Overview of statistical issues in analysis of TB prevalence surveys

Sian Floyd
Babis Sismanidis

WHO Task Force on TB Impact Measurement
Meeting of the sub-group on TB prevalence surveys
28-30 January, WHO Geneva
Overview

• Current guidance in “red book”

• 2 key issues considered further here:
  A. Analysis at individual level – allowing for correlation among individuals in the same cluster
  B. Missing data
Red book guidance
A) Cluster-level analysis

- Described in detail, using example from Cambodia
- Method used:
  1. Calculate point prevalence in each cluster
  2. Calculate the overall point prevalence
     (with each cluster weighted according to its size)
  3. Estimate design effect
     (by comparing observed between-cluster variation to the variation predicted if individual observations are statistically independent)
  4. Estimate 95% confidence interval for overall TB point prevalence estimate, corrected for the design effect
A) Cluster-level analysis (cont)

• Points out that a complete analysis would use individual-level data

• i.e. do the entire analysis with each individual as the unit of analysis, rather than each cluster as the unit of analysis

• Says this can be done in Stata using svy (survey) commands

• No detailed guidance on implementation
B) Missing data

- If proportion of individuals that are enumerated but do not participate in the survey is not low (e.g. >15%), must consider implications

- If those that do not take part are at higher or lower risk of TB than those that do, this will introduce bias

- Suggests investigating if participants have a different age-sex distribution compared with non-participants

- No detailed guidance on how to handle missing data, in the situation that participants and non-participants differ according to characteristics known to be associated with TB (e.g. age and sex)
Individual-level analysis, allowing for correlation among individuals in same cluster
Why is there within-cluster correlation in TB prevalence surveys?

• Within-cluster correlation = between-cluster variation

• Individuals within the same cluster are likely to be more similar to each other than they are to individuals in other clusters:
  – infectious disease tends to cluster in time and space
  – shared risk factors for infection, and for progression from infection to disease (genetic, environmental, socio-economic)
  – TB cases vary in their infectiousness (may be more infectious, and/or more or closer contact with non-cases, in some clusters than others)
Implication for statistical analysis

• Each individual provides less information than they would if their response was independent of other individuals in the same cluster.

• 2 extremes:

  • If correlation = 0, then responses of individuals within the same cluster are no more alike than those of individuals from different clusters. Individual observations are statistically independent – and can use standard methods of analysis.

  • If correlation = 1, then all responses in the cluster are the same. Effectively have only one observation per cluster, even if there are several hundred individuals in the cluster.

• Reality: somewhere in-between.
Implication for statistical analysis (cont)

• If we ignore correlation among individuals in the same cluster / between-cluster variation, then the main problem is:
  
  – estimated standard errors are too small

  – so confidence intervals are too narrow

  – and p-values are too small
  (if we make comparisons between groups)
4 possible methods to account for within-cluster correlation

• There is one method in which the analysis is performed at the level of the cluster – i.e the cluster becomes the unit of analysis (explained in “red book”)

• Analysis with the cluster as the unit of analysis is necessary if the number of clusters is <20, and may be the best option also if there are 20-29 clusters

• There are 3 methods in which the analysis is performed at the level of the individual – 2 key advantages are a) enables imputation of missing data b) risk-factor analysis (of individual-level characteristics)

• The simplest of the 3 individual-level methods is expected to work well if the number of clusters is 30+, and the more sophisticated methods should be tried if there are 30+ clusters

• All the TB prevalence survey protocols have 30+ clusters
3 methods to account for within-cluster correlation, individual-level analysis

(1) “Robust” standard errors

(2) Generalised estimating equations

(3) Random-effects (multi-level) model

= hierarchical logistic model in the case of the analysis of a cluster sample survey of TB prevalence

• Will concentrate on (1) and (3)

• Method (2) is a “half-way house” and is neither simple (like Method 1) nor does it provide a full probability model for the data (like Method 3)
Method (1) – “robust” standard errors

• Ordinary logistic regression model

• But with “robust” standard errors that are calculated based on the observed variability in TB prevalence among clusters

So:

• gives the same estimate of TB prevalence as a logistic model assuming statistical independence of individuals

• But standard errors, CIs and p-values are corrected for clustering
Method (1) – “robust” standard errors (cont)

Advantages

• Straightforward and intuitive - uses standard logistic regression method to obtain point estimate of TB prevalence, giving equal weight to each individual in the sample

• Always works – i.e. can always fit this model to the data

Disadvantages

• Does not take account of within-cluster correlation in the point estimate of TB prevalence

• However, this is relatively less important when the size of each cluster is similar (as in the design of TB prevalence surveys)
Sample weights, or self-weighting sample?

- No need for “sample weights”, if the sample size in each cluster is the same or similar, and the clusters are selected with probability proportional to size

- Recommended that sample size in each cluster is the same or at least similar because this makes the analysis both more straightforward and more intuitive
Method (3) – “random-effects” model

• Full probability model for the data – i.e. between-cluster variation is modelled explicitly (not “fixed” as in Method 1)

• In an ordinary logistic regression model, the probability model for individual j in cluster i ignores the clustering:

\[
\text{Log(odds of TB disease)}_{ij} = \text{baseline log odds}
\]

where baseline log odds = overall log(odds of TB disease) in the population

• In a random-effects model, the probability model for individual j in cluster i allows explicitly for the clustering:

\[
\text{Log(odds of TB disease)}_{ij} = (\text{baseline log odds}) + u_i
\]

• And each cluster has its own value of \( u_i \)

• So the log(odds) in each cluster is shifted from the population mean by an amount \( u_i \)
Method (3) – “random-effects” model (cont)

• This makes individuals in the same cluster similar to each other to some extent (within-cluster correlation) and different to some extent to individuals in other clusters (between-cluster variation)

• For logistic regression, the $u_i$ are usually assumed to follow a normal distribution with mean 0 and variance $\sigma^2$ (to be estimated)

• The bigger is the between-cluster variation, the bigger is the variance of the $u_i$

• Because it is assumed that the $u_i$ follow a probability distribution (in this case a normal distribution with mean 0), they are termed “random” effects
Method (3) – “random-effects” model (cont)

Disadvantages

• May not be able to fit the model satisfactorily

• The equations that need to be solved to fit this model are very complicated. Numerical approximations are required to solve them, and sometimes they do not work well

• Can check the reliability of the approximations – this is important when the clusters are quite large (as they are in TB prevalence surveys)

• If approximations are not reliable, then should not accept the result of the model, and instead use simple, pragmatic method i.e. Method 1 above (robust standard errors)
Method (3) – “random-effects” model (cont)

Advantages

• Clustering is taken into account in the point estimate of TB prevalence, as well as in the standard errors, CIs, and p-values

• We also obtain an estimate of the between-cluster variation, and within-cluster correlation

• Method can now be implemented in combination with multiple imputation for missing values, in Stata (since early 2008)

So:

This is the method of choice
- if we are able to fit the model
Individual-level analysis for correlated data - Key messages

• **Method 1** – easy to understand, always works, expected to provide unbiased estimate of TB prevalence

• **Method 3** – mathematically / statistically “satisfying”, because it is based on a full probability model for the data including the within-cluster correlation – but it does not always work

• **So**: recommend starting with Method 1, and then trying Method 3

• **And** recommend reporting both – results are more convincing if they are similar for the 2 methods

*(since then we know the results are robust to different assumptions)*
Multiple levels of clustering 
(household within cluster)

- Sufficient in the analysis of TB prevalence surveys to allow for just the “primary” level of clustering – the variance among the primary sampling units incorporates lower levels of clustering.

- However, worth examining and describing the extent to which TB cases cluster in households, as well as within clusters.

- A simple way to summarise is:

- Among the TB cases, for what proportion were they:
  1) the only TB case in the household
  2) in a household with one other TB case
  3) in a household with 3+ TB cases
Missing data
There will ALWAYS be missing data…

**Missing outcome data**

- Individuals do not participate in the screening survey
- Individuals do not attend for chest X-ray screening
- Individuals with an X-ray suggestive of TB, and/or TB symptoms, do not provide sputum
- Individuals provide sputum but it is lost, or contaminated

**Missing data on “explanatory” variables**

- Should be possible to collect virtually complete data on age, sex, stratum, and cluster

How to handle missing data in the analysis depends on what we believe is the reason for the missingness
3 important types of missing data

1) Missing completely at random (MCAR)

2) Missing at random (MAR)

3) Missing not at random (MNAR)
Missing completely at random (MCAR)

The probability of an individual being missed in the survey is NOT related to either:

a) the outcome (TB case yes/no) or

b) an individual characteristic that is associated with the outcome (e.g. age, sex, stratum, cluster, TB symptoms)

• If this is true, we can restrict our analysis to the individuals who DO participate in the survey, and we will obtain an unbiased estimate of true TB prevalence

• However, this is unlikely to be true?
Missing at random (MAR)

- The probability of an individual being missed in the survey \textit{is} related to individual characteristics
- (e.g. age, sex, stratum, cluster, TB symptoms)
- If these individual characteristics are also associated with the outcome variable, then an analysis that is restricted to individuals who participate in the survey will \textit{not} provide an unbiased estimate of true TB prevalence in the population
Missing at random (MAR)

However:

• If, WITHIN strata defined by individual characteristics such as age, sex, cluster, TB symptoms, the probability of an individual having missing data for the outcome variable is NOT associated with the outcome value (TB case yes / no)

and:

• We use observed patterns of TB prevalence within each stratum (defined by all possible combinations of individual characteristics), to predict the outcome value for individuals for whom data are missing:

We can still obtain an unbiased estimate of true TB prevalence
Missing at random (MAR) - imputation

- Imputation is done using regression models

- We can in a single analysis impute missing values across several variables with missing data
  e.g.
  TB symptoms
  chest X-ray (positive or negative)

- Should do the imputations separately for each of the following key outcomes:

  1. Chest X-ray (positive or negative)
  2. Bacteriologically confirmed pulmonary TB (yes or no)
  3. Smear-positive pulmonary TB (yes or no)

- For 2. and 3., would include chest X-ray result as a predictor
Missing at random (MAR) - imputation

The idea is as follows:

- Assign “starting values” to the missing data – for each variable, these starting values are a random sample from individuals WITH data

- For each variable with missing data (e.g. Chest X-ray, TB symptoms), fit a model with this particular variable as the outcome variable

  e.g. fit a logistic regression model with chest X-ray (positive or negative) as the outcome variable, and with age, sex, stratum, TB symptoms as predictors

- Use the fitted model to obtain, for each individual, a predicted probability that the chest X-ray outcome is positive

- From these predicted probabilities, impute a value of 0 (negative) or 1 (positive) for each individual with missing data for the chest X-ray result
Missing at random (MAR) - imputation

The idea is as follows:

• Next use the observed + imputed data on chest X-ray, combined with other predictors such as age, sex, and stratum, to predict a second variable with missing data – e.g. cough in the last 3 weeks yes or no

• Complete this process so that an imputed dataset is created in which all variables with missing data have been “filled in” through imputation based on a regression model

• If there is >1 variable with missing data to be imputed (e.g. chest X-ray and TB symptoms), cycle through this process several times in order to obtain one “reliable” imputed dataset (meaning one imputed dataset in which the imputed data are not dependent on the starting values – as the starting values are selected at random and are not optimal)

• 10-20 cycles usually sufficient, but should check that using more makes little difference to the results
Missing at random (MAR) - imputation

• Then repeat the process several times – i.e. create several imputed datasets

• Recommended to use at least 5 imputed datasets – 5 is often sufficient, 10 usually sufficient

• Can create the imputed datasets in Stata using the ice command (imputation using chained equations)

• With several imputed datasets, we can estimate the additional uncertainty introduced into the point estimate of TB prevalence by the missing data
Missing at random (MAR) - imputation

- Then combine the results from the 5 (or 10) imputed datasets, to obtain an overall estimate of TB prevalence, including a confidence interval that is corrected for both clustering in the survey design and the additional uncertainty introduced by missing value imputation.

- Can do this in Stata using the `micombine` command (or, since early 2008, using `mim`).

- Important to check that number of imputed datasets is sufficient (want the results from e.g. 5 imputed datasets to be similar to those from 10; if not then try 20 imputed datasets, and compare 10 to 20 etc.).

- Stop when increasing the number of imputed datasets makes little difference to the point estimate and confidence interval – i.e. if result from 5 imputed datasets similar to the result from 10, use 5.

- We should compare the point estimate of TB prevalence with and without imputation of missing values, to assess if an analysis ignoring the missing data introduces bias.
Missing not at random (MNAR)

- Even when we stratify our data on individual characteristics that are associated with participation in the survey, the probability of an individual having missing data on the outcome variable (TB yes/no) is different for individuals who have TB than for individuals who do not have TB.

- In this case, we cannot “correct” the estimate of TB prevalence simply by using missing value imputation based on the patterns in the observed data.

- A sensitivity analysis is appropriate for this situation.
### Example – prevalence of bacteriologically confirmed TB, Philippines TB prevalence survey 2007

**TABLE**

TB prevalence as estimated based on culture and/or smear positive results; individuals aged 10 or more years old

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>% (95% CI)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall point prevalence</td>
<td>0.66 (0.52-0.81)</td>
<td>0.66 (0.53-0.80)</td>
<td>0.60 (0.46-0.73), P=0.004</td>
</tr>
<tr>
<td><strong>Point prevalence by stratum</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metro Manila</td>
<td>0.67 (0.10-1.24)</td>
<td>0.64 (0.16-1.12)</td>
<td>0.58 (0.20-1.10)</td>
</tr>
<tr>
<td>Other urban</td>
<td>0.66 (0.47-0.86)</td>
<td>0.68 (0.50-0.86)</td>
<td>0.61 (0.39-0.83), P=0.004</td>
</tr>
<tr>
<td>Rural</td>
<td>0.66 (0.45-0.87)</td>
<td>0.65 (0.46-0.85)</td>
<td>0.59 (0.40-0.77)</td>
</tr>
<tr>
<td><strong>Overall crude prevalence</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>136/ 20,544</strong> (0.66, 0.56-0.78)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Model 1**: CI corrected for clustering  
**Model 2**: CI corrected for clustering, plus missing data imputed  
**Model 3**: Point prevalence and CI corrected for clustering
Missing data – Key messages

• Must describe the percentage of individuals with missing data, for each key variable
  e.g. symptom screening questionnaire
  chest X-ray
  smear and culture results available among individuals
  with a chest X-ray suggestive of TB

• Must describe (and tabulate) how the percentage of individuals with missing data on key outcome variables varies according to characteristics known to be associated with TB (e.g. age, sex, stratum, cluster)

• Unless data are “missing completely at random”, an analysis restricted to individuals with complete data may introduce bias

• Very important to keep missing data on outcome variables (X-ray, smear and culture results), and key explanatory variables (individual characteristics known to be associated with TB) to a minimum

• Repeat visits and tracing of missed individuals is very worthwhile even if we can try later to adjust for missing data in the analysis
Missing data – Key messages

• But there will always be some missing data

• To reduce bias, should impute data (based on patterns in the observed data)

• If data are “missing at random”, the percentage of individuals with missing data is not too high, appropriate imputation models are used, and the data from the imputed datasets are combined in an appropriate way, then we will obtain an unbiased estimate of TB prevalence

• Can calculate a confidence interval for TB prevalence that allows for the clustering in the survey design and the uncertainty introduced by the imputation
Missing data – key messages

• With a rare disease, such as TB, it can be important to use imputation even when the percentage of individuals with missing data is low.

• Multiple imputation using chained equations can be implemented in Stata (and R).

• In Stata, the key commands (ice, micombine, mim) are quite user-friendly and flexible.

• However, multiple imputation must be done with care – in particular, specification of appropriate imputation models is essential.
A postscript – Post-stratification

• Sample survey population may not match the known total population distribution
  E.g. proportion of survey population in each region may not match known regional population distribution

• May arise if census list from which clusters were sampled is no longer accurate

• May also occur by chance – as number of clusters selected is relatively small

• Can address by fitting a logistic regression model that includes the variable which categorises clusters according to the stratum they are in (e.g. urban vs rural)

• Estimate, from this model, the prevalence in each stratum

• Then weight the stratum-specific prevalences estimated by the regression model, according to the known population distribution, with corresponding 95% CI