TB surveillance and surveys: a training workshop for consultants

TB prevalence surveys:
Sampling and sample size calculation

Day 2 – Wednesday, 25 May 2011
Geneva

Babis Sismanidis
with acknowledgements to Sian Floyd

The recommended 9 steps to sampling design and sample size calculation

1. A prior guess of the true population prevalence

2. The relative precision around the estimate of prevalence

3. A prior guess of the magnitude of the "design effect"

4. Apply the recommended equation for the sample size calculation of a TB prevalence survey

5. A prior guess of the participation rate
The recommended 9 steps to sampling design and sample size calculation (cont.)

6. Stratification to ensure a representative and precise overall estimate of prevalence

7. Cluster selection

8. Selection of individuals within cluster

9. Eligible survey population

1. A prior guess of the true population prevalence

   • First step is to make a prior guess for the true population prevalence of TB

   • Use national surveillance data (e.g. WHO Global TB Report) AND other available research data for plausible values

   • Use previous prevalence survey results (if surveys comparable) and guess by how much TB decreased

   • Must be done in close collaboration within a team including a statistician and local TB experts
2. The relative precision

- This (relative) precision refers to "how far away" we are allowing the survey's estimate of prevalence to be from the true national prevalence.
- In statistical terms this translates into the width of the 95% confidence interval around the TB prevalence estimate we expect the survey to give us.
- The more precise the estimate the larger the sample size.
- Recommended precision is between 20% and 25%.

E.g. If prevalence 200, then 95% CI (160, 240).

3. A prior guess of the "design effect"

- We sample groups of people (clustered-random design), and not individuals (simple-random design).
- Clustered-sampled (CS) surveys provide more imprecise estimates compared to individually-sampled ones (for the same set of parameters).
- Therefore, we need an increased sample size for CS (multiply sample size for SRS by a factor called the "design effect").
- We estimate it from previous surveys OR the likely variation of prevalence between clusters.
- Design effect gets bigger the bigger the:
  - difference in prevalence between clusters,
  - cluster size (recommend 400-1000 individuals),
  - expected TB prevalence is.
4. Sample size calculation based on precision

Sample size for clustered sampled survey =
(simple size for simple random sampling) x (design effect)

\[ N = \left[ 1.96^2 \left( \frac{1 - \pi_g}{d^2 \pi_g} \right) \right] \times \left[ 1 + (m-1) \frac{k^2 \pi_g}{1 - \pi_g} \right] \]

<table>
<thead>
<tr>
<th>( N )</th>
<th>Number of people included in the survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi_g )</td>
<td>“Prior guess” of the true population prevalence of pulmonary TB (expressed as a proportion)</td>
</tr>
<tr>
<td>( d )</td>
<td>Relative precision (expressed as a proportion). Recommended 0.20 or 0.25</td>
</tr>
<tr>
<td>( m )</td>
<td>Cluster size (=number of targeted individuals), assumed to be constant across clusters</td>
</tr>
<tr>
<td>( k = \frac{\sigma_B}{\pi} )</td>
<td>Coefficient of between-cluster variation. Recommended to assume is in the range 0.4 – 0.6</td>
</tr>
</tbody>
</table>
## 4. Sample size calculation; examples (cont.)

<table>
<thead>
<tr>
<th>TB prevalence</th>
<th>Cluster size</th>
<th>precision=0.2</th>
<th>precision=0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Difference in prevalence between clusters increase</td>
<td>Difference in prevalence between clusters increases</td>
</tr>
<tr>
<td></td>
<td></td>
<td>k=0.4 k=0.5 k=0.6</td>
<td>k=0.4 k=0.5 k=0.6</td>
</tr>
<tr>
<td>100 per 100,000</td>
<td>500</td>
<td>103612 107925 113197</td>
<td>66312 69072 72446</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>107453 113927 121849</td>
<td>68770 72914 77978</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>111295 119930 130484</td>
<td>71229 76755 83510</td>
</tr>
<tr>
<td>150 per 100,000</td>
<td>500</td>
<td>71598 75912 81183</td>
<td>45823 48583 51957</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>75440 81914 89827</td>
<td>48282 52425 57489</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>79282 87917 98470</td>
<td>50740 56267 63021</td>
</tr>
</tbody>
</table>

Estimating $k = \text{think about plausible values of 2.5\% and 97.5\% centiles of distribution of cluster-level TB prevalence}$

(based on knowledge of the country)

1) $k = \text{SD} / \text{true mean prevalence}$, by definition
2) A value of $k=0.5$ would mean SD=0.5*364=181
4. Sample size calculation; cluster size (cont.)

• This is recommended to be between 400-1000. It is a difficult choice we must make considering opposing parameters
• The smaller the cluster size, the larger the number of clusters required (larger number of field teams) and the more difficult standardization becomes
• The larger the cluster size, the larger the design effect (larger total sample size) AND the longer cluster operations become (recommend to complete a single cluster’s operations within a week)
• Population size of the smallest available geopolitical unit (i.e. census enumeration area) is usually a good starting point to guide target cluster size
• Final cluster size is a difficult balancing act of all these considerations

5. Adjust sample size for non-participants

• In a field survey, some people will either not attend the initial screening, or will drop out during the survey (say a proportion of p)
• Initially calculated sample size (say \(N_1\)) should be increased to allow for non-participation in the survey
• Corrected sample size (say \(N_2\)) for non-participation is:

\[
N_2 = \frac{N_1}{1-p}
\]

• Recommended value for participation percentage is 85–90% (the lower this is the more questionable the representativeness of the sample becomes):

\[\text{e.g. } N_2 = \frac{N_1}{0.85}\]
6. Stratification

- TB prevalence will typically vary across different geographical regions of a country (AKA strata)
- A PPS stratified design should be used to increase the precision and representativeness of the overall country estimate of TB prevalence
- This means number of clusters from each stratum is proportional to population size in the stratum
- By design the approach of stratification allows the estimation of stratum-specific estimates of TB, but their precision is smaller compared to the overall nationwide estimate

If Java-Bali/Sumatera/Eastern Indonesia population split is 50%/40%/10%, and we need to randomly select 100 clusters then 50 should come from Java-Bali, 40 from Sumatera and 10 from Eastern Indonesia.

7. Cluster selection

- Once cluster size (m) is chosen and sample size (N) calculated; number of clusters=N/m
- Examples of clusters are villages, household blocks, census enumeration areas
- Cluster selection is a multi-stage process starting from larger to smaller sampling units, to promote geographical representation
- Selection of clusters in every stage is proportional to population size (PPS)
- Selection at final stage (where sampling units are roughly equally sized) is random
e.g. Nigeria: Zones, states, LGAs, EAs
7. Cluster selection

8. Selection of individuals within cluster

- Once a cluster is selected, the target sample size of eligible survey individuals is identified and invited

- Cluster is split into household groups (roughly equally sized) using e.g. paths, roads, natural boundaries

- Household groups are selected at random

- Number of participants in each cluster should be as similar as possible across clusters enrolled in the survey (this follows from the use of PPS to select clusters AND simplifies the analysis; not essential to apply weights)
9. Eligible survey population

- Eligible survey individuals should be representative of the target population.

- Eligibility is based only on:
  - age (aged 15 years or older)
  - residency status in the household (e.g. people living in the household for the past 4 weeks between pre-census and census visits)

- All eligible individuals should be classified as:
  (i) participants, (ii) absentees and (iii) non-consenters

- Enumerating those under 15 years is also important for post-hoc adjustments to demographic changes at the analysis stage.

Sample size calculation based on repeat surveys

\[
N = \frac{(1.65 + \beta)^2 \times \{\pi(1-\pi) + (m_2 - 1)\pi^2 k_2^2\} \times N_1}{(\pi - p_1)^2 \times N_1 - (1.65 + \beta)^2 \times \{p_1(1-p_1) + (m_1 - 1)p_1^2 k_1^2\}}
\]

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Number of people included in the second survey</td>
</tr>
<tr>
<td>(\pi)</td>
<td>“Prior guess” of the true population prevalence of pulmonary TB for the second survey, expressed as reduction from the first (p_1)</td>
</tr>
<tr>
<td>(\beta)</td>
<td>The z-value of the test power = 1- (\beta) (0.84 or 1.24 for power of 80% or 90%)</td>
</tr>
<tr>
<td>(m_1, m_2)</td>
<td>Cluster size, for first and second survey respectively (assumed to be constant across clusters within each survey)</td>
</tr>
<tr>
<td>(k_1, k_2)</td>
<td>Coefficient of between-cluster variation, for first and second surveys respectively. Recommended to assume is in the range 0.4 – 0.6</td>
</tr>
</tbody>
</table>