A Bayesian approach to estimating disease prevalence using information from multiple sources: HIV and HCV in England and Wales

Daniela De Angelis

Health Protection Agency, Centre for Infections, London
and
MRC Biostatistics Unit, Cambridge

Open University, 20th May 2009
Outline

1. Motivation for the work
2. General problem
3. Statistical formulation
4. Illustration: HCV prevalence estimation
5. Results
6. Considerations/open questions
Motivation

Implementation and evaluation of public health policies aimed to control and prevent epidemics rely crucially on the knowledge of fundamental aspects of the disease of interest, such as prevalence and incidence in a particular age group and location.
These characteristics are typically not easily measurable as little direct data is available on them.

There is plenty of indirect information on functions of these quantities.

Estimation through the synthesis of these diverse and fragmented sources of evidence feasible.

This has been common problem underlying most of the work I have been recently involved with, to reconstruct characteristics of blood borne virus epidemics:

Interest: estimation on $\boldsymbol{\theta} = (\theta_1, \theta_2 \ldots, \theta_k)$ on the basis of a collection of data $\mathbf{y} = (y_1, y_2 \ldots, y_n)$

Each $y_i$ provides information on
- a component of $\boldsymbol{\theta}$, or
- a function of such component (or more than one), i.e. on a quantity $\psi_i = f(\boldsymbol{\theta})$

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

**Note:** it is likely that $n > k$. This might create inconsistency between sources!

[Eddy, 1992], [Ades,Sutton 2006]
Inference

- From a **likelihood** point of view:
  
  choose \( \theta \) to maximise
  
  \[
  L = \prod_{i=1}^{n} L_i(y_i \mid \theta)
  \]

  where \( L_i(y_i, \theta) \) is the likelihood for the \( i^{th} \) data item.

- From a **Bayesian** point of view:
  
  if \( p(\theta) \) is the prior distribution for \( \theta \), derive the posterior
  
  \[
  p(\theta \mid y) \propto p(\theta) \times L
  \]
Infection with the Hepatitis C virus (HCV): background

- Acquired through exchange of blood with infected individual (e.g. injecting drug use)
- Disease with long incubation; progressive fibrosis of the liver to cirrhosis, hepatocellular carcinoma and death
- Antiviral therapy very effective
- Planning for resources for prevention and treatment implementation needs reliable estimates of the number currently infected
HCV prevalence estimation

No prevalence study/surveillance representatively covering the general population exists

- Estimates of proportion of infected derived in specific (opportunistic) groups:
  - Injecting drug users (IDUs) attending treatment clinics
  - Women attending antenatal/neonatal screening
  - Individuals attending Genito-Urinary Medicine (GUM) clinics (residual sera)
  - Individuals of all ages having routine diagnostic blood tests (residual sera)
• Resulting estimates are not interpretable in terms of prevalence in the general population as these groups are mixtures of sub-groups (e.g. IDUs) with different HCV risk

• To relate these estimates to HCV prevalence in the general population information is needed on mixture composition and size of the groups at risk of HCV in the population
Coherent framework needed

- Formally expresses the nature of available data (i.e. as mixture of groups with different HCV risk)
- Incorporates additional (uncertain) information on the mixture composition
- Allows incorporation of data on size of the risk groups
- Allows inclusion of potential biases
Proposed Approach

- Information from any available study, either on HCV prevalence or on the size of the groups at risk for HCV, is expressed in terms of 3 main risk groups $g$:
  - current injecting drug users ($g = CUR$)
  - Ex-injecting drug users ($g = EX$)
  - Non-IDUs ($g = NON – IDU$)

specified for 3 different regions $r$: London, North West England, Rest of England and Wales;
3 age groups $a$: $[15, 29)$, $[30, 44)$, $[45, 59)$. 
Parameters of interest

\[ \rho_{\text{CUR} r s a}, \rho_{\text{EX} r s a}, \rho_{\text{NON-IDU} r s a} \]

prevalence (i.e. the proportion) of the current IDU, ex-IDU, and Non-IDU risk-groups in the population at a given time \( T \) for region \( r \), gender \( s \), and age-group \( a \).

- \( \rho_{\text{NON} r s a} = 1 - \rho_{\text{CUR} r s a} - \rho_{\text{EX} r s a} \)
- \( \rho_{\text{IDU} r s a} = \rho_{\text{CUR} r s a} + \rho_{\text{EX} r s a} \)

\[ \pi_{\text{CUR} r s a}, \pi_{\text{EX} r s a}, \pi_{\text{NON-IDU} r s a} \]

corresponding prevalence of HCV.
Derived quantities

- **HCV prevalence in a particular group** (e.g. current IDUs)

\[
\pi_{CUR} = \frac{\sum_r \sum_s \sum_a N_{rsa} \rho_{CURrsa} \pi_{CURrsa}}{\sum_r \sum_s \sum_a N_{rsa} \rho_{CURrsa}}
\]

- **HCV prevalence in region** \(r\), **gender** \(s\) and **age-group** \(a\)

\[
\pi_{rsa} = \rho_{CURrsa} \pi_{CURrsa} + \rho_{EXrsa} \pi_{EXrsa} + \rho_{NON-IDUrsa} \pi_{NON-IDUrsa}
\]

- **Overall HCV prevalence**

\[
\pi = \frac{\sum_r \sum_s \sum_a N_{rsa} \pi_{rsa}}{\sum_r \sum_s \sum_a N_{rsa}}
\]

where \(N_{rsa}\) is the size of the general population in region \(r\), gender \(s\) and age-group \(a\)
Data on $\rho_{grsa}$

- **Capture re-capture study** in 15-44 years old in London estimate of number of current IDUs

- **Household surveys**:
  - **British Crime Survey** (HO)
    - use of heroin - ever - past year
    - time since starting in ex-users
  - **Survey of Psychiatric Morbidity** (ONS)
    - use of heroin - ever - past year
  - **Offending Crime and Justice Survey** (HO)
    - use of heroin ever - past year
    - length of heroin use and time since starting in ex-users
  - **National Survey of Sexual Attitudes and Lifestyles** (NATSAL)
    - use of non-prescribed IDU drugs - ever - past year
Data on $\pi_{grsa}$

- **UA programme in genito-urinary medicine (GUM) clinics**
  - HCV prevalence in ever IDUs and non-IDUs

- **UA programme in current IDUs attending specialist clinics**
  - HCV prevalence in current IDUs

- **UA programme in antenatal clinics and neonatal samples**
  - HCV prevalence in pregnant women

- **Studies in blood donors**
  - HCV prevalence in low risk population

- **Sentinel laboratory surveillance**
  - HCV prevalence in populations testing for HCV
Challenges

- data structure simple as mostly of the form \( \{ r_{grsa}, n_{grsa} \} \)
- But the observed proportions are typically
  
  - biased estimates of the true proportions of interest (e.g. size of the populations)
  
  - only interpretable as mixtures of proportions
- lack of direct information on specific proportions of interest (e.g. the size of the ex-IDU population)
Challenges: example (1)

- Household studies
  
  $i^{th}$ study provides information on $\rho_{grsa}$ in the form of $\{r_{grsa}^i, n_{grsa}^i\}$. We assume that

  $$ r_{grsa}^i \sim \text{Binomial}(n_{grsa}^i, \rho_{grsa}^i) $$

  and

  $$ \text{logit}(\rho_{grsa}^i) = \text{logit}(\rho_{grsa}) + b $$

  where $b$ is a bias parameter.
Mixture of proportions

the HCV prevalence in GUM clinic attendees who have ever injected

\[ \pi_{\text{GUM IDUsa}} \]

estimated using \( \{r_{\text{GUM IDUsa}}, n_{\text{GUM IDUsa}}\} \) can only be interpreted as being

\[ \pi_{\text{GUM IDUsa}} = \psi \pi_{\text{GUM CURRs}} + (1 - \psi) \pi_{\text{GUM EXsa}} \]

where the mixture coefficient \( \psi \) is informed by the NATSAL survey. Thus the contribution to the likelihood is of the kind

\[ r_{\text{GUM IDUsa}} \sim \text{Binomial}(n_{\text{GUM IDUsa}}, \pi_{\text{GUM IDUsa}}) \]
Challenges: examples (3)

- Lack of direct data on $\rho_{EXrsa}$
  - $\rho_{EXrsa}$ can be written as function of $\rho_{IDUrsa}$ as
    \[
    \rho_{EXrsa} = \rho_{IDUrsa} \kappa_{EXrsa}
    \]

  where $\kappa_{EXrsa}$ is the probability of being an ex-user at the current time $T$, conditionally on being an ever IDU, and can be expressed as

  \[
  \kappa_{EXrsa} = \sum_{t=0}^{T_{max}} F_{D|r_{rsa}}(t)f_{TSS|r_{rsa}}(t)
  \]

  with $F_D$ and $f_{TSS}$ indicating the cumulative distribution function and the density function of the injecting duration $D$ and the time since starting $TSS$. 
Analogously

\[ \rho_{CURsa} = \rho_{IDUsa} (1 - \kappa_{EXsa}) \]

Since the \( f_{TSS|r,s,a}(t) \) can be written in terms of its marginal distribution \( f_{TSS}(t) \) and the age at first use distribution \( f_{AAFU} \), then

\[ \rho_{EXsa} = g_1(\rho_{IDUsa}, f_D, f_{TSS}, f_{AAFU}) \]

and

\[ \rho_{CURsa} = g_2(\rho_{IDUsa}, f_D, f_{TSS}, f_{AAFU}) \]
Similarly for the $\pi_{EXrsa}$

$$\pi_{EXrsa} = g_3(\pi_{IDUrsatss}, f_D, f_{TSS}, f_{AAFU})$$

$$\pi_{CURrsa} = g_4(\pi_{IDUrsatss}, f_D, f_{TSS}, f_{AAFU})$$
Illustration: HCV prevalence estimation

Graphical model

- **Direct information**
- **Indirect information**

Parameters:
- $f_{AAFU}$
- $f_{TSS}$
- $f_D$

Nodes:
- $\rho_{IDU}$
- $\rho_{CUR}$
- $\rho_{EX}$
- $\pi_{IDU}$
- $\pi_{CUR}$
- $\pi_{EX}$
- $\pi_{NON-IDU}$

Studies:
- Capture Re-capture Study
- House-hold Studies
- Additional information
- Tested population
- Ante/neo Natal surveys
- UA IDU
- GUM UA
- Blood transfusions

Tested population

- UA IDU
- GUM UA
- Blood transfusions
Estimation

- Bayesian estimation not feasible analytically
- simulation (Markov chain Monte Carlo) used to generate the posterior distributions
Posterior distributions for the number of IDU, E&W 2003

Ex-injecting drug users

Current injecting drug users

Number of IDUs (thousands)
Results

Posterior distributions for the number of individuals with anti-HCV antibodies, E&W 2003
Methodological issues

Powerful approach that allows use of all available information inevitably leading to complex probabilistic models

- How do we assess these complex models?
  - adequacy as representation of the data
  - appropriateness compared to alternative models
  - identification of the ‘drivers’ of the resulting inference
Model assessment: instruments

Deviance

\[ D(\theta) = -2[\log p(y | \theta) - \log p(y | \hat{\theta})] \]

- Posterior mean deviance (\( \bar{D} \))

\[ \bar{D} = E_{\theta|y}[D(\theta)] \]

- Deviance Information Criteria (\( DIC \))

\[ DIC = \bar{D} + 2p_D \]

[Spiegelhalter et al, 2002]
## Model assessment: model choice

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
<th>$\rho_{CUR}$ (%)</th>
<th>$\rho_{EX}$ (%)</th>
<th>$\pi_{CUR}$ (%)</th>
<th>$\pi_{EX}$ (%)</th>
<th>$\pi_{NON}$ (%)</th>
<th>$\pi$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No bias $b = 0$</td>
<td>1022</td>
<td>0.26</td>
<td>0.73</td>
<td>33.7</td>
<td>19.9</td>
<td>0.094</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.22, 0.30)</td>
<td>(0.65, 0.81)</td>
<td>(30.3, 37.3)</td>
<td>(17.2, 22.8)</td>
<td>(0.048, 0.152)</td>
<td>(0.27, 0.39)</td>
</tr>
<tr>
<td>Common bias, surveys &amp; risk-groups $b = b$</td>
<td>976</td>
<td>0.67</td>
<td>2.69</td>
<td>32.7</td>
<td>18.9</td>
<td>0.098</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.49, 0.93)</td>
<td>(1.83, 4.04)</td>
<td>(29.2, 36.5)</td>
<td>(16.3, 21.7)</td>
<td>(0.048, 0.157)</td>
<td>(0.60, 1.16)</td>
</tr>
<tr>
<td>Risk-group specific bias $b = b_g$</td>
<td>978</td>
<td>0.68</td>
<td>1.41</td>
<td>33.0</td>
<td>19.7</td>
<td>0.091</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.49, 0.96)</td>
<td>(0.58, 3.19)</td>
<td>(29.3, 37.2)</td>
<td>(16.8, 22.7)</td>
<td>(0.046, 0.150)</td>
<td>(0.39, 0.97)</td>
</tr>
<tr>
<td>Survey specific bias $b = b^i$</td>
<td>981</td>
<td>0.70</td>
<td>2.79</td>
<td>32.6</td>
<td>18.8</td>
<td>0.098</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.50, 1.01)</td>
<td>(1.84, 4.27)</td>
<td>(29.2, 36.5)</td>
<td>(16.2, 21.6)</td>
<td>(0.049, 0.158)</td>
<td>(0.60, 1.21)</td>
</tr>
<tr>
<td>Survey &amp; risk-group specific bias $b = b^i_{g}$</td>
<td>986</td>
<td>0.69</td>
<td>1.45</td>
<td>33.2</td>
<td>19.6</td>
<td>0.091</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.49, 0.96)</td>
<td>(0.61, 3.31)</td>
<td>(29.4, 37.3)</td>
<td>(16.7, 22.7)</td>
<td>(0.046, 0.151)</td>
<td>(0.39, 0.99)</td>
</tr>
</tbody>
</table>
Considerations/open questions

Model assessment: conflict between data sources

\[ \bar{D} = \sum_{i}^{n} \bar{D}_i \]

assuming independence between the \( n \) data sources becomes the sum of the item specific deviance contributions

- can be usefully employed to identify conflict between data sources
## Considerations/open questions

### Conflict between data sources

<table>
<thead>
<tr>
<th>Data source left out</th>
<th>$D$ for data source $i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>None (Model B3)</td>
<td>2.6</td>
</tr>
<tr>
<td>1 (C-R)</td>
<td>-</td>
</tr>
<tr>
<td>2 (BCS)</td>
<td>2.1</td>
</tr>
<tr>
<td>3 (CJS)</td>
<td>3.7</td>
</tr>
<tr>
<td>4 (SPM)</td>
<td>2.7</td>
</tr>
<tr>
<td>5 (NATSAL)</td>
<td>2.0</td>
</tr>
<tr>
<td>6 (UA IDUs)</td>
<td>2.5</td>
</tr>
<tr>
<td>7 (UA GUM)</td>
<td>2.4</td>
</tr>
<tr>
<td>8 (UA Antenatal)</td>
<td>2.6</td>
</tr>
<tr>
<td>9 (Neonatal sample)</td>
<td>2.5</td>
</tr>
<tr>
<td>10 (Blood donors)</td>
<td>2.5</td>
</tr>
<tr>
<td>11 (Tested population)</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Model assessment: main drivers

Data source left out

HCV antibody prevalence

Daniela De Angelis (HPA, MRC-BSU)
Co-authors

- Statisticians
  - Michael Sweeting, MRC-BSU
  - Tony Ades, Bristol University

- Epidemiologists
  - Matthew Hickman, Bristol University
  - Vivian Hope, HPA and LSHTM,
  - Mary Ramsay, HPA
For more details

M. Sweeting, D. DeAngelis, M. Hickman, A.E. Ades. 

D. De Angelis, M. Sweeting, A.E. Ades, V. Hope, M. Ramsay. 
An evidence synthesis approach to estimating Hepatitis C Prevalence in England and Wales. 
*Statistical Methods in Medical Research*, 

M. Sweeting, D. DeAngelis, M. Hickman, A.E. Ades. 
Estimating the prevalence of ex-injecting drug use in the population. 
*Statistical Methods in Medical Research*, 
For background reading

D.M. Eddy, V. Hasselblad, R. Shachter.
Meta-Analysis by the Confidence Profile Method.

Multiparameter evidence synthesis in epidemiology and medical
decision-making: current approaches.

D.J. Spiegelhalter, N.J. Best, B.P. Carlin, A. van der Linde.
Bayesian measure of model complexity and fit.
For other examples

N.J. Welton, A.E. Ades.
A model of toxoplasmosis incidence in the UK: evidence synthesis and consistency of evidence.

Estimates of HIV prevalence and proportion diagnosed based on Bayesian multi-parameter synthesis of surveillance data.

A. Presanis, D. De Angelis, D.J. Spiegelhalter, S. Seaman, A. Goubar, A.E. Ades.
Conflicting evidence in a Bayesian synthesis of surveillance data to estimate HIV prevalence