Model criticism, comparison and selection in dynamic transmission models for HIV: Bayesian evidence synthesis

Anne Presanis
in collaboration with Daniela De Angelis
and the Health Protection Agency

MRC Biostatistics Unit, Cambridge

16 March 2011
All models are wrong
Groningen
Outline

1 Introduction
   - Motivation: HIV
   - Evidence synthesis

2 Modelling HIV prevalence and incidence
   - Prevalence model
   - Incidence model
   - A joint model for incidence and prevalence

3 Transmission modelling
   - Mixing patterns

4 Results
   - Posterior distributions

5 Model criticism & comparison
   - Deviance summaries
   - Influence & Identifiability

6 Concluding comments
Motivation

Human Immunodeficiency Virus

Estimates of HIV prevalence and incidence are essential for understanding and monitoring the epidemic, as well as for assessing the impact of public health interventions.

Challenges

- HIV has a long asymptomatic incubation period, so many infections undiagnosed
- Surveillance systems available only for certain risk groups and populations
- Surveillance and other survey/ad-hoc data subject to biases
- Data sometimes tell us only indirectly about the quantities of interest
Evidence synthesis - a long-established idea

Methods for combining evidence are *not* new:

- **The Bayesian paradigm**
  - combining prior knowledge with new [Bayes (1763), Efron (2010)]
- **Meta-analysis**
  - combining studies of same type
- **Confidence Profile Method** [Eddy *et al* (1992)]
  - combining information of different types/study designs (medical-decision making literature)
- **Multi-parameter evidence synthesis** [Spiegelhalter *et al* (2004), Ades & Sutton (2006)]
  - epidemiology
Introduction
Evidence synthesis

Statistical formulation

- Interest: estimation of $\theta = (\theta_1, \theta_2 \ldots, \theta_k)$ on the basis of a collection of data $y = (y_1, y_2 \ldots, y_n)$
- Each $y_i$ provides information on
  - a **single** component of $\theta$, or
  - a **function** of one or more components, i.e. on a quantity $\psi_i = f(\theta)$

Thus inference is conducted on the basis of both **direct** and **indirect** information.

- Maximum likelihood: $L = \prod_{i=1}^{n} L_i(y_i \mid \theta)$
- Bayesian: $p(\theta \mid y) \propto p(\theta) \times L$
Modelling HIV prevalence and incidence

Prevalence model

\[ \pi \rho (1 - \delta) \rho \pi \delta \sum_g \rho \pi \delta \]

Proportion in risk group

HIV prevalence

Proportion diagnosed

Stratified by time, risk group and region

NATSAL
Proportion of men who are MSM

UA surveys
Prevalence of undiagnosed infection

SOPHID
Proportion of diagnosed infection attributable to each group

SOPHID
Total number of diagnosed infections

A. M. Presanis (MRC BSU)

Bayesian dynamic transmission models

16 March 2011

6 / 23
Incidence from prevalence

- $e(t)$: Not at risk
- $s(t)$: Susceptible
- $u(t)$: Infected Undiagnosed
- $d(t)$: Diagnosed

\[
\begin{align*}
e(t) &= 1 - \rho(t) \\
s(t) &= (1 - \pi(t))\rho(t) \\
u(t) &= (1 - \delta(t))\pi(t)\rho(t) \\
d(t) &= \delta(t)\pi(t)\rho(t)
\end{align*}
\]

- $\rho(t)$: Proportion in risk group
- $\pi(t)$: HIV prevalence
- $\delta(t)$: Proportion diagnosed

\[
\begin{align*}
\rho(t) &= s(t) + u(t) + d(t) \\
\pi(t) &= (u(t) + d(t))/\rho(t) \\
\delta(t) &= d(t)/(u(t) + d(t))
\end{align*}
\]
A multi-state model

\[ s(t) \rightarrow \psi \rightarrow u(t) \rightarrow \alpha(t) \rightarrow d(t) \]

- \( \alpha(t) \): New 15 year olds
- \( \psi \): New MSM
- \( \lambda(t) \): Incidence
- \( \kappa(t) \): Diagnosis rate
- Migration in/outwards
- Exits due to age/death
Combined incidence and prevalence model

Initial state of system

\[ c \in \{e, s, u, d\} \]

\( \rho \)

\( \pi \)

\( \delta \)

\( \pi(1 - \delta) \)

\( \frac{\rho \pi \delta}{\sum_g \rho \pi \delta} \)

NATSAL
Proportion of men who are MSM

time \( t_1 \)

UA surveys
Prevalence of undiagnosed infection

SOPHID
Proportion of diagnosed infection attributable to each group

\( \theta \)

data

times \( \{t_2 \ldots t_K\} \)

\( \rho \)

\( \pi \)

\( \delta \)

\( \pi(1 - \delta) \)

\( \frac{\rho \pi \delta}{\sum_g \rho \pi \delta} \)

NATSAL
Proportion of men who are MSM

UA surveys
Prevalence of undiagnosed infection

SOPHID
Proportion of diagnosed infection attributable to each group
**Results**

Prior distribution: \( \lambda(t), \kappa(t) \sim \text{Unif}(0, 1) \)

Posterior distribution:

---

**Incidence rate**

**Diagnosis rate**
Parameterisation of $\lambda(t)$

Incidence $\lambda_J(t)$ is a function of *prevalence*, the *contact* structure and the *probability of transmission* given a contact.

Random mixing

$$\lambda_J(t) = \chi_J(t) \left\{ \tau_D \delta_L(t) \pi_L(t) g_{JL} + \tau_U (1 - \delta_L(t)) \pi_L(t) g_{JL} \right\}$$

Presanis et al, Biostatistics, *in press*
More realistic mixing patterns

Preferential mixing: avoiding diagnosed partners, so that prevalence of diagnosed infection in chosen partners is smaller than in all MSM:

$$\lambda(t) = \chi(t) \{\tau_D \phi \delta(t) \pi(t) + \tau_U (1 - \delta(t)) \pi(t)\}$$

Model 1 Random mixing: $\phi = 1$
More realistic mixing patterns

**Preferential mixing:** avoiding diagnosed partners, so that prevalence of diagnosed infection in chosen partners is smaller than in all MSM:

\[ \lambda(t) = \chi(t) \{ \tau_D \phi \delta(t) \pi(t) + \tau_U (1 - \delta(t)) \pi(t) \} \]

**Model 1** Random mixing: \( \phi = 1 \)

**Model 2** Completely avoid diagnosed partners ("serosorters"): \( \phi = 0 \), \( \lambda(t) = \chi(t) \{ \tau_U (1 - \delta(t)) \pi(t) \} \)
More realistic mixing patterns

Preferential mixing: avoiding diagnosed partners, so that prevalence of diagnosed infection in chosen partners is smaller than in all MSM:

$$
\lambda(t) = \chi(t) \{ \tau_D \phi \delta(t) \pi(t) + \tau_U (1 - \delta(t)) \pi(t) \}
$$

Model 1 Random mixing: $\phi = 1$

Model 2 Completely avoid diagnosed partners ("serosorters"): $\phi = 0$, $\lambda(t) = \chi(t) \{ \tau_U (1 - \delta(t)) \pi(t) \}$

Model 3 No information on proportion who avoid diagnosed partners: $\phi \sim \text{Unif}(0, 1)$
More realistic mixing patterns

**Preferential mixing:** avoiding diagnosed partners, so that prevalence of diagnosed infection in chosen partners is smaller than in all MSM:

\[
\lambda(t) = \chi(t) \{ \tau_D \phi \delta(t) \pi(t) + \tau_U (1 - \delta(t)) \pi(t) \}
\]

**Model 1** Random mixing: \( \phi = 1 \)

**Model 2** Completely avoid diagnosed partners ("serosorters"): \( \phi = 0 \), \( \lambda(t) = \chi(t) \{ \tau_U (1 - \delta(t)) \pi(t) \} \)

**Model 3** No information on proportion who avoid diagnosed partners: \( \phi \sim \text{Unif}(0, 1) \)

**Model 4** Informative prior: \( \phi \sim \text{Beta}(10, 40) \)
Incidence by diagnosis status of contact

Grey: $\lambda_U(t)$, incidence due to undiagnosed contacts
Magenta: $\lambda_D(t)$, incidence due to diagnosed contacts

Model 1
Model 2
Model 3
Model 4
Transmission probabilities

Prior distribution: $\tau_U \sim \text{Unif}(0, 0.3), \quad \tau_D \sim \text{Unif}(0, \tau_U)$

Posterior distribution:
### Deviance Information Criteria (DIC)

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>$\bar{D}$</th>
<th>$D(\bar{\theta})$</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>174</td>
<td>175.7</td>
<td>22.3</td>
<td>153.4</td>
<td>329.1</td>
</tr>
<tr>
<td>2</td>
<td>174</td>
<td>176.7</td>
<td>22.9</td>
<td>153.8</td>
<td>330.5</td>
</tr>
<tr>
<td>3</td>
<td>174</td>
<td>175.4</td>
<td>22.4</td>
<td>153.0</td>
<td>328.4</td>
</tr>
<tr>
<td>4</td>
<td>174</td>
<td>176.0</td>
<td>22.5</td>
<td>153.4</td>
<td>329.4</td>
</tr>
</tbody>
</table>

What influences the estimates of $\lambda_U$, $\lambda_D$?

$$\lambda(t) = \chi(t) \{\tau_D \phi \delta(t) \pi(t) + \tau_U (1 - \delta(t)) \pi(t)\}$$

- Equal fit to data informing prevalences, transition rates, so data *don't* have strong influence on estimates of $\lambda_U$, $\lambda_D$?
- Only *prior* information on $\tau_U$, $\tau_D$, $\phi$ and model *structure* having effect?
Transmission probabilities

**Red: Prior**  **Black: Posterior**

**Model 1**
Pr(T | undiagnosed contact)

**Model 2**
Pr(T | undiagnosed contact)

**Model 3**
Pr(T | undiagnosed contact)

**Model 4**
Pr(T | undiagnosed contact)

Pr(T | diagnosed contact)

Pr(T | diagnosed contact)

Pr(T | diagnosed contact)
Where is the information coming from?

Bounds are well identified, so $\tau_U, \tau_D$ partially identified

\[
\tau_D = \frac{\lambda(t) - \chi(t)\tau_U(1 - \delta(t))\pi(t)}{\chi(t)\phi\delta(t)\pi(t)}
\]

and

\[
0 \leq \tau_D \leq \tau_U
\]

\[
\Rightarrow \quad \frac{\lambda(t)}{\chi(t)(1 - \delta(t)(1 - \phi))\pi(t)} \leq \tau_U \leq \frac{\lambda(t)}{\chi(t)(1 - \delta(t))\pi(t)}
\]
$\tau_U$ with limits, e.g. in year 2004

- Lower limit
- $\Pr(\text{Transmission} \mid \text{Undiagnosed contact})$
- Upper limit

<table>
<thead>
<tr>
<th>Model</th>
<th>Lower limit</th>
<th>$\Pr(\text{Transmission} \mid \text{Undiagnosed contact})$</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lower limit informed by $\phi$

Red: Prior  Black: Posterior

Model 3

Model 4
Influence of model structure on incidence

\[ \lambda(t) = \lambda_U(t) + \lambda_D(t) \]

\[ = \chi(t) \{ \tau_D \phi \delta(t) \pi(t) + \tau_U (1 - \delta(t)) \pi(t) \} \]
Concluding comments

Summary

- DIC only take us so far in discriminating between models, when *priors/model structure* have more influence on differences in inferences than data?
- Importance of understanding *influence* of data, priors and model structure on inference (O’Hagan, 2003)
- Idea that models/model structure are “*priors*/indirect evidence” (Efron, Stat. Sci. 2010)
- How to *judge/discriminate between* different model assumptions in this context, other than by presenting sensitivity analyses?
- Here we conclude that *further data required* on behaviour change once diagnosed (φ) - important endpoint.
- *Partial* identifiability (Gustafson, Greenland) - but still able to infer \( \lambda_U, \lambda_D \) based on *mechanistic* model assumptions.
Further work

- Investigating sources of data for *behaviour*
- **Expansion** of model to 3 levels of risk amongst MSM?
- Understanding *influence* of each part of model (data, priors, structure)
  - e.g. *loosening* assumed model structure, by using beta-binomial and negative binomial likelihoods instead of binomial and Poisson
Acknowledgements

- Daniela De Angelis (MRC Biostatistics Unit & Health Protection Agency)
- Tony Ades (Bristol)
- Aicha Goubar (INVS, Paris)
- David Spiegelhalter, Paul Birrell, Chris Jackson (MRC Biostatistics Unit)
- Graham Medley (Warwick)
- HIV Department, Health Protection Agency
An aside on partial identifiability - Model 3
Further or less information

Model 5: $\tau_U \sim N(0.4, 0.18^2)T(0, 1)$ from meta-analysis of Baggaley (2006)

Model 6: $\tau_U \sim N(0.05, 0.005^2)T(0, 1)$

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>$\bar{D}$</th>
<th>D($\bar{\theta}$)</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>174</td>
<td>175.40</td>
<td>22.18</td>
<td>153.24</td>
<td>328.64</td>
</tr>
<tr>
<td>6</td>
<td>174</td>
<td>175.67</td>
<td>22.16</td>
<td>153.46</td>
<td>329.13</td>
</tr>
</tbody>
</table>
Transmission probabilities - densities

Model 5

Model 6

Model 5

Model 6
Transmission probabilities - traces

Pr(T | undiagnosed contact) for Model 5, showing variability over iterations.

Pr(T | diagnosed contact) for Model 5, also showing variability over iterations.