic and demographic quantities of interest were more precise compared to estimating each outcome independently using the same data.

Through data about age-specific prevalence and mortality, we were able to infer peaks in HIV incidence that occurred prior to the establishment of direct observation of HIV incidence. HIV incidence trends were well-resolved by observed prevalence and mortality data up to around 7-9 years before the establishment of each cohort. Prior to this incidence trends required informative prior information about near-zero prevalence at time points in the 1970s, which are realistic and available from other characterizations of how HIV epidemics spread more generally across sub-Saharan Africa.

Compared to other approaches to estimating HIV incidence from prevalence data, using flexible smooth functions more effectively shares information than treating consecutive age groups or time-periods independently (Hallett et al. 2008), but provides greater flexibility compared to approaches that make strong parametric assumptions about the distribution of incidence by age or the epidemic shape over time (Mossong et al. 2013).

Further work on this paper includes:

- Testing different specifications for the model for HIV incidence rate and comparing results obtained using second-order difference penalties to the first-order penalties used here. Specifically, we hypothesise that second-order difference penalties may provide better estimates and projections for HIV incidence and age-specific mortality. This is because second-order penalties place prior weight on backwards extrapolation of the log-linear incidence trend, instead of maintaining the epidemic level as the first-order penalties. A log-linear incidence trend is consistent with epidemic theory of exponential epidemic growth. For mortality at old ages where data are more sparse, extrapolating a linear trend is likely more appropriate than a constant mortality level.

- Development of methods for projecting HIV incidence to completion of current cohorts with statistical uncertainty.

- Analysis of patterns in age-specific HIV incidence and incidence by birth cohort

A further limitation is that the current model does not account for uncertainty in the distribution of survival after seroconversion as a function of age at infection. This could potentially be relaxed if it proves feasible to simultaneously estimate parameters across
sites using a hierarchical structure (e.g. assuming this survival distribution is the same, but incidence parameters are different across sites). Penalty variance parameters could also potentially be learned jointly in a hierarchical specification, generating generalizable information about the appropriate level of smoothness for estimating and projecting HIV and mortality trends in southern and eastern Africa.
References


Todd, Jim et al. 2007. “Time from HIV Seroconversion to Death: A Collaborative Analysis of Eight Studies in Six Low and Middle-Income Countries before Highly Active


Figure 1: Estimates for trends in the probability of dying between age 15 and 60 ($45_{15}$). Black lines and shaded areas represent trends for the entire population (posterior means and 95% CIs). Green lines indicate estimates for non-HIV $45_{15}$. Red lines illustrate a counterfactual estimate for $45_{15}$ in the absence of ART. Points and vertical line segments illustrate annual Kaplan-Meier estimates for $45_{15}$ using solely mortality data.
Figure 2: Cumulative HIV incidence (the probability of becoming infected) between ages 15 and 60 for men (blue) and women (pink) in each site. Lines shaded areas represent posterior means and 95% credible intervals. Points and vertical segments represent annual non-parametric Kaplan-Meier estimates for cumulative incidence based on directly observed incidence cohort data.
Figure 3: Sensitivity of HIV epidemic estimates to strength of prior information in zero-prevalence prior. Figure illustrates cumulative incidence between ages 15 and 60 over time (posterior mean and 95% CI). Colors represent different sample sizes for zero-prevalence prior, ranging from sample sizes of zero (no prior) to 451. The thick vertical dashed line indicates the year in which population cohort data began. The thin grey dashed line indicates the year in which the zero prevalence prior is imposed.