Using multiple sources to estimate disease burden

Daniela De Angelis

MRC Biostatistics Unit, Cambridge
and
Health Protection Agency, London
Outline

1. Motivations for the work
2. Case study - general formulation
3. Results
4. Discussion/current challenges
**Ultimate motivation**

Planning, implementation and evaluation of public health policies in the UK *e.g.*

- Department of Health (DoH) National Strategy for Sexual Health
- DoH Action Plan for Hepatitis C
- Pandemic influenza Preparedness Strategy

... rely on the monitoring of fundamental aspects of the disease of interest, such as

- prevalence (undiagnosed prevalence)
- incidence

- by age groups and locations
- at regular intervals (real time!)
Motivations for the work

Methodological motivation

- These characteristics are typically not easily directly measurable (if at all) with little *direct* information available on them.
- There is plenty of *indirect* information on these quantities (e.g. on functions) from diverse sources (surveillance, ad hoc surveys etc).
- Usually this *indirect* information is:
  - discarded as not directly relevant
  - used in a very ad hoc way
- Estimation from the *synthesis* of both *direct* and *indirect* information within a coherent framework feasible.
- This has been common problem underlying most of the work I have been recently involved with:

  [Goubar *et al*, 2008], [Sweeting *et al*, 2008], [De Angelis *et al*, 2009],
  [Conti *et al*, 2012], [Jackson *et al*, 2012], [Price *et al*, 2012].
Motivations for the work

Evidence synthesis - a long-established idea

Methods for combining evidence are not new:

- The **Bayesian** paradigm
  - combining prior knowledge with new evidence
- **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)]
  - health technology assessment
  - epidemiology
Case study

Estimation of HIV prevalence and incidence

- how tackled
  - conventionally
  - within an evidence synthesis approach
  - general formulation
  - epidemiological/methodological considerations and open questions
Infection with HIV

- HIV is a long incubation, asymptomatic disease and many infections undiagnosed.
- Undiagnosed infections contribute to transmission - lack of access to treatment.
- Reliable estimates of the number infected, particularly undiagnosed required to evaluate and plan interventions to reduce transmission - complicated precisely due to lack of symptoms.
- Holy grail: incidence - even more problematic.
Estimation problem: notation

$G$ non-overlapping risk groups - $R$ regions

Total population, region $r$, time $t$

$N_{t,r} \sum_{g} \rho_{t,g,r} \pi_{t,g,r}$

Sum over risk groups

Proportion of region $r$ in group $g$ at time $t$

HIV prevalence in group $g$, region $r$, at time $t$

Diagnosed

$N_{t,r} \sum_{g} \rho_{t,g,r} \pi_{t,g,r} \delta_{t,g,r}$

Undiagnosed

$N_{t,r} \sum_{g} \rho_{t,g,r} \pi_{t,g,r} (1 - \delta_{t,g,r})$
Estimation problem: basic parameters

- $\rho_{t,g,r}$ prevalence (i.e. the proportion in the population) of risk-group $g$ in the population at time $t$ for region $r$
- $\pi_{t,g,r}$ corresponding prevalence of HIV
- $\delta_{t,g,r}$ proportion of infections diagnosed in risk-group $g$, region $r$

Any other quantities can be derived from these
13 risk groups, 3 regions, over time

**MSM:** Men who have sex with men
- Current MSM, current STI clinic attendees
- Current MSM, not current STI attendees
- Past MSM

**IDUs:** Injecting drug users (non-MSM)
- Current IDU (men and women)
- EX IDU (men and women)

**SSA-born:** Heterosexual individuals born in Sub-Saharan Africa (non-IDUs)
- SSA-born (men and women)

**STI:** Heterosexual individuals (non-SSA), current STI attendees
- Current STI-attending (men and women)

**LR:** Lower risk heterosexual individuals (non-SSA, non-STI)
- Lower risk (men and women)
Traditionally

- Conventional approaches - ‘direct methods’- concentrate on the estimation of undiagnosed infections

\[ N_{t,r} \sum_{g} \rho_{t,g,r} \pi_{t,g,r} (1 - \delta_{t,g,r}) \]

- Simple idea: estimate \( \rho_{t,g,r}, \pi_{t,g,r} \) and \( \delta_{t,g,r} \) and multiply them to derive

\[ N_{t,r} \sum_{g} \hat{\rho}_{t,g,r} \hat{\pi}_{t,g,r} (1 - \hat{\delta}_{t,g,r}) \]

This was the approach used in the UK and other countries (e.g. Karon et al, 1998; Ramon et al, 2002; McGarrigle et al., 2006)
Problems with ‘direct methods’ (1)

Each parameter is informed by a single data item

- Direct information to estimate $\rho_{t,g,r}$, $\pi_{t,g,r}$ and $\delta_{t,g,r}$ not available for each $g$, $r$
- Wealth of indirect information (e.g. on mixtures of groups; on diagnosed individuals)
- Discarded as not easily incorporated in such a simplistic framework
- Not efficient use of available data with potential for biased results due to selective information
- Ad hoc assumptions made to compensate for lack of data
Problems with ‘direct methods’ (2)

- No explicit model - no notion of model fit
- No easy results validation
- No ability to quantify formally uncertainty around resulting estimates

Need for an alternative approach

- Synthesizes direct and indirect information
- Appropriately accounts for uncertainty
Evidence synthesis: general formulation

- Interest: estimation of \( \theta = (\theta_1, \theta_2 \ldots, \theta_K) \) basic parameters on the basis of a collection of \( n \geq K \) independent data items \( y = (y_1, y_2 \ldots, y_n) \)

- Each \( y_i \) provides
  - direct information on a single component \( \theta_k \) of \( \theta \), or
  - indirect information on one or more components.

- Denote by \( \psi_i = \psi_i(\theta) \) a generic function of \( \theta \) (i.e. \( \psi_i = \theta_i \), \( \psi_i = \psi_i(\theta_i) \) or \( \psi_i = \psi_i(\theta) \)). From the independence of the \( y_i \), the full data likelihood is

\[
L(\theta; y) = \prod_{i=1}^{n} L_i(\psi_i(\theta); y_i)
\]

where \( L_i(\psi_i(\theta); y_i) \) the likelihood contribution of \( y_i \) to the basic parameter vector \( \theta \)
Evidence synthesis: general formulation

Inference is conducted on the basis of both direct and indirect information.

- Maximum likelihood: \( L(\theta; y) \)
- Bayesian: \( p(\theta | y) \propto p(\theta) \times L(\theta; y) \)

with \( p(\theta) \) indicating a prior distribution on \( \theta \)
Direct Acyclic Graph (DAG) representation

\[ \theta_1 \cdots \theta_i \theta_{i+1} \cdots \theta_k \]

\[ \psi_1 \cdots \psi_j \psi_{j+1} \cdots \psi_n \]

\[ y_1 \cdots y_j y_{j+1} \cdots y_n \]
Case study - general formulation

Availability of data: 13 risk groups, 3 regions, over time

<table>
<thead>
<tr>
<th>Risk group</th>
<th>N</th>
<th>ρ</th>
<th>π</th>
<th>δ</th>
<th>ψ(ρ, π)</th>
<th>ψ(π, δ)</th>
<th>ψ(ρ, π, δ)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSM</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>IDUs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Born sub-Saharan Africa</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>STI clinic attendees</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lower risk</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDUs</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Born sub-Saharan Africa</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>STI clinic attendees</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower risk</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ALL</strong></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
Distributional assumptions

- Information \( y_{t,g,r} \) comes in the form \( \{x_{t,g,r}, n_{t,g,r}\} \) and can be naturally assumed to be realisations of a Binomial random variable

\[
X_{t,g,r} \sim \text{Binomial}(n_{t,g,r}, \psi_{t,g,r})
\]

\( \psi_{t,g,r} \) equals any of \( \rho_{t,g,r}, \pi_{t,g,r} \) and \( \delta_{t,g,r} \) if \( y_{t,g,r} \) provides direct information or is a function of these basic parameters otherwise.
Distributional assumptions

- Information $y_{t,g,r}$ comes in the form $\{x_{t,g,r}, n_{t,g,r}\}$ and can be naturally assumed to be realisations of a Binomial random variable

$$X_{t,g,r} \sim \text{Binomial}(n_{t,g,r}, \psi_{t,g,r})$$

$\psi_{t,g,r}$ equals any of $\rho_{t,g,r}, \pi_{t,g,r}$ and $\delta_{t,g,r}$ if $y_{t,g,r}$ provides direct information or is a function of these basic parameters otherwise

- or as counts: observed diagnoses in men and women $x_{t,m,r}$ and $x_{t,f,r}$ as $X_{t,m,r} \sim \text{Poisson}(\mu_{t,m,r})$ and $X_{t,f,r} \sim \text{Poisson}(\mu_{t,f,r})$ where

$$\mu_{t,m,r} = N_{t,m,r} \sum_{g_m} (1 - \nu_{t,g_m}) \delta_{t,g_m,r} \pi_{t,g_m,r} \rho_{t,g_m,r}$$

$$\mu_{t,f,r} = N_{t,f,r} \sum_{g_f} (1 - \nu_{t,g_f}) \delta_{t,g_f,r} \pi_{t,g_f,r} \rho_{t,g_f,r}$$

Note: $\nu_{t,g_m}, \nu_{t,g_f}$ are ‘bias’ parameters.
## Sparseness of information

<table>
<thead>
<tr>
<th>Risk group</th>
<th>N</th>
<th>$\rho$</th>
<th>$\pi$</th>
<th>$\delta$</th>
<th>$\psi(\rho, \pi)$</th>
<th>$\psi(\pi, \delta)$</th>
<th>$\psi(\rho, \pi, \delta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men</strong></td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>MSM</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDUs</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>Born sub-Saharan Africa</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STI clinic attendees</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower risk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>ALL</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td>✔</td>
</tr>
<tr>
<td>IDUs</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Born sub-Saharan Africa</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STI clinic attendees</td>
<td></td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower risk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALL</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sparseness of information

- This sparsity can be addressed by sharing information between men and women.
Sparseness of information

- This sparsity can be addressed by sharing information between men and women.
- $\pi_{t,g,r}$ and $\delta_{t,g,r}$ are expected to vary by gender and by location.
Sparseness of information

- This sparsity can be addressed by sharing information between men and women.
- $\pi_{t,g,r}$ and $\delta_{t,g,r}$ are expected to vary by gender and by location.
- Reasonable to assume that their male-to-female odds ratio might be similar between regions.

\[
\begin{align*}
\logit(\pi_{t,g,m,r}) &= \lor.\pi_{t,g,r} + \logit(\pi_{t,g,f,r}); \quad \lor.\pi_{t,g,r} \sim N(P_{t,g}, \sigma_{t,\pi}^2) \\
\logit(\delta_{t,g,m,r}) &= \lor.\delta_{t,g,r} + \logit(\delta_{t,g,f,r}); \quad \lor.\delta_{t,g,r} \sim N(D_{t,g}, \sigma_{t,\delta}^2)
\end{align*}
\]

with a further hierarchy over risk groups:

\[
P_{t,g} \sim N(\Pi_t, \omega_{t,\pi}^2); \quad D_{t,g} \sim N(\Delta_t, \omega_{t,\delta}^2).
\]
Priors

- Diffuse Uniform priors on $\pi_{t,g,r}$ and $\delta_{t,g,r}$.
- Dirichlet priors on the proportions of the male and female populations in each risk group $\rho_{t,gm,r}$ and $\rho_{t,gf,r}$.
- Informative Normal or Uniform priors are assigned to bias parameters such as $\nu_{t,gm}$ and $\nu_{t,gf}$.
- The means $\Pi_t$ and $\Delta_t$ are a priori distributed as $\text{Normal}(0, 100^2)$. The standard deviations $\sigma_{t,\pi}$, $\sigma_{t,\delta}$ and $\omega_{t,\delta}$ are given informative priors.
Inference

- Performed on the basis of the posterior

\[ p(\theta_t \mid y_t) \propto p(\theta_t) \times L_t(\theta_t; y_t) \]

- Samples from \( p(\theta_t \mid y_t) \) obtained through MCMC
### Results

Posterior median (95% CrI) number of adult HIV infections in E&W in 2008, by group and diagnosis status

<table>
<thead>
<tr>
<th>Group</th>
<th>Total Infections</th>
<th>Diagnosed Infections</th>
<th>Undiagnosed Infections</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSM men</td>
<td>16,050</td>
<td>7,050</td>
<td>8,950</td>
</tr>
<tr>
<td>MSM women</td>
<td>15,000</td>
<td>7,000</td>
<td>8,000</td>
</tr>
<tr>
<td>SSA men</td>
<td>4,650</td>
<td>2,350</td>
<td>2,300</td>
</tr>
<tr>
<td>SSA women</td>
<td>4,600</td>
<td>2,300</td>
<td>2,300</td>
</tr>
<tr>
<td>STI + LR men</td>
<td>11,450</td>
<td>7,450</td>
<td>3,950</td>
</tr>
<tr>
<td>STI + LR women</td>
<td>11,450</td>
<td>7,450</td>
<td>3,950</td>
</tr>
<tr>
<td>IDU men</td>
<td>800</td>
<td>350</td>
<td>450</td>
</tr>
<tr>
<td>IDU women</td>
<td>550</td>
<td>400</td>
<td>150</td>
</tr>
</tbody>
</table>

Total infected = 57,600 (53,650 – 62,200)
Total infected diagnosed = 40,300 (38,950 – 41,750)
Total infected undiagnosed = 17,250 (13,500 – 21,750)
Trends in prevalence in MSM by diagnosis status

Daniela De Angelis (MRC-BSU, HPA)
Rennes, September 2012
Prevalence snapshots

\[ e_t \rightarrow \text{Not at risk} \]
\[ s_t \rightarrow \text{Susceptible} \]
\[ u_t \rightarrow \text{Infected Undiagnosed} \]
\[ d_t \rightarrow \text{Diagnosed} \]

Proportion in risk group: \( \rho_t = s_t + u_t + d_t \)
HIV prevalence: \( \pi_t = (u_t + d_t)/\rho_t \)
Proportion diagnosed: \( \delta_t = d_t/(u_t + d_t) \)
A compartmental model

Transition rates piecewise constant:

\[ \lambda_t = \{ \lambda^e_t, \lambda^{e,s}_t, \text{demographics}, \lambda^{s,u}_t, \lambda^{u,d}_t \} \]

\[ \lambda_t = \text{piecewise constant over time} \]
Joint incidence and prevalence model

\[ c_t = \{ e_t, s_t, u_t, d_t \}, \quad \theta_t = \{ \rho_t, \pi_t, \delta_t \} \]
Inference

- Performed on the basis of the posterior

\[ p(\theta_t, \lambda_t \mid y_t, z_t) \propto p(\theta_t)p(\lambda_t) \times L_t(\theta_t, \lambda_t; y_t, z_t) \]

where \( L_t(\theta_t, \lambda_t; y_t, z_t) = \prod_{t=1}^{T} L_t(\theta_t; y_t) L_t(\lambda_t; z_t) \)

**Note:** \( L_t(\theta_t; y_t) \) depends on \( c_t \), numerically evaluated at each MCMC iteration.
Joint incidence and prevalence model
Incidence

HIV incidence rate

2001 2002 2003 2004 2005 2006 2007 2008 2009
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Results

Diagnosis rate

HIV diagnosis rate

Daniela De Angelis (MRC-BSU, HPA)
Rennes, September 2012
Prevalence

HIV prevalence

Daniela De Angelis (MRC-BSU, HPA)
Rennes, September 2012
Discussion

- Example of a successful application of Bayesian evidence synthesis
- Since 2005 this is the method adopted in the UK to produce official estimates of HIV burden
- The approach further developed to tackle changes in surveillance systems
- International interest
Discussion

- The method responds formally to the need to ‘triangulate’ diverse and multiple sources
- Demonstrated how it can make full use of all available information (minimising selection biases; leading to more accurate inference)
- Accounts for all uncertainties, reflected in the resulting posterior distribution
- Why Bayesian?
Current challenges

- Use of multiple sources of evidence leads to complex probabilistic models
- Increasingly expert at formulating and estimating such models
- The availability of a well-defined iterative process of model criticism lags behind this expertise
- Model criticism becomes more crucial but harder as the number of data sources increases and the model becomes more complex
- Work on approaches to conflict detection, model choice and assessment needed
Co-authors

- **Statisticians**
  - Anne Presanis, David Spiegelhalter MRC-BSU
  - Tony Ades, Bristol University

- **Epidemiologists**
  - Noel Gill and HIV Department, HPA