APPENDIX

Modified Back-projection Approach for Estimating HIV incidence

Methodology used in this study described elsewhere [11]. Briefly, back-projection methodology was modified to estimate HIV incidence based on a parametric formulation of the duration of time between the time of acquisition of HIV infection and the time of earliest diagnosis of HIV infection from HIV surveillance systems. Models were constructed based on two testing forces: “testing during asymptomatic infection” and “testing driven by clinical symptoms at late stage of HIV progression”. This distribution is estimated through the combination of two separate progression rate sub-models. AIDS diagnostic data were used to estimate HIV infections using the back-projection technique until 1987 and were not used to simulate HIV/AIDS trends for subsequent years.

Sub-model 1: HIV testing during asymptomatic infection

It is assumed that a proportion of people infected with HIV will be diagnosed with HIV prior to clinical symptoms or AIDS. It was initially assumed that during the asymptomatic infection, each individual will have a constant testing rate \( \lambda \), since testing rate may be heterogeneous and may vary across individuals, \( \lambda \) follows a standard exponential distribution with probability density function (p.d.f.), \( g(\lambda) = e^{-\lambda} \):

\[
\int_0^\infty f_a(x \mid \lambda)g(\lambda)d\lambda = 1/(x+1)^2
\]

This mixed exponential model is the p.d.f. of the Pareto distribution for the duration \( X \) between HIV infection and HIV diagnosis, which steps down over time giving the following survivor and hazard functions:
\[ S_a(x) = \Pr\{X > x\} = \frac{1}{1+x} \text{ and } h_a(x) = \frac{1}{1+x} \] (1)

The probability of testing \( x \) years after infection then defined as follows:

\[ f_a(x) = S_a(x) - S_a(x + 1) = 1/(x + 2)(x + 1) , \ x = 0,1,2,... \]

**Sub-model 2: HIV testing driven by clinical symptoms at late stage of HIV progression**

A proportion of HIV diagnoses are assumed to be made at a late stage of HIV infection, as a result of clinical symptoms close to, or at, AIDS diagnosis. It was assumed that the progression from HIV infection to the earliest HIV diagnosis follows a distribution similar to the progression to CD4 counts of <200 cells/μL without any treatment. A Weibull distribution was used, with median time to HIV diagnosis of 6.5 years and shape parameter 2.08 with the following survivor and hazard functions:

\[ S_b(x) = \exp\{-0.014x^{2.08}\} \text{ and } h_b(x) = 0.029x^{1.08} \] (2)

The probability of testing \( x \) years after infection is defined by:

\[ f_b(x) = S_b(x) - S_b(x + 1) \]

\[ = \exp\{-0.014x^{2.08}\} - \exp\{-0.014(x + 1)^{2.08}\} , \ x = 0,1,2,... \]

**Overall rate of progression to HIV diagnosis**

The overall rate of progression to HIV diagnosis \( f(x | t; \varphi) \) was then formulated based on combining the two sub-models (i.e. \( f_a(x) \) and \( f_b(x) \)) described above by using a mixture distribution model as follows:

\[ f(x | t; \varphi) = S(x | t; \varphi) - S(x + 1 | t; \varphi) \]

where \( S(x | t; \varphi) \) is the survival function:
\[ S(x \mid t; \varphi) = m_t(\varphi)S_a(x) + [1 - m_t(\varphi)]S_b(x) \]  

(3)

where \( m_t(\varphi) = \frac{\pi e^{\delta + \gamma(t-t_0)}}{1 + e^{\delta + \gamma(t-t_0)}}, t \geq t_0, \) is a mixing function with \( \varphi = (\pi, \delta, \gamma); \) \( \pi \) represents the proportion of infected individuals who were not tested because of clinical symptoms; \( \delta \) determines the overall historical trend of the epidemic and \( \gamma \) denotes the rate of increase in infection at time \( t. \) When HIV testing became available at \( t = t_0, \) \( m_{t_0}(\varphi) = \frac{\pi e^{\delta}}{1 + e^{\delta}}, \) as \( m_t(\varphi) \) increases with time \( t \) at rate \( \lambda \) to a saturation level \( \pi = \lim_{t \to \infty} m_t(\varphi). \) We assume that there will be a proportion \( 1 - \pi \) of infected individuals that are driven by clinical symptoms to be tested (as specified by sub-model 2).

Prior to the availability of HIV testing in 1985, HIV diagnosis was only made on the basis of AIDS symptoms. This information was incorporated into our model by allowing the model vary overtime, so that the proportion of diagnoses due to clinical symptoms would decrease after 1985. Therefore, the mixture distribution, \( f(x \mid t; \varphi) \) results in an overall “bath-tub” shaped hazard, with a relatively high rate of HIV diagnosis in the first year following HIV infection, which then decreases over time, before increasing again as clinical symptoms would appear. The two sub-models given by (1) and (2) are then mathematically connected based on HIV diagnostic data. For this purpose, we first define the following distribution functions by using (3):

\[ F(t \mid t; \varphi) = 1 - S(t \mid t; \varphi) = \Pr\{\text{HIV diagnosed during } (t, t + \tau) \mid \text{HIV infection at time } t; \varphi\} \]

and

\[ r(\tau \mid t) = \Pr\{\text{infection in } [t - \tau, t) \mid \text{diagnosis at time } t\}. \]
Where \( F(\tau \mid t; \varphi) \) is the distribution of “time to testing” and \( r(\tau \mid t) \) is the distribution of “time-since-infection”. For small \( \tau \), one has the approximation 
\[ r(\tau \mid t) i_2(t) \approx i_1(t) F(\tau \mid t), \]
implying the number of individuals who are infected and diagnosed during a small duration of time \( \tau \) around the infection time \( t \). Therefore, the back-calculation predicts
\[ \tilde{r}(\tau \mid t; \varphi) = \frac{\tilde{i}_1(t, \varphi)}{\tilde{i}_2(t, \varphi)} F(\tau \mid t; \varphi). \]

The data on “recent infections” \((k_i)\) among newly diagnosed individuals \((n_i)\) were used to identify the parameters in \( \varphi \). Since the pair \((k_i, n_i)\) follow a binomial distribution, the likelihood function for \( \varphi \) can be written as
\[
L(\varphi \mid \tilde{i}_1(t; \varphi)) = \tilde{r}(\tau \mid t; \varphi)^{k_i} [1 - \tilde{r}(\tau \mid t; \varphi)]^{n_i - k_i}.
\]

The Expectation-Maximization-Smoothing (EMS) algorithm is used to back-calculate HIV incidence from HIV diagnostic data determine the final estimate for the HIV incidence. For observed values of \((k_i, n_i)\), methodology searches all possible values in the parameter space for \( \varphi = (\pi, \delta, \gamma) \) to generate \( \tilde{r}(\tau \mid t; \varphi) \) that is in best agreement with observed proportion \( \frac{k_i}{n_i} \). This also implies that, for each combination of \((\pi, \delta, \gamma)\) in the parameter space, the following back-calculations take place in order to determine the final incidence, \( \hat{i}(t) = \tilde{i}_1(t, \hat{\varphi}) \), (i) based on AIDS diagnostic data; (ii) based on HIV diagnostic data.