Capture-recapture studies

Laura Anderson

Centre for Infections
Health Protection Agency
UK
Reiterating underlying assumptions

1) No misclassification of records
   (perfect record linkage)

2) Closed population
   (no immigration or emigration in time period studied)

3) Homogenous population
   (same chance of being observed and re-observed)

4) Independent registers
   (probability of being on one is not affected by being/not being on the other).
Efforts can be made to reduce violation of assumptions

- Complete data
- Good personal identifiers
- Short time period to collect data
- Stratification
- Covariates in model.

- Adding an interaction

- Allows linkage of patients from different registers
- Ensures there is little change in the population
- Dealing with heterogeneity
- Identifying and controlling interdependency
Describe two capture-recapture studies:

1) Egypt study – middle income setting
2) England study – resource rich setting

• Why were these studies successful?
• What were the limitations?
• Can they be considered representative of the population of the country as a whole?
Objectives:
To estimate case detection rates in Egypt in 2007 using record-linkage and capture-recapture and

Explore this method as a tool for periodic evaluation of the WHO tuberculosis control strategy in a resource-limited setting.
Methodology

• 4 randomly selected governorates: Cairo, Dakahlia, Fayum and Matrouh = 11.1% of national population.

• Longitudinal surveillance Oct-Dec 2007 within public and private non-NTP sector plus data from NTP.

• Case definition – patients registered with the NTP (new and retreatment cases) or confirmed on NTP criteria for non-NTP providers.

• Three source log-linear capture-recapture on all and smear positive cases.
Data sources and linkage

1) NTP register
2) Public non-NTP
3) Private non-NTP

- NTP identical register introduced to non-NTP providers (demographic and clinical data) plus identical NTP lab register.

- Record linkage by name

- Misclassification was corrected by examining NTP register Jan 2007-April 2007.
Results

Total number of tuberculosis cases = 410
Capture-recapture analysis

• Log-linear model included main effects plus the following expected interactions:
  positive NTP*Public non-NTP
  negative Public non-NTP*Private non-NTP

• CDR of NTP surveillance = 55% (95% ci 46% - 68%)

• Completeness of case ascertainment = 62% (95% CI 52%-77%).

• Sputum smear-positive was higher at 66% (55%-75%) and 72% (60%-82%), respectively.
Strengths and limitations

• Case definition identical for all registers.

• Perfect record-linkage – hierarchic combination of three Arabic names used to minimise overmatching (routine in Egypt).

• Improved completeness through weekly visits (minimise undermatching) and follow up of NTP register.

• Data collected in a short period of time (but small numbers which may limit the overlap of data collection).

• Two source interdependency controlled by log linear model using interaction terms.

• Three interdependency assumed to be negligible.
• Exclusion of false positive cases outside the study period or those not having TB.

• Longitudinal surveillance = more expensive and longer than expected but

  1. was required in a resource poor setting to obtain a more accurate estimate.

  2. allowed control of possible interdependencies which cannot be carried out in two-source studies).

• Results available only months after a short surveillance period (seasonal variation?)

• Cluster random sampling = good representation of the whole population.
England study
(van Hest et al., 2008)

Objectives:

Estimate annual incidence of TB in England

Assess the completeness of reporting between 1999-2002
Data Sources

HOSPITAL EPISODE STATISTICS
cases admitted to NHS hospitals with 1st or 2nd hospital discharge code as TB (International Classification of Diseases-10)

ENHANCED TUBERCULOSIS SURVEILLANCE
Notification system which collects additional information on demographic and clinical characteristics. Includes both culture confirmed and clinically diagnosed cases of TB.

MycobNet
Culture confirmed cases of MTB complex from reference laboratories
Time period

Interval of more than 1 year between entries in data sources were considered to be separate episodes.

Record linkage

Duplicates removed.

ETS and MycobNet routinely linked then match HES.

Matched on core identifiers: age, sex, d.o.b., postcode, name using soundex (not available for HES).

Sophisticated software using a points system (human element required)
False positive cases

Removed from ETS using Treatment Outcome Monitoring data.

Matching of HES to mycobacteria other than tuberculosis (MOTT) database.

Excess of unlinked HES cases removed using the Bernoulli parameter (probability of being a true TB cases based on the covariates: days of admission, number of admissions in TB episode, ICD code, rank number of TB diagnosis.)
**Interdependency**

3 source log linear model allows interactions to be examined first and then added to the model as an interaction term.

This can be used to correct for interdependency between 2 sources but not if there is dependency between 3 sources.
## Results

<table>
<thead>
<tr>
<th>Source</th>
<th>% of cases captured</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETS</td>
<td>84.1</td>
</tr>
<tr>
<td>MycobNet</td>
<td>54.3</td>
</tr>
<tr>
<td>HES</td>
<td>41.6</td>
</tr>
</tbody>
</table>

15.9% undernotification to ETS.

This improved from 18.8% in 1999 to 13.3% in 2002.
• 28678 total TB cases observed.
• 14291 unobserved

• $\frac{28678}{42969} = 66.7\%$ complete cases ascertainment.

• Overall undernotification = 43.8%
Strengths

• Good completeness of data from cross validation.

• Excellent record linkage – removal of duplicates and false positive cases. 94.9% had an association with a high score suggesting true links.

• Short time period.

• Interdependency likely controlled for using log linear model.
Strengths

- Good database management.
- Access to data.
- Expertise in computing, epidemiology and statistics – important to work as a team.
- Data from whole country so representative. (However, TB varies by region of the UK and therefore perhaps not representative of Scotland, Wales or Northern Ireland).
Limitations

• Case definition – varies for different data sources and therefore there are variations in specificity. Still trying to capture all cases in ETS and MycobNet.

• Cases excluded due to no d.o.b therefore could be more complete.

• Homogenous population likely violated (age, site of disease). Couldn’t stratify due to incompleteness of identifiers from all sources.
Conclusion

• ALWAYS place a capture-recapture study in the context of limitations of the study.

• Important to estimate undetected burden in the country

HOWEVER

• May be a more useful to use as a tool to evaluate a surveillance system- e.g England study
Usually all underlying assumptions are violated and therefore the study is only as good as the data sources.
Thank you!