Use of EPP-ASM for CSAVR

Jeff Eaton, Guy Mahiane, Rob Glaubius, John Stover

Mon Jun 25 12:42:31 2018

Introduction

Following discussion at the October 2018 UNAIDS Reference Group meeting, we showed that the code for the EPP-ASM model was able to take annual HIV incidence rates as inputs and reproduce Spectrum model ouputs for the numbers of new HIV infections, PLHIV, and AIDS deaths for the age 15+ population. This report extends that analysis to explore combining the EPP-ASM and CSAVR models for epidemic inference from case surveillance and vital registration data.

There a number of motivations for exploring a combined model:

- Consider opportunities for more efficient estimation for CSAVR countries.
- Explore influence of modelling incidence rate or transmission rate for estimates of recent incidence trends and potential harmonization of assumptions for CSAVR and EPP.
- Work towards incorporating case surveillance data into HIV estimates for SSA.

Example data: Netherlands and Chile 2017 Spectrum files

As an example case study, we use data about new diagnoses and numbers of AIDS deaths from the Netherlands and Chile 2017 Spectrum files. For the Netherlands, the Spectrum file only includes a partial time series of AIDS deaths, so we updated these with the full time series of AIDS deaths to adults aged 15+ from the WHO Mortality database. Moreover, Netherlands uses the ECDC model for direct incidence estimates, not CSAVR, and so it is uncertain the source or accuracy of new case inputs in the Spectrum file. In short, consider any outputs a proof of concept only.

The tables below summarize the data for the Netherlands and Chile datasets.

Table 1: Netherlands case surveillance and vital registration data

	plhiv	undercount	new_cases	undercount	aids_deaths	undercount
1983					8	
1984					16	
1985					30	
1986					63	
1987					105	
1988					132	
1989					201	
1990					267	
1991					291	
1992					409	
1993					424	
1994					444	
1995					435	
1996					320	
1997			813		181	
1998			647		134	
1999			687		137	

	plhiv	undercount	new_cases	undercount	aids_deaths	undercount
2000			843		132	
2001			960		127	
2002			1027		89	
2003			1065		87	
2004			1174		85	
2005			1219		80	
2006			1136		48	
2007			1233		66	
2008			1315		53	
2009			1214		70	
2010			1218		50	
2011			1158		54	
2012			1073		44	
2013	22400		1035		34	
2014	22600		897		38	
2015	22900		866		33	

Table 2: Chile case surveillance and vital registration data

	plhiv	undercount	new_cases	undercount	aids_deaths	undercount
1987			93	39		
1988			234	39		
1989			388	31		
1990			597	31		
1991			651	24		
1992			897	24		
1993			1025	18		
1994			1083	18		
1995			1282	18		
1996			1669	18		
1997			1651	18	410	10
1998			1723	13	383	10
1999			1959	13	474	10
2000			2116	13	458	10
2001			2123	13	552	10
2002			1993	13	440	10
2003			2005	13	423	3
2004			2029	8	399	5
2005			2084	8	395	7
2006			2200	8	422	6
2007			2477	8	398	8
2008			2717	8	392	8
2009			2787	8	435	8
2010			2982	5	435	7
2011			3159	5	472	7
2012			3395	5	456	9
2013			4014	5	523	8
2014			4080	2	506	10
2015			4301	2		
2016			4964	2		

Model for case surveillance data

Model for new diagnoses

Adding new diagnoses to EPP-ASM

We extended the EPP-ASM model to explicitly stratify the untreated HIV positive population (`hivpop`) population as those who are diagnosed and not diagnosed, in order to model data about numbers of new HIV diagnoses and output estimates for the proportion diagnosed in the HIV care cascade.

It is assumed that diagnoses does not affect disease progression, mortality, or HIV transmission. To implement diagnosis in such a way to minimally affect other model processes, a separate array `diagnpop` of the same dimension as `hivpop` in order to track the number diagnosed within each stratum of the untreated HIV positive population. Individuals are not removed from `hivpop` once they become diagnosed, but all calculations of population progression, disease progression, and mortality that affect `hivpop` are replicated for the `diagnpop`.

Upon ART initiation, individuals are removed from `diagnpop` (similar to removal from `hivpop`. Presently, there are no changes to the allocation of new ART initiations by age, sex, or CD4 category among the untreated HIV population. New ART initiations are presumed to come from the 'diagnosed' population where available. If there are more persons initiating ART in a given age/sex/CD4 stratum than have been previously diagnosed, then individuals are taken from the undiagnosed population and added to the number of new diagnoses in that time step (resulting in a higher number of diagnoses than specified by the diagnosis rate parameter).

Persons who dropout from ART are returned into the `diagnpop`.

There are two additional model input parameters that must be specified as part of the `fp` input list:

- `diagn_rate`: A four-dimensional array (same dimensions as `hivpop`) specifying the diagnosis rate per year for undiagnosed HIV positive persons stratified by CD4 category, HIV age groups, sex, and projection year.
- `t_diagn_start`: An integer indicating the first projection year in which HIV diagnoses may occur. This does not affect the model output, but improves computational efficiency by preventing looping over diagnosed population calculations for years prior to the first diagnosis. `diagn_rate` should be 0 for all years before `t_diagn_start`. Setting `t_diagn_start` to be equal to or greater than the number of projection years should result in no calculations related to the diagnosed population and result in a projection with identical results and no additional computational burden as the EPP-ASM model without diagnosis.

There are three additional outputs, stored as attributes to the return value of `simmod()`:

- `diagnpop`: The number diagnosed within each CD4 stage / HIV age / sex stratum of the untreated HIV positive poplation. `diagnpop` is a subset of `hivpop`.
- `diagnoses`: The number of new HIV diagnoses occurring by stratum in each projection year (mid-year to mid-year).
- `artinits`: The number of new HIV ART initiations by stratum in each projection year (mid-year to mid-year).

Three additional functions are added reporting for calculating outputs related to diagnoses and ART intiations. These and other functions to be added may be used for likelihood calculations.

- `calc_diagnoses()` returns the total number of new diagnoses in each year.
- `diagnosis_median_cd4()` returns the median CD4 count among newly diagnosed persons in each year.
- `artinit_median_cd4()` returns the median CD4 count among new ART initiations persons in each year.

Parametric model for diagnosis rate

Modelling new HIV diagnoses requires specifying a parametric model for the `diagn_rate` input, the HIV diagnosis rate by CD4 stage m, age group a, and sex s, at time t, a function $\Delta_{s,a,m}(t)$. The relative diagnosis rate across groups at time t is assumed to be proportional to the HIV mortality rate $\mu_{s,a,m}$ and the trend in diagnosis rate of the course of the epidemic is modelled by a cumulative gamma distribution function with shape parameter 1 and rate parameter θ. That is

$$\Delta_{s,a,m}(t) = \mu_{s,a,m} \cdot \gamma_{max} \cdot \int_{t_0}^{t} e^{-\theta\tau} d\tau \tag{1}$$

where t_0 is the start time of the epidemic (e.g. $t_0 = 1970$).

This is similar to the approximation by Mahiane, except that we explicilty track size of the undiagnosed population over the course of the epidemic rather than using the approximation calcuated based on time since infection.

For Bayesian inference, we define diffuse prior distributions on the parameters for the diagnosis rate over time:

$$\log(\gamma_{max}) \sim \text{normal}(3, 5)$$
$$\log(\theta) \sim \text{normal}(-3, 5)$$

Models for incidence and transmission rate

We considered three parameteric models for the HIV incidence rate, denoted $\lambda(t)$, or HIV transmission $r(t)$. For directly modelling $\lambda(t)$, we implemented the single logistic (two parameters) and double logistic (five parameters) models implemented by the current CSAVR software. The single logistic model is

$$\lambda(t) = c \cdot \frac{e^{\alpha(t-t_0)}}{1 + e^{\alpha(t-t_0)}} \tag{2}$$

where c is the equilibrium incidence rate and *alpha* determines the rate of increase in incidence. The double logistic model is

$$\lambda(t) = \frac{e^{\alpha(t-t_{mid})}}{1 + e^{\alpha(t-t_{mid})}} \cdot \left(2 \cdot a \cdot \frac{e^{\beta(t-t_{mid})}}{1 + e^{\beta(t-t_{mid})}} + b\right) \tag{3}$$

b is the equilibrium incidence reate, $(a + b)/2$ is the incidence rate at the inflection point t_{mid}, α determines the rate of increase and β determines the rate of convergence to the asymptote.

Secondly, instead of directly modelling the HIV incidence rate $\lambda(t)$, we considered modelling the transmission rate $r(t)$, as in the EPP model. In this case, the incidence rate is

$$\lambda(t) = r(t) \cdot \frac{I(t)}{N(t)} \cdot \left(1 - 0.7 * \frac{A(t)}{I(t)}\right). \tag{4}$$

The expression $I(t)/N(t)$ is the HIV prevalence at time t, $A(t)/I(t)$ is the ART coverage and 0.7 is the average reduction in transmission per additional person on ART. We use a logistic function to model the logarithm or $r(t)$, termed the *rlogistic* model, with four parameters

$$\log r(t) = r_0 - (r_\infty - r_0) \cdot \frac{1}{1 + e^{-\alpha \cdot (t-t_{mid})}} \tag{5}$$

where e^{r_0} is the initial exponential growth rate of the epidemic, e^{r_∞} is the equilibrium value for $r(t)$, α is the rate of change in $\log r(t)$ and t_{mid} is the inflection point. For this model we additionally specify a fifth parmaeter ι as the incidence rate at time $t = t_0$ providing the initial pulse of infections.

For Bayesian inference, we defined diffuse prior distributions on all parameters. For the single and double logistic models, we defined the following prior distributions

$$\log \alpha, \log \beta \sim \text{normal}(-1, 5)$$

$$\log a, \log b \sim \text{normal}(-10, 5)$$

$$t_{mid} \sim \text{normal}(1995, 10)$$

For the rlogistic model, we used prior distributions

$$r_0 \sim \text{normal}(\log(0.35), 0.5)$$

$$r_\infty \sim \text{normal}(\log(0.09), 0.3)$$

$$\log \alpha \sim \text{normal}(\log(0.2), 0.5)$$

$$t_{mid} \sim \text{normal}(1993, 5)$$

$$\log \iota \sim \text{normal}(-13, 5)$$

Semi-parametric model variants for incidence rate or transmission rate (segmented polynomials, p-splines, random walk, etc.) remain to be implemented and tested.

Likelihood

The total number of expected diagnoses in year t is the sum of expected diagnoses over all sex, age, and CD4 groups of undiagnosed persons

$$n_I(t) = \sum_s \sum_a \sum_m \Delta_{s,a,m}(t) \cdot U_{s,a,m}(t)$$

where $\Delta_{s,a,m}(t)$ is the diagnosis rate describe above and $U_{s,a,m}(t)$ are the number HIV positive undiagnosed, recalling that for simplicity we assumed this was equal to the untreated population. A Poisson distribution is used as the likelihood for the reported number of diagnoses $y_I(t)$ given the expected number of diagnoses $n_I(t)$ and the proportion estimated underreporting of new diagnoses $u_I(t)$

$$y_I(t) \sim \text{Poisson}(n_I(t) \cdot (1 - u_I(t))) \tag{6}$$

A Poisson likelihood is also used for the reported number of AIDS deaths $y_D(t)$ given the expected number of AIDS deaths $n_D(t)$ predicted by the Spectrum model and the estimated underreporting of AIDS deaths $u_D(t)$:

$$y_D(t) \sim \text{Poisson}(n_D(t) \cdot (1 - u_D(t))) \tag{7}$$

We have not yet implemented the likelihood for the mean CD4 count at diagnosis developed by Mahiane.

Model estimation

To fit the model, we first create a model fitting object consisting of the Spectrum model inputs and the CSAVR data inputs.

```
## Create fitting objects
nl <- list(fp = nl_fp, csavrd = nl_csavrd)
cl <- list(fp = cl_fp, csavrd = cl_csavrd)
```

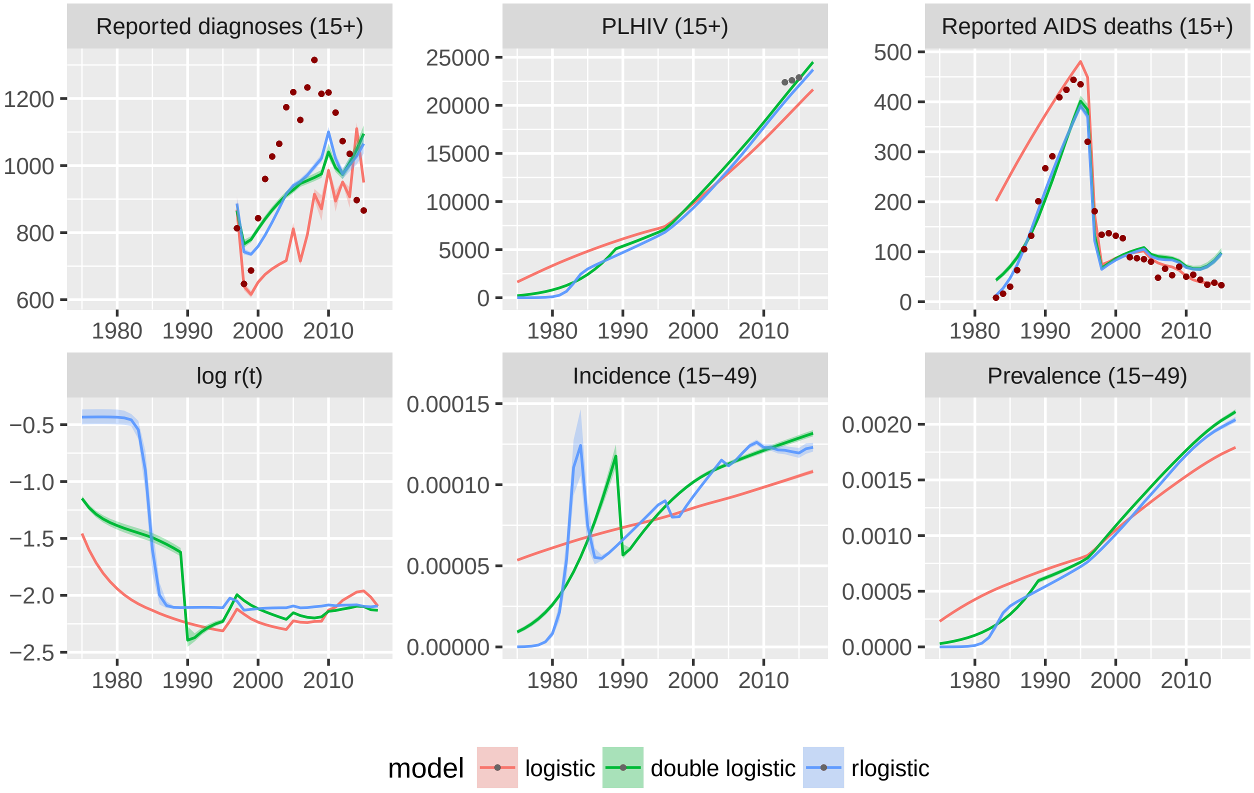
The code below illustrates fitting each of the three models to the Netherlands dataset using either optimization (`optfit = TRUE`) or full Bayesian inference with IMIS. The function `fitmod_csavr(...)` will prepare the model fit and data, and then call the requested model fitting routine.

```
## fit single logistic model for incidence rate
nl_opt1 <- fitmod_csavr(nl, incid_func = "ilogistic", B0=1e4, optfit=TRUE)
nl_fit1 <- fitmod_csavr(nl, incid_func = "ilogistic", B0=1e4, B=1e3, B.re=3e3, opt_iter=1:3*5)

## fit double logistic model for incidence rate
nl_opt2 <- fitmod_csavr(nl, incid_func = "idbllogistic", B0=1e2, optfit=TRUE)
nl_fit2 <- fitmod_csavr(nl, incid_func = "idbllogistic", B0=1e4, B=1e3, B.re=3e3, opt_iter=1:3*5)

## fit logistic model for transimssion rate (r(t))
nl_opt3 <- fitmod_csavr(nl, eppmod="rlogistic", B0=1e4, optfit=TRUE)
nl_fit3 <- fitmod_csavr(nl, eppmod="rlogistic", B0=1e4, B=1e3, B.re=3e3, opt_iter=1:3*5)
```

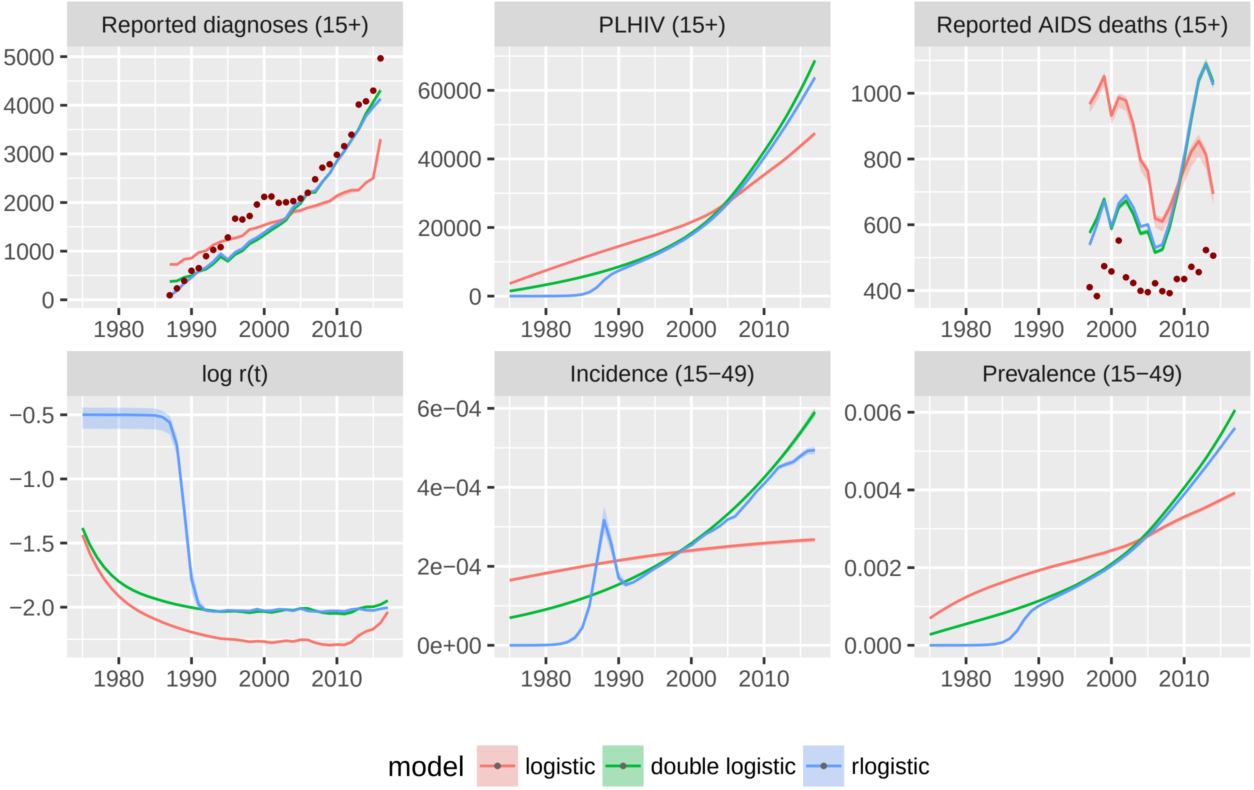
Fit model to Chile dataset.

```
## fit single logistic model for incidence rate
cl_opt1 <- fitmod_csavr(cl, incid_func = "ilogistic", B0=1e4, optfit=TRUE)
cl_fit1 <- fitmod_csavr(cl, incid_func = "ilogistic", B0=1e4, B=1e3, B.re=3e3, opt_iter=1:3*5)

## fit double logistic model for incidence rate
cl_opt2 <- fitmod_csavr(cl, incid_func = "idbllogistic", B0=1e3, optfit=TRUE)
cl_fit2 <- fitmod_csavr(cl, incid_func = "idbllogistic", B0=1e4, B=1e3, B.re=3e3, opt_iter=1:3*5)

## fit logistic model for transimssion rate (r(t))
cl_opt3 <- fitmod_csavr(cl, eppmod="rlogistic", B0=1e4, optfit=TRUE)
cl_fit3 <- fitmod_csavr(cl, eppmod="rlogistic", B0=1e4, B=1e3, B.re=3e3, opt_iter=1:3*5)
```

The figure below shows model results for the Netherlands from IMIS.



The figure below shows model results for Chile.



We should be cautious about over interpreting these results because the parametric models implemented thus far are relatively inflexible to matching data-driven trends. A few observations:

- Both the double logistic model and rlogistic model, which each have five parameters, are better able to fit the data compared to the two parameter single logistic model.
- The uncertainty ranges are very narrow. This is likely due to both using relatively few parameters and not considering any uncertainty in other model processes such as disease progression or changes in testing and new diagnoses.
- Both the double logistic model and rlogistic model give fairly similar fits to observed new diagnoses and AIDS deaths, but result in qualitatively different estimates for recent HIV incidence trends. This conclusion should be reconsidered after implementing more flexible models for the incidence rate and transmission rate.

Benchmarking

The table summarizes the time in seconds for model fitting via optimization and IMIS.

Country	Model	Optimization	IMIS
Netherlands	logistic	0.29	85.3
Netherlands	double logistic	1.08	119.8
Netherlands	r(t) logistic	0.24	299.8
Chile	logistic	0.36	84.1
Chile	double logistic	1.02	118.2
Chile	r(t) logistic	2.02	366.1