Inventory and Capture Recapture Studies

Ana Bierrenbach
WHO consultant
Nov/Dec 2010

Based on lecture elaborated by Dr. Udo Buchholz
Introduction

• No surveillance system captures ALL cases … but one can try to estimate the real numbers

• Possible sources of data to cross-check for presence of TB cases:
  – HIV/AIDS notification data
  – Hospital discharge or admission data
  – Data from vital registration
  – Health insurance data
  – Data from primary care physicians registers within a national health service
  – Drug prescriptions data
  – Data from microscopy services (for AFB+ cases)
Two data sources

• Assume:
  – Health insurance: 100 cases
  – National TB program: 200 cases

• True number
  – 200?
  – 300? (100 + 200)
  – > 300?

• Two ways to tackle the problem:
  – Minimum number: inventory method
  – Estimated number: capture - recapture
Inventory method
step 1: construct the lists

**Lista 1: Health insurance**
- Mary Kruger, 23 years, female
- Joseph Formell, 23 years, male
- Ernest Kalahari, 17 years, male
- Mario Maradona, 33 years, male
- Victor Limpopo 49 years, male
- Shaka Shakira, 28 years, female

**List 2: National TB program**
- Mary Kruger, 23 years, female
- Feng Shui, 2220 years, female
- Michel Ballack, 53 years, male
- Mario Maradona, 33 years, male
- Laure Grajaes, 98 years, female
- Hector Sanchez, 44 years, male
- Paolo Almodovar, 45, years, male

**Lista 3: Complete list**
- Mary Kruger, 23 years, female
- Joseph Formell, 23 years, male
- Feng Shui, 2220 years, female
- Michel Ballack, 53 years, male
- Ernest Kalahari, 17 years, male
- Mario Maradona, 33 years, male
- Laure Grajaes, 98 years, female
- Hector Sanchez, 44 years, male
- Shaka Shakira, 28 years, female
- Paolo Almodovar, 45, years, male

List 1: 6
List 2: 7
List 3: 11

Common list (between 1 & 2): 2
Inventory method
step 2: count the cases

Simply count the cases:

Health insurance
(n=100)

National TB program (n=200)

80 + 20 + 180 = 280
Calculate multiplier

Proportion of missing cases, multiplier:

Proportion of missing cases:
\[ \frac{200}{280} = 71\% \]

Multiplyer:
\[ \frac{280}{200} = 1.4 \]

- May be calculated separately for different age groups
Principle of capture - recapture

How many birds/mosquitoes... are there in the central park?
- Difficult to count!

Capture

Mark the birds

Time elapses

Recapture
Estimation

All of the birds

Marked birds

Recaptured birds
### 2 x 2 table

<table>
<thead>
<tr>
<th></th>
<th>Birds recaptured</th>
<th>- not recaptured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marked birds</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>-not marked</td>
<td>180</td>
<td>??</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>??</td>
</tr>
</tbody>
</table>

All of the birds
2 sources: formula

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$N_{AB}$</td>
<td>80</td>
</tr>
<tr>
<td>-</td>
<td>180</td>
<td>??</td>
</tr>
<tr>
<td>$N_B$</td>
<td>??</td>
<td>$N$</td>
</tr>
</tbody>
</table>

- Only if the two sources are independent: $P(A \text{ and } B) = P(A) \times P(B)$, so that:

$$\frac{N_{AB}}{N} = \frac{N_A}{N} \times \frac{N_B}{N} \quad \Rightarrow \quad N = \frac{N_A \times N_B}{N_{AB}} \quad \Rightarrow \quad N = \frac{(N_A + 1) \times (N_B + 1)}{(N_{AB} + 1)} - 1$$

More exact
Estimation for two sources

<table>
<thead>
<tr>
<th></th>
<th>Birds recaptured</th>
<th>- not recaptured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birds marked</td>
<td>20</td>
<td>80</td>
</tr>
<tr>
<td>- not marked</td>
<td>180</td>
<td>??</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>??</td>
</tr>
</tbody>
</table>

\[ N = \frac{(N_A + 1) \times (N_B + 1)}{(N_{AB} + 1)} - 1 \]

\[ = \left(\frac{(100+1) \times (200+1)}{(20+1)}\right) - 1 = 966 \]

Remember: the inventory method yielded 280 cases
Assumptions 1

- **Closed population**: no birth, death or migration
- **Perfect linkage**: no matching error
- **No misdiagnosis**: correct case definition
- **Homogeneity capture**: for a given source, every case has the same chance of being captured, the same "catchability"
  - For example: male and female cases, rich and poor cases should have the same chance of being captured
Assumptions 2

• **Source independence:**
  I.e. cases captured in system 1 should have the same probability to show up in system 2 as those not captured in system 1

  – **Negatively dependent:**
    • Cases with a health insurance have less chance of being notified to the NTP, but cases without health insurance (e.g. treated in governmental services for the poor) have a greater chance
    • Mutually exclusive data sources based on geographical criteria

  – **Positively dependent:**
    • You are more likely to get on to NTP register if you are on a public hospital
Overestimation, underestimation

- **Over-estimation:**
  - When cases captured in both sources are not recognized as being the same, or **negative** dependence

- **Under-estimation:**
  - When there is **positive** dependence between the two sources
Example for overestimation: the difference in the estimates can be great if matching cases are missed

![Table showing 20 and 80 matches example](image)

\[
\text{20 matches: } \frac{(100+1) \times (200+1)}{(20+1)} - 1 = 966
\]

\[
\text{80 matches: } \frac{(100+1) \times (200+1)}{(80+1)} - 1 = 250
\]

What a difference!!
The inventory method would have yielded: **280 cases**
Tackling dependence (1)

Stratification helps but does not solve the problem

- It seems that the blue source systematically "loses" children older than 5 years
- Regarding gender, there seems to be no large difference
Tackling dependence (2)

Modelling: several models

- When there are only two sources one cannot check this assumption in a proper way (other than qualitatively)

- Requirement of at least 3 data source to be able to check for dependencies and to mathematically account for them while computing results

- Can be resource intensive (matching, stats, etc)
Capture Recapture Recommendations

- Should never take precedence over routine surveillance strengthening
- Best when adequately resourced
- Should have at least three data sources
- May provide useful information on data sources and how they relate to each other
- Provides a broad estimate of under-notification that should always be interpreted with local knowledge
Thank you