TB surveillance and surveys: a training workshop for consultants

Analysis of TB prevalence and drug resistance surveys

Day 4 – Friday, 27 May 2011
Geneva

Babis Sismanidis
with acknowledgements to Sian Floyd

Part I

Accounting for clustering

Background

Prevalence = TB cases / Number of eligible participants
(95% CI of a proportion) - *NO*

- Only correct if participants randomly sampled individually from general population
- Cluster sampling used instead to simplify logistics and reduce cost
- Cluster sampling needs to be accounted for in analysis

Overview

(New) key issues covered:

A. Analysis at cluster level (classical approach)

B. Analysis at individual level – allowing for correlation among individuals in the same cluster

C. Accounting for missing data

A) Cluster level analysis

Method used to:

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
   (observed between-cluster variation to the variation predicted if individuals independent)
4. Calculate 95% confidence interval for overall TB point prevalence estimate, corrected for the design effect

BUT…it does not allow for missing value imputation and investigation for association with risk factors (e.g. age)

B) Individual level analysis

- 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)
- Between-cluster variation = within-cluster correlation (how similar individuals from the same cluster are)

Implication for statistical analysis

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

2 recommended methods to account for cluster sampling

(1) “Robust” standard errors

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

= "hierarchical" logistic model in the case of the analysis of a cluster sampled TB surveys of disease prevalence and drug resistance

Method (1) – "robust" standard errors I

- Ordinary logistic regression model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1$$

- but with "robust" standard errors that are calculated based on the observed variability in TB prevalence among clusters

So:

- gives the same estimate of interest 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 II

Advantages

- Straightforward and intuitive - uses standard logistic regression method to obtain point estimates, 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 clustering in the point estimate
- However, this is relatively less important when the size of each cluster is similar

Method (2) – "random-effects" model I

- In an ordinary logistic regression model, the estimation ignores the clustering
- In a random-effects model, the estimation allows explicitly for the clustering, by allowing a term u_i for each of the i clusters

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_{1i} + u_i$$

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

Method (2) – "random-effects" model II

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
- Produces "shrunken" estimates of the intercept α when between-cluster variability is large and cluster size variable. In that case use method (1)
- Approximations are sometimes not reliable. If so, results of the model should not be accepted, and instead use the simpler, more pragmatic method (1) robust standard errors

Method (2) – "random-effects" model III

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 correcting for missing values (STATA 10, since early 2008)

Individual-level analysis for correlated data; key messages

- Method 1 – easy to understand, always works, expected to provide unbiased estimate of TB prevalence
- Method 2 – 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*
- Recommend starting with Method 1, and then trying 2
- Report both – results are more convincing if they are similar for the different methods
- Individual-level preferred over cluster-level analysis because they allow for investigating association between TB prevalence and risk factors AND correcting for missing data (*more to follow*)

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 these would be to look among TB cases what proportion are 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

Part II

Accounting for missing data

C) (There is always) 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 sample is lost, or contaminated so no bacteriological result is available

Missing data on "explanatory" variables

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

Missing data could provide biased (under- or over-) point estimates and associations with explanatory variables. How to handle missing data in the analysis depends on what we believe is the reason for the missingness

3 types of missing data

1) Missing completely at random (MCAR): no adjustment required, ignore missing data

2) Missing at random (MAR): missing value imputation required. Compare the point estimates with and without imputation of missing values, to assess if an analysis ignoring the missing data introduces bias

3) Missing not at random (MNAR): sensitivity analysis required to study the extend of bias

None of these assumptions can be proven *based on observed data*

Missing completely at random (MCAR)

The probability of an individual having missing data is NOT related to either:

a) the outcome (TB or MDR 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 (complete case analysis), and we will obtain unbiased point estimate
- However, this is unlikely to be true? MCAR is the strongest assumption one could make about missingness

Missing at random (MAR)

- The probability of an individual having missing data IS related to individual characteristics

  (e.g. age, sex, stratum, urban/rural, 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 most possibly provide a biased estimate of true TB or MDR-TB prevalence

Missing at random (MAR)

However:

- If within strata defined by individual characteristics the probability of an individual having missing data is NOT associated with the outcome value (i.e. MCAR within strata)

- And we use observed patterns of data 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 (due to missingness) estimate of true TB prevalence

MAR – introduction to imputation process

- Imputation is done using regression models

- We can in a single analysis impute missing values across several variables with missing data ("imputation of chained equations")

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)
  4. MDR TB (yes or no)

MAR – imputation process I

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 fit a model with this particular variable as the outcome variable

e.g. fit a logistic regression model with MDR (positive or negative) as the outcome variable, and with age, sex, cluster as predictors

- Use the fitted model to obtain, for each individual, a predicted probability that the MDR outcome is positive
- From these predicted probabilities, impute a value of 0 (negative) or 1 (positive) for each individual with missing data for MDR

MAR – imputation process II

- Next use the observed + imputed data on MDR combined with other predictors, to predict a second variable with missing data
- 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 by using more cycles and seeing this makes little difference to the results

MAR – imputation process III

- 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

MAR – imputation process IV

- Then combine the results from the 5 (or 10) imputed datasets, to obtain an overall estimate, 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). A new series of mi commands exist in STATA 11
- Important to check that number of imputed datasets is sufficient
- Stop when increasing the number of imputed datasets makes little difference to the point estimate and confidence interval
- We should compare the point estimate of TB or MDR-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 is different for individuals who have TB (or MDR-TB) than for individuals who do not have TB (or MDR-TB)
- In this case, we cannot "correct" estimates of TB or MDR-TB prevalence simply by using missing value imputation based on the patterns in the observed data
- A sensitivity analysis is appropriate for this situation

Missing data; key messages I

- There will always be some missing data
- With a rare disease, such as TB or MDR-TB, it can be important to use imputation even when the percentage of individuals with missing data is low
- To address potential bias impute data (based on patterns in the observed data) if missing at random
- Can calculate a confidence interval for point estimates that allows for *BOTH* the clustering in the survey design *AND* the uncertainty introduced by the imputation

Missing data; key messages II

- Must describe the percentage of individuals with missing data, for each key variable

  e.g. symptom screening questionnaire, chest X-ray, smear, culture results, resistance testing

- 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 or MDR-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 (over- or under- estimate TB prevalence)

Missing data; key messages III

- 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

- 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. NEW mi commands available

- 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: (i) if census list from which clusters were sampled is no longer accurate, or (ii) 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

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
- Always good to confirm this assumption holds by performing a weighted analysis and confirming it is similar to a non-weighted one