Estimating HIV Epidemics At Fine Scales

Le Bao
Department of Statistics
The Pennsylvania State University

joint work with UNAIDS reference group

March 29, 2016
CTSI BERD Seminar
Outline

Background

Estimation and Projection of HIV Epidemics

Understanding HIV Epidemic at Finer Scales

Results

Summary
Overview of HIV Epidemics

The global HIV epidemic is one of the greatest threats to human health and development:

▶ The number of people living with HIV continued to grow, reaching 36.9 million, about three times more than in 1990.
▶ 2 million people became newly infected with HIV.
▶ There have been 76 million infected with HIV and over 34 million AIDS-related deaths so far.
Great progress has been made:

- Prevention: 38% decline in new infections since 2001;
- PMTCT: 58% drop of new infections in children since 2002;
- Treatment: 35% fall in the AIDS-related death since 2005.
Great progress has been made:

- Prevention: 38% decline in new infections since 2001;
- PMTCT: 58% drop of new infections in children since 2002;
- Treatment: 35% fall in the AIDS-related death since 2005.

However, there is still no cure or low cost treatment; certain areas and key populations are left behind.

Sustainable Development Goals: ending the AIDS epidemic by 2030.
Fewer than 200,000 new HIV infections globally.
Ending the AIDS epidemic by 2030

“To reach the visionary goal of ending the AIDS epidemic after three decades of the most serious epidemic in living memory, countries will need to use the powerful tools available, hold one another accountable for results and make sure that no one is left behind.”

It requires understanding of the HIV-epidemic at finer scales.

- Integrating multiple data sources;
- Sharing information across areas and key populations.
Outline

Background

Estimation and Projection of HIV Epidemics

Understanding HIV Epidemic at Finer Scales

Results

Summary
The Estimation and Projection Package (EPP)

Spectrum/EPP was developed by UNAIDS reference group in 1997.

In 2015, 163 countries used Spectrum/EPP to estimate the impact of HIV on their population at the national level.

A susceptible-infected model divides the adult population at time $t$ into two groups:

- $Z(t)$: not-infected group
- $Y(t)$: infected group
Differential equations are used to describe the disease dynamic:

\[
\begin{aligned}
\frac{dZ(t)}{dt} &= E(t) - r(t)\rho(t)Z(t) - \mu(t)Z(t) - a_{50}(t)Z(t) + M(t)Z(t), \\
\frac{dY(t)}{dt} &= r(t)\rho(t)Z(t) - \text{HIVdeath}(t) - a_{50}(t)Y(t) + M(t)Y(t),
\end{aligned}
\]

which involve

- A set of fixed parameters obtained from external data sources;
- A set of free parameters, \( \theta \), with informative priors;
- Once given \( \theta \), EPP outputs the epidemic details over time.

(Bao et. al., 2010; Hogan et. al. 2010)
Prevalence Data

Commonly used datasets are the proportions of HIV positive cases estimated among antenatal clinic or STD clinic patients.
Likelihood for Clinic Data

The observed prevalence, $y_{st}$, and the EPP output prevalence, $\rho_t$, are linked through a mixed effects model:

$$
\Phi^{-1}(y_{st}) = \Phi^{-1}(\rho_t) + \beta + b_s + \epsilon_{st}, \quad s = 1, \ldots, S.
$$

$$
b_s \sim N(0, \sigma^2),
$$

$$
\epsilon_{st} \sim N(0, \nu_{st}).
$$

- $y_{st}$: the proportion of HIV+ patients at clinic $s$ in year $t$;
- $\rho_t$: EPP outcomes which are yearly HIV prevalence rates at the national level.
- $\beta$: the bias of the data source w.r.t. prevalence data from national population-based household surveys (NPBS).
- $b_s$: the clinic-specific random random effect.
Model Outputs

EPP provides estimates of many key epidemic indicators.
A Bayesian Model for EPP

- $q(\theta)$: the prior for inputs, coming from expert knowledge.
- EPP: maps $\theta$ to epidemic details such as prevalence $\rho$, incidence $I$, mortality $\mu$.
- D (Data): are sample proportions of HIV positive individuals.
- $L(D|\rho)$: likelihood is calculated via a linear mixed effect model with site-specific random effects.

(Alkema, Raftery and Clark 2007)
Outline

Background

Estimation and Projection of HIV Epidemics

Understanding HIV Epidemic at Finer Scales

Results

Summary
Estimating HIV Sub-Epidemics

- More accurate information at sub-national and sub-population levels are needed.

- The epidemic models have been applied *independently* to each sub-epidemic.

- Imbalanced data quality and availability may lead to unreliable results derived from the independent models.
Example of Sparse Data

Thailand Indirect Sex Workers Area 4

- EPP w/o imputation
- Training data
- Imputed data

Year
Prevalence(%)
Hierarchical Model Setup

Parameters $\theta$'s across areas might be correlated:

$$E(\theta_{\text{urban}}) = \theta_{\text{country}}$$

$$E(\theta_{\text{rural}}) = \theta_{\text{country}}$$

Urban Data

Rural Data

Figure: 1. A hierarchical model for input parameters of EPP.

The sub-national parameter estimates affect each other through the national parameter estimates.

(Bao et. al., 2014)
Hierarchical Model Setup

\[ \theta_{\text{country}} \]

\[ E(\theta_{\text{urban}}) = \theta_{\text{country}} \]

\[ E(\theta_{\text{rural}}) = \theta_{\text{country}} \]

Urban Data

Rural Data

Figure: 1. A hierarchical model for input parameters of EPP.

Challenge:

- Estimating multiple dynamic systems simultaneously;
- The dynamic system has no close form solution!
We propose a simple and innovative way to incorporate the hierarchical information into the dynamic systems:

- Apply hierarchical models to datasets across sub-epidemics without using any dynamic system.
- Incorporate results in EPP via auxiliary data.
Generalized Linear Mixed Effect Models (GLMM)

GLMM comes in handy when analyzing data with the hierarchical structure or the spatial dependence.

In a scenario of multiple areas and multiple surveillance sites within each area, let $i$, $a$, $t$ indicate site, area, and time.

$$Y_{ia}(t) \sim \text{Binomial}(n_{iat}, \rho_{ia}(t))$$

$$\text{logit}(\rho_{ia}(t)) = f_a(t) + b_a + c_{i(a)},$$

where $f_a(t)$ determines the area level time trend, $b$ and $c$ are the random coefficients at area and site level.

Ideally, $f_a(t)$ could be the prevalence trend produced by EPP.
GLMM with natural splines

Here, we model the prevalence trends by splines which introduces great flexibility and relieves the pressure of over-fitting.

\[ f_a(t) = \beta_{a,0} + \beta_{a,1} f_1(t) + \beta_{a,2} f_2(t) + \ldots + \beta_{a,K+3} f_{K+3}(t). \]

The natural cubic spline with equally spaced knots are used. \( \beta \)'s include both fixed effects and random effects.

Posterior distributions of \( f_a(t) \) can be approximated efficiently via MCMC.

Different models can be compared by DIC.
Thailand Indirect Sex Workers Area 4

- **EPP w/o imputation**
- **Training data**
- **Imputed data**


Prevalence (%): 0, 2, 4, 6, 8, 10, 12

Prevalence Fitted by GLMM
Bayesian melding: the prior distributions are defined on both

- the input parameters of the dynamic system, \( \theta \);
- the output quantities such as HIV prevalence, \( \rho(t) \).

We take the posterior distribution of the prevalence from the GLMM as the prior information of \( \rho(t) \).

The procedure is further simplified by replacing the prevalence prior by auxiliary data.
Auxiliary data approach

Let \((\mu_{a,g,t}, \nu_{a,g,t})\) be the posterior mean and variance of HIV prevalence in area \(a\), group \(g\), and year \(t\) from GLMM.

Convert the posterior to a binomial observation with prevalence \(\mu_{agt}\) and sample size \(n_{agt} = \frac{\mu_{at}(1-\mu_{at})}{\nu_{at}}\).
Auxiliary data approach

It simplifies the communication between GLMM and EPP:

- GLMM runs outside EPP;
- A new row of auxiliary data is added to EPP spreadsheet.
- Apply EPP to both original data and auxiliary data.

The auxiliary data contains the information from other areas and risk groups.
Auxiliary data approach

The strength of prior derived from GLMM can be easily adjusted by alternating the sample size of auxiliary data.

Let $n_{a,g} = \sum_t n_{a,g,t}$ be the total auxiliary data sample size for a sub-epidemic. It can be set to any predetermined number $K$:

- $K = 0$ corresponds to fitting Spectrum/EPP to data without borrowing information from other areas or groups.
- As $K$ increases, the “prior” prevalence distribution becomes more informative.
- The default value of $K$ will be determined based on test data evaluations in multiple countries.
Model Evaluation

For each sub-epidemic, randomly partition observations into training and test sets.

Compare the observations in the test data with their corresponding predictive distributions:

- $y_{it}$: an observed prevalence at the probit scale in the test data.
- $y_{it}^{(j)}$: the $j$th posterior sample of $y_{it}$ generated from the predictive distribution $P(Y_{it}|\text{training data})$.

Mean absolute error (MAE): $|y_{it} - y_{it}^{(j)}|$ averaged overall all observations in test data and posterior samples.

Coverage and the average width of the 95% prediction intervals.
Outline

Background

Estimation and Projection of HIV Epidemics

Understanding HIV Epidemic at Finer Scales

Results

Summary
Nigeria Example (37 areas)

Nigeria represents countries with high (>1%) HIV prevalence.

1. For each clinic, randomly assign 50% to the training and 50% to the test.
2. Apply GLMM to training data across all areas, and create auxiliary prevalence data for each area.
3. For each area, apply the EPP model to the training data with the addition of auxiliary data of size, $K = 0$ (original EPP), 50, 100, 200, 500, 1, 000, 2, 000, 5, 000, 10, 000.
4. Evaluate MAE, coverage and width of 95% prediction intervals on test data.
Nigeria Example

Nigeria 11

- EPP w/o imputation
- Training data

Year
Prevalence (%)
Nigeria Example

Nigeria 11

Prevalence (%)

- EPP w/o imputation
- EPP with imputation
- Training data
- Test data
- Imputed data
We calculate the MAE reduction as \( \frac{\text{MAE}_0 - \text{MAE}_K}{\text{MAE}_0} \).

(a) Reduction of Mean Absolute Error

Each gray line corresponds to one area in one training-test split. The black solid line provides the median and the 95% interval.
(b) Width Reduction of 95% Credible Intervals

New approaches narrows the credible interval.
For each training-test split, we combine the test datasets of 37 areas, and define the coverage as the proportion of all test data points that fall within the 95% credible intervals.
Thailand Example (4 areas, 3 key populations)

Thailand represents epidemics where the HIV prevalence < 1% and largely concentrated among specific sub-populations.

Reliable surveillance data is available for pregnant women, direct and indirect sex worker. Data ranges from 1989 to 2011.

1. Randomly select 3 sites and then 3 years of data from each selected site as training data.
2. Apply GLMM to training data across all sub-epidemics, and generate auxiliary prevalence data.
3. Within each area and each group, apply EPP model to the training data and corresponding auxiliary data with $K = 0, 50, 100, 200, 500, 1, 000, 2, 000, 3, 000, 4, 000, 5, 000, 10, 000$.
4. Evaluate model performance on test data.
(a) MAE Reduction, Pregnant Women

Each gray line corresponds to one area in one training-test split. The black solid line and dashed lines provide the median and (2.5th, 97.5th) quantiles.
(b) MAE Reduction, Indirect Sex Workers

Each gray line corresponds to one area in one training-test split. The black solid line and dashed lines provide the median and (2.5th, 97.5th) quantiles.
Each gray line corresponds to one area in one training-test split. The black solid line and dashed lines provide the median and (2.5th, 97.5th) quantiles.
Thailand Indirect Sex Worker Example

Thailand Indirect Sex Workers Area 1

- EPP w/o imputation
- Training data
- Imputed data

Year

Prevalence (%)

0 5 10 15 20


Year

Prevalence (%)
Thailand Indirect Sex Worker Example

Thailand Indirect Sex Workers Area 1

- EPP w/o imputation
- EPP with imputation
- Training data
- Test data
- Imputed data
Thailand Indirect Sex Worker Example

Thailand Indirect Sex Workers Area 2

- EPP w/o imputation
- Training data
- Imputed data

Year
Prevalence (%)

0 5 10 15 20
Thailand Indirect Sex Worker Example

Thailand Indirect Sex Workers Area 3

- EPP w/o imputation
- EPP with imputation
- Training data
- Test data
- Imputed data

Prevalence (%)

Year

Thailand Indirect Sex Worker Example

Thailand Indirect Sex Workers Area 4

- EPP w/o imputation
- Training data
- Imputed data
Thailand Indirect Sex Worker Example

Thailand Indirect Sex Workers Area 4

- EPP w/o imputation
- EPP with imputation
- Training data
- Test data
- Imputed data

Prevalence(%) vs. Year (1970-2010)
Outline

Background

Estimation and Projection of HIV Epidemics

Understanding HIV Epidemic at Finer Scales

Results

Summary
The auxiliary data improve the prediction accuracy in many cases, and never substantially reduce the accuracy.

Only need to fit the GLMM once to extract the useful information: 5～30 minutes.

It is applicable to all models (EPP, r-spline, r-trend) and all data sources (HIV incidence, HIV deaths).

Other forms of GLMM can be considered (geo-spatial model).

UNAIDS reference group has recommended implement the above method into EPP.
Acknowledgements

The research work was supported by UNAIDS, NSF grant BCS0941553, and NSF IGERT, DGE1144860.

Thanks to Xiaoyue Niu, Ben Sheng and Yuan Tang for exploratory analysis and manuscript writing.

Thanks to Tim Brown, Jeff Eaton, Peter Ghys, Mary Mahy and Keith Sabin for discussion and suggestions.

Thanks to all for your attendance and advice!