Assessing the coverage of TB notification using capture recapture methods

Ibrahim Abubakar
Outline

• Introduction
• Basic Method
• Assumptions
• Bernoulli Census
• Models
  – Log Linear
  – Truncated
• Advantages and Implementation Issues
• Application to UK TB epidemiology
  – Estimating Incidence
  – Estimating Mortality
Introduction

• First used - Laplace to estimate the population of France in 1786
• Petersen 1894 (tagged fish)
• Widely used in monitoring the abundance of wildlife and estimating human population
• Sekar and Deming 1949 Human Population
• Wittes 1968 Epidemiology
• WHO DIAMOND project (diabetes care) 1990
Introduction

Notifications

Hospital

Primary care

Deaths

Combine sources and remove duplicates

Notifications

Hospital

Primary care

Deaths
Population (N) =?

Sample (M) = 15

Population (N) =?

Population (N) with marked sample (15)

New sample (n) = 10
Marked (m) = 5

...and if I get one a second time, I can use that to estimate how many of them there are.
Capture recapture

N = (n x M) / m : Petersen estimator
N = [(M+1)(n+1)/(m+1)] - 1 : Corrected

N* = [(15+1)(10+1)/(5+1)] – 1 = 28

*Var(N) = ((M+1)(n+1)(M-m)(n-m))/((m+1)^2(m+2))
~ when samples are large
# nearly unbiased estimator (Chapman)

Duplicate elimination

N = (M+n) – m = 20

<table>
<thead>
<tr>
<th>Source A</th>
<th>Source B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
</tr>
<tr>
<td>Observers</td>
<td>m</td>
</tr>
<tr>
<td>Not observed</td>
<td>m-n</td>
</tr>
<tr>
<td>Total Source B</td>
<td>n</td>
</tr>
</tbody>
</table>
Assumptions

• Population closed to addition or deletion
• Individual identified are not lost (good matching)
• Independence
  – Population should be homogenous i.e. for a given source each case is equally likely to be listed by that source and…..
  – concept of trap fascination or avoidance – visibility or invisibility to the health care system
  – Two samples are independent registers
  – …..if direction of bias known, it may be used for boundary limits

Correct case definition (High specificity and PPV)
Small numbers
Tackling dependence

- Stratification can help with lack of homogeneity
- Modelling:
  - Bernoulli census
  - Log-linear models
  - Truncated models
Bernoulli census

- Requires multiple data sources (at least 3)
- Pairwise comparison
- If relatively similar then no dependence
- Subsequent merging
- If different, however......

- Not a very efficient method

Wittes et al 1974
Log Linear models

- Requires at least three data sources
- fit model to 2*k contingency table and predict number in missing (unascertained cell). K=data sources~
  ~ where the dependence between all k sources is assumed to be zero
Log Linear models

• For example – 3 list
  – all possible combinations of capture are arranged as an incomplete contingency table
  – Data assumed to follow a distribution
  – Fit log linear model
  – Assume that the fitted model also applies to the missing cell of the complete contingency table
Log Linear models

- For a contingency table
  - Three sources a, b, c with i, j, k levels respectively

\[
\ln u_{ijk} = u + u_a^i + u_b^j + u_c^k + u_{ab}^{ij} + u_{ac}^{ik} + u_{bc}^{jk} + u_{abc}^{ijk}
\]

- Where Log F is natural log of the expected frequency for cell ijk

- Model with eight terms
  - Number expected in all list (u)
  - Three “main effects” – log odds against appearing on each list \( u_a, u_b, u_c \)
  - Three two factor interactions \( u_{ab}, u_{ac}, u_{bc} \)
  - One three factor interaction \( u_{abc} \)

- Assumptions
  - that \( u_{abc} \) is zero
Model selection

• As with all model selection
  – We will never know if the final model is “correct” but can exclude wrong ones
  – Using likelihood ratio tests
  – Akaike's Information Criterion (AIC)
    • AIC = -2 × [log(L) - q]
    • Log(L) is the log-likelihood evaluated at the maximum likelihood estimates of the model parameters, and q is the number of parameters in the model
  – Bayes information criterion (BIC)
    • BIC = -2 × log(L) + q × log(n)
    • Where n is the sample size
Truncated models

- Based on assumptions about distribution of data e.g.
  - Poisson
  - Binomial
  - Mixture of distributions
  - Zeltermans truncated Poisson mixture models
  - Chao’s heterogeneity and bias corrected models

Wilson RM Biometrika 1992
Suggested advantages

- Allow monitoring - ensuring that changes are not a result of “an increase in the efficiency of registration”
  - NOTE: CI do not correct for wrong assumptions
- Enhances comparison between geographical areas
- Cheaper than repeating population surveys annually
But.....

- Can be resource intensive (matching, stats etc)
- Requires at least three routine data source (if dependencies expected)
Estimating TB under-notification

• Objective
  – to assess the use of capture–recapture methodology for estimating the completeness of TB reporting in England between 1999 and 2002

• Staff and expertise
  – sourcing data (costs, confidentiality, bureaucracy)
  – record linkage (suitable identifiers)
  – database management
  – Analysis and interpretation

• Methods:
  – Sources of data
    • Enhanced Tuberculosis Surveillance
    • Laboratory Database (MycobNet)
    • Hospital Discharge (HES)
  – Matching (deterministic)

van Hest et al 2008
Estimating TB under-notification

• Methods:
  – Model – Log Linear Models
  – Comparison with truncated models: Poisson Mixture Models, Structural Source model

• Results
  – overall estimated completeness of case ascertainment was 66.7% (28 678/42 969)
  – Overall observed under-notification was 15.9%.

van Hest et al 2008
Estimating TB under notification

• Limitations
  – Record Linkage – Matching works?
  – Dependence
  – Data sources and case definitions
  – Under-notification

• Implications for national surveillance
  – Audit of laboratory notifications
  – Automated reminder to local units

van Hest et al 2008
Estimating TB mortality

- **Objective:** to obtain a better estimate of mortality among TB cases
- **Data sources:**
  - Death Certificates
  - National Health Service Register
  - Surveillance Register (Treatment Outcome)
- **Data linkage:** Deterministic matching
- **Analysis:** A three-list saturated log-linear model was used to undertake capture-recapture

Crofts et al 2008
Estimating TB mortality

- Treatment Outcome: 853, 77%
- NHS Register: 888, 80%
- Death Certificates: 527, 48%

1169 (1108+61), (95% CI 1140-1224)

Crofts et al 2008
Estimating TB mortality

- Assumptions
  - Matching
  - Homogeneity

- Implications for national surveillance
  - Improving TOM
  - Reusing the methods for continuous monitoring of mortality trends
### Other published CRC TB studies

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sources</th>
<th>Aim/place</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferrer E et al</td>
<td>Notification and lab registers (2)</td>
<td>Incidence in Valencia, Spain</td>
</tr>
<tr>
<td>Ivanez G et al</td>
<td>Notification and lab registers (2)</td>
<td>Incidence in Seville, Spain</td>
</tr>
<tr>
<td>Sanghavi et al</td>
<td>Interviews and lab smear results (2)</td>
<td>Incidence in Peru</td>
</tr>
<tr>
<td>Perez C et al</td>
<td>Notification and lab register (2)</td>
<td>Incidence in Huesca, Spain</td>
</tr>
<tr>
<td>Tocque et al</td>
<td>Notification, lab records and hospital admissions (3)</td>
<td>Incidence in Liverpool, UK</td>
</tr>
<tr>
<td>Iglesias et al</td>
<td>Notification and lab records (2)</td>
<td>Incidence in Zaragoza, Spain</td>
</tr>
<tr>
<td>Mayoral et al</td>
<td>Notification and hospital discharge (2)</td>
<td>Incidence in Seville, Spain</td>
</tr>
</tbody>
</table>

van Hest et al 2008
### Other published CRC TB studies

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sources (number)</th>
<th>Aim/place</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tejero et al</td>
<td>Notification and hospital admission (2)</td>
<td>Incidence in Valladolid, Spain</td>
</tr>
<tr>
<td>Inigo et al</td>
<td>Epi contact and DNA fingerprinting data (2)</td>
<td>Recent transmission in Madrid, Spain</td>
</tr>
<tr>
<td>Cailhol et al</td>
<td>Notifications and drug resistance survey (2)</td>
<td>Meningitis incidence in France</td>
</tr>
<tr>
<td>Guernier et al</td>
<td>Lab, hospital and TB control databases (3)</td>
<td>Incidence in French Guyana</td>
</tr>
<tr>
<td>Baussano et al</td>
<td>Notification, lab register and hospital records (3)</td>
<td>Incidence in Italy</td>
</tr>
<tr>
<td>Van Hest et al</td>
<td>Notification, lab register and hospital records (3)</td>
<td>Incidence in England and Wales</td>
</tr>
<tr>
<td>Van Hest et al</td>
<td>Notification, lab register and hospital records (3)</td>
<td>Incidence in the Netherlands</td>
</tr>
</tbody>
</table>

van Hest et al 2008
Summary

- Method can estimate under-notification but note assumptions
- It is not a cheap (quick and dirty) method
- Statistical, data management and matching as well as epidemiological expertise required
- We are not fishes!
Acknowledgements

- Andrew Grant
- Jonathan Crofts
- Rob van Hest
<table>
<thead>
<tr>
<th>List A</th>
<th>List B</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Not Observed</td>
<td>Total List</td>
</tr>
<tr>
<td></td>
<td>n11</td>
<td>n01</td>
<td>na</td>
</tr>
<tr>
<td></td>
<td>n10</td>
<td>n00</td>
<td></td>
</tr>
<tr>
<td>Total List B</td>
<td>nb</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>

- Assuming being on list is random then
  - \( E[n_a] = N*p_a \)
  - \( E[n_b] = N*p_b \)
  - \( E[n_{ab}] = N*p_{ab} \)
  
  \((E=\text{expected random variable}, \ p=\text{probability})\)

- Assuming proportion in list A is roughly same as list B then, \( n_a/N = n_{11}/n_b \)
  - \( Ne = (n_a*n_b)/n_{11} \) or
  - \( Ne = r+n_{00} \) where \( n_{00} = (n_{01}*n_{10})/n_{11} \)

\[
\begin{align*}
  r &= n_{11} + n_{10} + n_{01} \\
  n_b &= n_{11} + n_{01} \text{ and } n_a = n_{11} + n_{10}
\end{align*}
\]
Dependence

- For large populations* (N):
  - \( E[Ne] = E[n_a] * E[n_b] / E[n_{11}] \) or
  - \( E[Ne] = N * p_a * p_b / p_{ab} \) or
  - \( E[Ne] = N * R \). Where \( R = (p_a * p_b / p_{ab}) \) or \( = p_b / p_{b|a} \)

- \( R=1 \) if list are independent and greater or less with dependence
- e.g. if being on list A increases probability of being on B, then \( p_{b|a} > p_b \) and \( R<1 \)
- Therefore \( Ne \) will underestimate \( N \)
Log Linear models

<table>
<thead>
<tr>
<th>List A</th>
<th>List B</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Not Observed</td>
<td>Total List A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>n11</td>
<td>n01</td>
<td>na</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not observed</td>
<td>n10</td>
<td>n00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total List B</td>
<td>nb</td>
<td></td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using the same notation in a two list example:

- Parameterizing a 2x2 contingency table:
  - $\log E[n_{11}] = u$
  - $\log E[n_{01}] = u + u_a$
  - $\log E[n_{10}] = u + u_b$
  - $\log E[n_{00}] = u + u_a + u_b + u_{ab}$

- Where $u_a$ and $u_b$ represent marginal log-odds of the probabilities of appearing in various lists