A small-area model for estimating subnational HIV incidence

Jeffrey W Eaton, Sumali Bajaj, and others

Sun Apr 29 10:23:18 2018

1 Introduction

We have previously described a model for estimating subnational HIV prevalence and ART coverage using survey data and routine health system data [@Eaton2017]. This note describes an extension of our model to estimate subnational HIV incidence including probabilistic uncertainty. We illustrate the proposed model for estimating the number of new HIV infections by subnational area for two-common cases of data availability:

1. Subnational data about HIV prevalence and ART coverage are available, but estimates of HIV incidence are only available at the national level, for example from national EPP/Spectrum estimates.
2. Direct data about subnational HIV incidence are available through inclusion of tests for recent infection in population survey data.

Finally, we discuss the model as a framework for identifying HIV transmission 'hotspots' for prioritising further HIV prevention activities.

2 Modelling HIV incidence

The key principle underpinning the modelling of infectious disease dynamics is that the risk of acquiring infection depends on (1) the rate of tranmsission from an infectious host and (2) the probability of contacting an infectious host [@Anderson1982]. For HIV, it is useful to express the HIV incidence rate, or *force of infection*, λ as the product of the transmission rate κ and the prevalence of unsuppressed HIV viral load in the population, that is

$$\lambda = \kappa \cdot \rho \cdot (1 - \omega \cdot \alpha),$$

where ρ is the HIV prevalence, α is the ART coverage, and ω is the relative reduction in tranmission for persons on ART. This relationship between HIV incidence and prevalence has long been used to estimate and project national HIV incidence trends with the EPP model [@EPP].

This theory suggests a log-linear model for the HIV incidence rate λ_i in region i

$$\log(\lambda_i) = \log(\kappa_0) + \log(\rho_i) + \log(1 - \omega \cdot \alpha_i) + u_i, \tag{1}$$

where κ_0 is the average transmission rate, ρ_i is the HIV prevalence in region i, α_i is the ART coverage, and u_i is a random effect capturing other region-level differences in the relative transmission rate. Removing the prevalence term from the linear model to directly model spatial patterns in the transmission rate κ_i may be more intuitive:

$$\log(\kappa_i) = \log(\kappa_0) + \log(1 - \omega \cdot \alpha_i) + u_i, \tag{2}$$

In this formulation, the incidence rate $\lambda_i = \kappa_i \cdot \rho_i$, noting that κ_i is the ratio of incidence over prevalence. The expression for the annual number of new infections I_i in region i is

$$I_i = \kappa_i \cdot \rho_i \cdot ((1 - \rho_i) \cdot N_i),$$

where N_i is the total population and observing that $(1 - \rho_i) \cdot N_i$ is the number susceptible in the population.

2.1 Case 1: No directly observed HIV incidence

In many cases, data about HIV prevalence and ART coverage are available at subnational levels from population surveys and routine programme data, but estimates for HIV incidence are only available at national or first administrative unit level, for example from the EPP model. In this case, the objective for a model for subnational HIV incidence is to estimate the most plausible subnational distribution of new HIV cases based on the available information about subnational prevalence and ART coverage ρ_i and α_i, and such that the aggregated subnational estimates of new infections are consistent with the exogenously estimated national (or larger area) esimates. The random effect terms u_i allow for additional uncertainty about subnational new infections arising from transmission being higher or lower than predicted solely by the prevalence of unsuppressed viral load.

Exogenous estimates and uncertainty about the national HIV incidence ratey λ^{Nat} may be summarized as a normal distribution for $\log(\lambda^{Nat})$:

$$\log(\lambda^{Nat}) \sim \text{Normal}(\log(\hat{\lambda}^{Nat}), \sigma_{\lambda^{Nat}}) \tag{3}$$

We now seek to relate the relate the national HIV incidence rate to the aggregation of the subnational incidnece rates. Defining $S_i = (1 - \rho_i) \cdot N_i$ as the number not infected with (susceptible to) HIV infection in region i, the national HIV incidence rate is expressed as

$$\lambda^{Nat} = \frac{\sum_i \lambda_i \cdot S_i}{\sum_i S_i} = \frac{\sum_i \kappa_0 \cdot (1 - \omega \cdot \alpha_i) \cdot \rho_i \cdot \exp(u_i) \cdot S_i}{\sum_i S_i}. \tag{4}$$

Taking the logarithm and rearranging terms yields that

$$\log(\lambda^{Nat}) = \log(\kappa_0) + \log\left(\sum_i (1 - \omega \cdot \alpha_i) \cdot \rho_i \cdot \exp(u_i) \cdot S_i\right) - \log\left(\sum_i S_i\right) \tag{5}$$

Substituting equation (5) into equation (3) defines a model for subnational incidence given estimates of prevalence, ART coverage, and the variance of the spatial random effects:

$$\begin{aligned} u_i &\sim \text{Normal}(0, \sigma_u) \\ \log(\kappa_0)|\rho_i, \alpha_i, u_i &\sim \text{Normal}\left(\log(\hat{\lambda}^{Nat}) - \log\left(\sum_i k_i \cdot S_i\right) + \log\left(\sum_i S_i\right), \sigma_{\lambda^{Nat}}\right) \\ \lambda_i &= \kappa_0 \cdot k_i, \end{aligned} \tag{6}$$

where $k_i = \rho_i \cdot (1 - \omega \cdot \alpha_i) \cdot \exp(u_i)$. The joint distribution for ρ_i, α_i is estimated using the model we have described previously for subnational HIV prevalence and ART coverage. In the absence of subnational HIV incidence data, a value for σ_u must be assumed or derived from other sources.

2.2 Case 2: Recent infection testing algorithm

A number of national HIV household surveys are now including biomarker-based laboratory algorithms to identify whether HIV infections were 'recently' acquired. Instead of simply providing estimates of the number HIV positve and number HIV negative in the population, these surveys furnish a trinomial observation $\{R_i, L_i, N_i\}$ of the number HIV positive and recently infected, the number of long-term (not-recent) HIV infected, and the number HIV negative, respectively, in each region i.

Kassanjee and colleagues describe an estimator for HIV incidence as a function of the HIV prevalence ρ, the proportion of HIV-positive cases that were recently infected p^R, the mean duration of recent infection (MDRI)

Ω_T, the expected proportion of long-term infections misclassified as recent β_T, and the recency cut-off period T:

$$\lambda = \frac{(p^R - \beta_T) \cdot \rho}{(1-\rho) \cdot (\Omega_T - \beta_T T)}. \tag{7}$$

Rearranging equation 7, the expected proportion of recent infections p_i^R in a region as a function of the prevalence ρ_i and force of infection λ_i is

$$\begin{aligned} p_i^R &= \frac{\lambda_i \cdot (1-\rho_i) \cdot (\Omega_T - \beta_T T) + \beta_T \rho_i}{\rho_i} \\ &= \kappa_i \cdot (1-\rho_i) \cdot (\Omega_T - \beta_T T) + \beta_T \end{aligned} \tag{8}$$

where the second equality follows from substituting $\lambda_i = \kappa_i \cdot \rho_i$.

Using this expression, we may express a likelihood for the number recently R_i of the total number HIV positive $P_i = R_i + L_i$:

$$R_i \sim \text{Binomial}(p_i^R, P_i) \tag{9}$$

By convention, we assume

$$\begin{aligned} \Omega_T &\sim \text{Normal}(\Omega_{T0}, \sigma_\Omega^2) \\ \beta_T &\sim \text{Normal}_{\beta_T \geq 0}(\beta_{T0}, \sigma_\beta^2), \end{aligned} \tag{10}$$

where the values of Ω_{T0}, σ_Ω^2, β_{T0}, σ_β^2, and T are specified for a given survey depending on the HIV subtypes in the survey population and the details of the RITA used.

In realistic applications of the model, data R_i and P_i most likely arise from complex survey data including clustered sampling with unequal weights, which must be accounted for in analysis. We use normalized weighted counts for R_i and P_i for evaluation of the likelihood in equation~ (9). A conventional approach for model-based analysis of survey data is to approximate the likelihood for the observed counts with design-based estimates and standard errors at the level of analytical interest, in this case a direct estimate of $\hat{p}_i^R$. This approach is infeasible here because recent infection is a rare event, and indeed we expect a large number of regions with $R_i = 0$. The use of weighted counts ensures that aggregated model-based estimates are consistent with design-based estimates, though may understate the uncertainty arising from the clustered sampling [1]. The bias will be minimal if the regions i are small enough that sampled clusters within a given region are relatively homogeneous. Optimal approaches for analysis of complex survey data with small counts remains an open research question. Modelling the full survey design, for example by post-stratifying the analysis by all factors used in defining analysis weights and including cluster random effects nested with area-level effects, has been used elsewhere, but is overly cumbersome for our exposition and may be impossible if full survey design details are not avaialble to the analyst.

3 Application to simulated data from Malawi

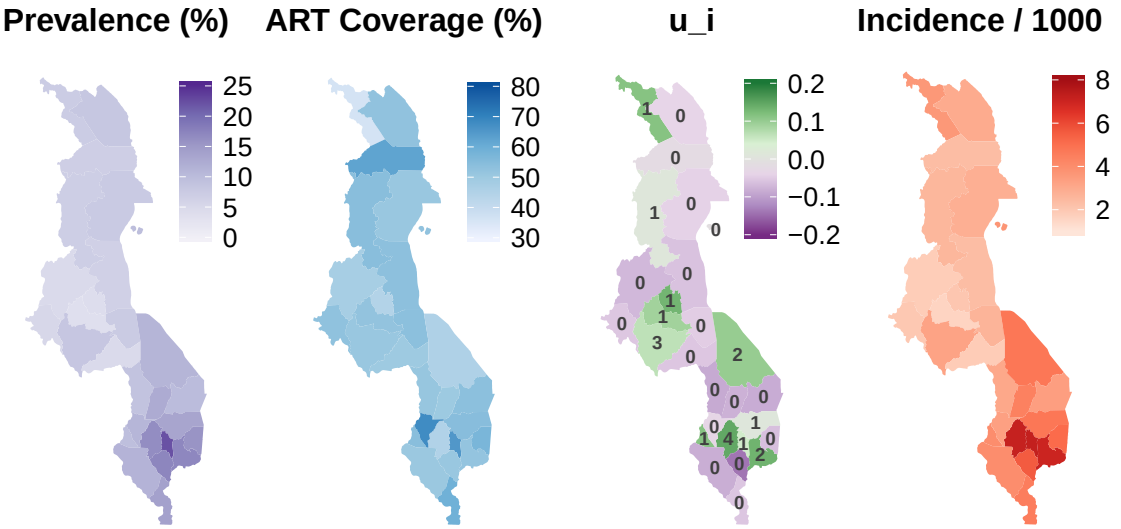
We demonstrate the models for subnational incidence estimates using *simulated data* representative of Malawi. National-level estimates for HIV prevalence, ART coverage, and incidence are available from the Malawi Population HIV Impact Assessment (MPHIA) survey conducted in 2016. We generate simulated subnational data consistent with national-level survey results for illustration. Further details of the simulated data are described in the Appendix and the simulated district-level survey results are presented in Table~**??**.

For 'case 1' in which only national incidence estimates are available, we used estimates for national HIV incidence among adults 15-49 years of 0.36% (SE 0.087%). Log-transforming these estimates yields estimates $\log(\hat{\lambda}^{Nat}) = -5.63$ and $\sigma_{\lambda^{nat}} = 0.241$. These

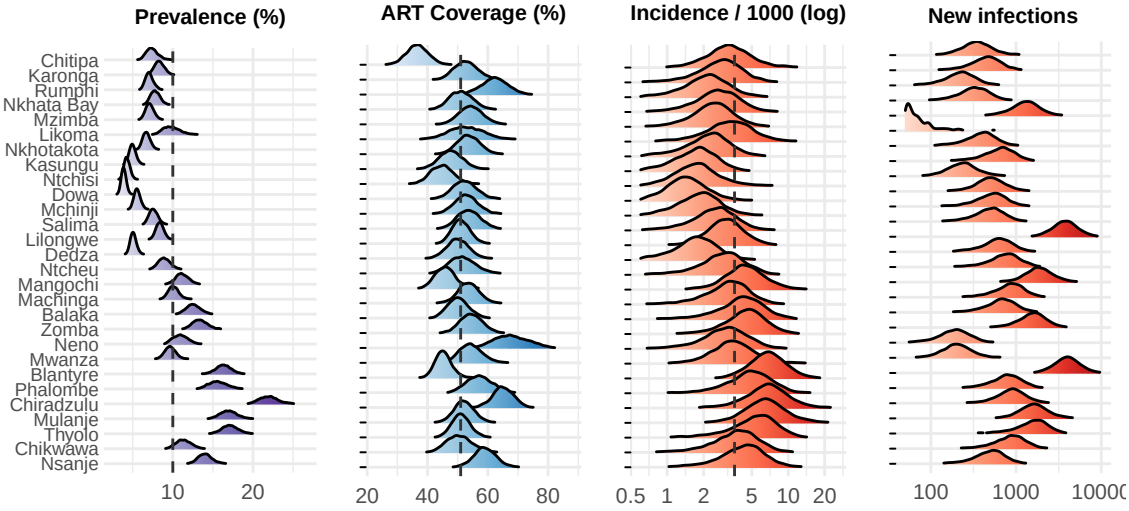
[1] This approach is consistent with the primary analysis of PHIA survey results which assumed a design effect of 1.0 for the proportion recent in published incidence estimates.

Throughout all analyses, we used the value $\omega = 0.7$ as assumed by the EPP model for the average reduction in transmission per percentage increase in ART coverage

The



The next figure shows the posterior distribution for HIV prevalence, ART coverage, HIV incidence rate, and the number of new infections by district. Vertical dashed lines in the first three panels demarcate the national estimate for each outcome. The figure illustrates the large heterogeneity in HIV prevalence across districts and district-level prevalence estimates are relatively precise, reflecting the large amount of available data about HIV prevalence. Estimates for the HIV incidence rate are more uncertain given that district estimates are based on a total of 18 cases of recent infection, but overall the model estimates higher incidence rate in high prevalence districts in Southern Malawi. The predicted number of new infections reflects the variation in incidence rate and the district population. The largely urban districts of Blantyre and Lilongwe account for an estimated 30% (21%–42%) of all new infections. Blantyre is one of the highest prevalence and incidence districts, but prevalence and incidence in Lilongwe are estimated to be slightly lower than the national levels.



TODO: Compare posterior estimates for national prior vs. RITA data.

TODO: Compare incidence estimates depending on assumption about standard deviation of u_i.

4 Identifying HIV transmission hotspots

5 Dicussion and future directions

We have proposed a small-area model for subnational HIV incidence that combines basic theory of infectious disease transmission and a stochastic model for the force of infection. We use the prevalence of unsuppressed HIV viral load to predict the relative levels of regional HIV incidence. Including this theoritical epidemiologic relationship in the model is important because in many cases there are not directly observed data about HIV incidence, and where data are available, the number of recent infections observed in subnational regions is small, in many cases zero. We anticipate that validation of this assumption through standard statistical approaches such as cross-validation may be challenging due to the paucity of subnational HIV incidence data. However, the assumption is supported by recent studies from Kwa-Zulu Natal South Africa and western Kenya that found that HIV incidence was highly correlated with the prevalence of unsuppressed HIV viral load in the loca area [@Vandormael2017, @Ndhiwa abstract].

The model for $\log(\lambda_i)$ may be enhanced in natural ways, including adding fixed effects that predict regions of higher or lower HIV transmission, or through more sophisticated modelling of the u_i, such as capturing spatial correlation. Fixed effects may include factors such as urbanization, distance from major roadways, or locations of risk enumerated through risk-mapping exercises. With sufficient data on subnational HIV incidence, these relationships may be learned through regression, or may be included in the model through informative prior distributions derived from epidemiological literature.

One appealing feature of the the proposed small-area incidence model is the interpretation of the u_i random-effect terms. Positive or negative values of u_i indicate regions where HIV transmission is higher or lower than would be expected based on the local prevalence of unsuppressed HIV viral alone. This presents a framework for statistically identifying HIV transmission 'hotspots', which may be useful for focusing HIV prevention programmes for greater impact. Given the small number of recent infection cases expected in any one subnational area, we do not anticipate that analysis of recent infection testing algorithms in household surveys is likely to provide sufficient statistical power to robustly identify transmission hotspots. However, more precise estimates for transmission hotspots may be possible through extending the model to incorporate other data about or covariates for HIV, such as case reports of new HIV diagnoses from routine programmatic data or estimates of HIV incidence derived from self-reported HIV testing history in household surveys [@Fellows].

5.1 Relationship between the force of infection and prevalence

- Expect higher force of infection in areas of higher prevalence.
- Not necessasrily true.
- Including $\beta \cdot \log(\rho_i)$ regression term in model for λ_i, hypothesise positive coefficient

5.2 Indirect inference about incidence from prevalence trends

Noting that the proposed small-area HIV incidence model for uses the exact same functional form for predicting HIV incidence as a function of the force of inection, prevalence, and ART coverage as the EPP model uses for inferring HIV incidence trends, we begin to see that we are conceptually only a small step from a in integrated model for jointly estimating spatio-temporal HIV prevalence, incidence, and service coverage. The final component requires specifying a model for HIV mortality, such that we can model the relationship between HIV incidence, prevalence, and mortality over time. However, acheiving this is anticipated to entail substantial research effort both in model development and computational aspects.

5.3

A Appendix: simulated subnational survey data for Malawi

The 2016 MPHIA survey included RITA-based incidence estimates. Regional data from this survey are not available for analysis, but it has been reported that among respondents aged 15-49 at the national level, there were 13727.9 HIV-negative, 1527.1 HIV positive, and 17.6 recently infected weighted cases [@MPHIA2016_FirstReport]. We use these national counts to generate simulated district-level counts of recent infections for the purpose of demonstrating the small-area incidence model.

The expected number of 'recent'

```
set.seed(7046752)
data(mwshest)
mw <- mwshest
mw@data[c("lat", "long")] <- coordinates(mw)

omega <- 0.7

## Number HIV positive in each district
mw$hivpos <- with(mw@data, pop15to49 * prev_mod5)

## Distribution of HIV infections by district, determined by population size,
## prevalence of unsuppressed viral load, and number susceptible.
mw$infections <- with(mw@data, hivpos * (1 - prev_mod5) * (1 - 0.7 * artcov_mod5))

## Distribute number sampled in each district proportional to population size
mw$nsamp <- c(rmultinom(1, 15255, mw$pop15to49))

## Sample number HIV+ proportional to distribution of HIV population
mw$npos <- c(rmultinom(1, 1527, mw$hivpos))

## Sample 18 observed recent cases proporitonal to new infections
mw$nrecent <- c(rmultinom(1, 18, mw$infections))

## Survey prevalence estimates and standard error, assume design effect = 2.0
mw$prev_est <- with(mw@data, npos/nsamp)
mw$prev_se <- with(mw@data, sqrt(2 * prev_est * (1 - prev_est)/nsamp))
```

Table 1: Simulated district-level data iputs

district	pop15to49	adultart	nsamp	npos	nrecent	prev_est	prev_se	ANC prev	ANC artcov
Chitipa	116000	3900	218	18	1	8.3	2.6	5.3	23.8
Karonga	175000	9600	322	24	0	7.5	2.1	6.4	50.6
Rumphi	113000	6200	209	5	0	2.4	1.5	4.4	55.8
Nkhata Bay	134000	6800	247	21	0	8.5	2.5	5.8	48.1
Mzimba	580000	28200	1045	80	1	7.7	1.2	4.6	45.5
Likoma	7000	400	14	1	0	7.1	9.7	6.5	41.2
Nkhotakota	191000	8500	332	25	0	7.5	2.0	4.8	49.5
Kasungu	397000	11600	764	34	0	4.5	1.1	3.3	36.2
Ntchisi	143000	3400	229	10	1	4.4	1.9	3.1	37.7
Dowa	367000	9300	674	28	1	4.2	1.1	2.6	46.0
Mchinji	302000	10900	596	28	0	4.7	1.2	4.1	49.8
Salima	215000	11000	351	30	0	8.5	2.1	5.5	51.3

district	pop15to49	adultart	nsamp	npos	nrecent	prev_est	prev_se	ANC prev	ANC artcov
Lilongwe	1358000	70500	2462	208	3	8.4	0.8	6.1	46.5
Dedza	377000	12200	712	30	0	4.2	1.1	3.5	41.9
Ntcheu	282000	16900	531	50	0	9.4	1.8	7.3	54.5
Mangochi	476000	31000	811	98	2	12.1	1.6	8.8	45.3
Machinga	293000	20100	529	57	0	10.8	1.9	7.4	50.3
Balaka	183000	14900	335	48	0	14.3	2.7	9.6	48.2
Zomba	401000	37000	769	86	1	11.2	1.6	11.0	53.0
Neno	70000	6500	131	18	0	13.7	4.3	8.8	76.9
Mwanza	64000	4100	114	16	1	14.0	4.6	7.2	53.4
Blantyre	702000	61500	1271	220	4	17.3	1.5	12.8	43.4
Phalombe	191000	21700	376	48	0	12.8	2.4	13.1	58.9
Chiradzulu	170000	31600	317	68	1	21.5	3.3	15.6	61.2
Mulanje	303000	34500	541	85	2	15.7	2.2	13.5	48.6
Thyolo	357000	39800	601	106	0	17.6	2.2	12.7	44.3
Chikwawa	259000	18500	478	50	0	10.5	2.0	9.6	52.4
Nsanje	135000	14500	276	35	0	12.7	2.8	10.0	53.8