Sampling design and sample size calculations in drug resistance surveys

Matteo Zignol
STOP TB Department
World Health Organization

TB surveillance and surveys: A training workshop for consultants
Geneva, 24-27 May 2011

Paragraph 5.4, pages 27-32
Sampling frame

Generally sputum-smear positive TB cases:

- No strong evidence to indicate that the proportion of drug resistance varies according to the smear status (positive or negative)

- Culture yield from smear negative cases is relatively low compared to smear positive cases → larger sample size → larger lab workload and implications

- As a minimum all patients in the public sector (NTP)
Sample size (I)

\[ n(SRS) = \frac{N \times z^2 \times p \times (1 - p)}{d^2 \times (N - 1) + z^2 \times p \times (1 - p)} \]

- **N** = total number of smear positive cases registered in the previous year area to be studied
- **p** = expected proportion of MDR-TB from available data
- **d** = desired absolute precision. It should be lower than 20% of **p**
- **z** = z-value that corresponds to the desired confidence level (if CI 95%, **z**=1.96)

- **n(SRS)***d.e.* if cluster sampling where d.e.=design effect (assumed to be 2 in drug resistance surveys)
Sample size (II)

- Sample size to be inflated of 20% to account for expected losses (culture not grow, contamination, DST unreadable,…)
  \[ N_1 = N / 0.8, \text{ where } N_1 = \text{ eligible, } N = \text{ sample size} \]

- Sample size for previously treated TB cases usually not feasible to reach → all previously treated TB cases presenting to the diagnostics unit in the intake period of new cases are enrolled

- In repeated surveys sample size calculated to detect a significant difference between proportions → sample size generally not feasible to reach
Sampling strategies (I)
100% sampling of diagnostic centers

All eligible patients enrolling at each diagnostic centre within the same limited intake period

Characteristics:
- Relatively small numbers of diagnostic centers
- Good infrastructure and transportation of samples
- Intake period = sample size divided by the number of sputum-smear positive diagnosed per year
- Enrolment either during the same period in all centers or on rotation (to prevent overload in the labs)
Sampling strategies (II)
Cluster sampling

Probability-proportional to size cluster sampling should be used

Characteristics:
- Minimum of 30 clusters → the number depends on inter-cluster variance and costs
- Cluster size between 10-40 patients → not too small (costs, logistics) or too large (sampling inefficiency)
Exercise 4: Calculation of a sample size for a drug resistance survey - new smear positive cases

We use data from Bulgaria, Senegal and Pakistan for this exercise.

A National Tuberculosis Programme wishes to estimate the prevalence of multi-drug resistant tuberculosis (MDR-TB) among new sputum smear positive (ss+) cases of TB cases in the country. How many new cases of pulmonary tuberculosis should be included in the sample so that the prevalence may be estimated to within one percentage point of the true value with 95% confidence?

1. From the description of the three settings below select a sampling approach

2. Part 1: Calculate the sample size for new cases taking into account: (i) the expected prevalence \( p \) of MDR*, (ii) the desired precision**, (iii) the desired confidence level in the estimate, e.g. 95%***, and (iv) the design effect****

Part 2: Inflate sample size to account for 20% expected loss

---

<table>
<thead>
<tr>
<th>Country</th>
<th>ss+/year</th>
<th>diagnostic units</th>
<th>Population size</th>
<th>Regular transport</th>
<th>Sampling approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgaria</td>
<td>1,266</td>
<td>30</td>
<td>7,585,131</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Senegal</td>
<td>7,883</td>
<td>66</td>
<td>12,534,228</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td>89,343</td>
<td>1,087</td>
<td>169,708,300</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Country</th>
<th>ss+/year</th>
<th>Expected prevalence of MDR (p)</th>
<th>1-p</th>
<th>Precision</th>
<th>Confidence interval (CI) *</th>
<th>Z-value corresponding to desired CI *</th>
<th>Design effect</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgaria</td>
<td>1,266</td>
<td>0.030</td>
<td>0.01</td>
<td>95%</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senegal</td>
<td>7,883</td>
<td>0.021</td>
<td>0.01</td>
<td>95%</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td>89,343</td>
<td>0.018</td>
<td>0.01</td>
<td>95%</td>
<td>1.96</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

For Question 1:
This is an exercise in feasibility. The variables considered should be number of diagnostic units in the area surveyed, regular transport to those facilities and ability to supervise the number of sites selected.

For Question 2:

For **simple random sampling** (SRS) or when the entire population is "sampled" use this formula:

\[ N = \text{smer-positive cases} \]

\[ z = \text{z-value that corresponds to the desired confidence level} \]

\[ d = \text{absolute precision (as a decimal, e.g. 0.01 or 0.02)} \]

\[ p = \text{expected prevalence of MDR} \]

For **cluster sampling** use this formula:

\[ n(\text{cluster-sampling}) = n(SRS) \times d.e. \]

The design effect (d.e.) for a cluster is assumed to be 2.