# Operational Studies in TB Control Programmes: Data Collection and Analysis

P.G. Gopi

## Abstract

Revised National Control Programme (RNTCP) through the application of Directly Observed Treatment-Short Course (DOTS) was implemented in India and aimed at least 85% cure 70% case detection rate (CDR). Operational studies are conducted on various components of the TB control programme with a view to monitoring and evaluating the programme. It gives insight in to the functioning and effectiveness of the programme. Policy makers and health planners use the findings to develop new strategies towards effective TB control.

First a protocol, a document which gives a description of the conduct of the study with a specific objective needs to be prepared. The design of the study should be appropriate in selecting the study population. Data collection and analysis are the two important components of the operational studies.

The methods adopted for data collection, standardization of the procedure, quality check, data entry & verification of the data, analysis, statistical test of significance and interpretation of the results will be discussed in detail with worked out example.

## Introduction

Tuberculosis (TB) remains a major cause of illness and death among adults despite being nearly 100% curable. Globally, the burden of TB was estimated to be 8.9 million new cases in 2004 of which about 3.9 million cases were sputum smear-positive. In India, the burden of TB disease was estimated to be 8.5 million for the year 2000. World Health Organisation (WHO) recommended and globally accepted Directly Observed Treatment- Short course (DOTS) strategy is considered as one of the most cost effective of all health interventions. The Revised National Tuberculosis Control Programme (RNTCP) was implemented in India in 1993 in a phased manner and the whole country was covered by March 2006. The aim of the RNTCP is to detect 70% of infectious TB cases and treat successfully atleast 85% of the detected infectious TB cases in the community. One of the challenges is to sustain the programme over the next several years.

National TB control programmes should not compromise on the quantity or quality of information. Operational research helps to evaluate and identify the areas that require strengthening in terms of quality and quantity. A good research question should be specific, concise and relevant to the programme. Operational studies are conducted on various components of TB control programme with a view to monitoring and evaluating the programme. Operational studies are essential to develop sustainable strategies, which will enhance the success of the programme and make TB control reality. These studies give an insight into the functioning and effectiveness of the programme. Policy makers, potential donors and health planners may use these findings of the studies towards achieving effective TB control.

Protocol

First of all, a protocol has to be prepared for a research question that needs to be answered. Protocol is a document which gives a description of the conduct of the study. It should start with a background which gives a description on why the study is undertaken based on some scientific support. The objective of the survey needs to be clearly mentioned in the protocol. The objective should be very specific and the conduct of the survey is to answer the objective. There can be more than one objective; primary and secondary. It is better not to have many objectives including one or two secondary objectives. The next part of the protocol is the study design and the methodology which depends on the objective of the study. The design should state the kind of study that will provide the relevant data to answer the research question. For example, if the objective is to estimate the prevalence of TB, a cross-sectional survey is usually undertaken in a representative random sample of the study population. For this, the sample size is estimated based on some basic assumptions as decided by the principal investigator and his team. Since we go for a sample survey, the sample size estimation is an important exercise. Small study sample fails to give precise estimate or detect important effects on the outcomes of interest. Large sample size results in waste of resources, time consumption and may not give the desired results. We should take a sample that gives a reasonable estimate of the parameter. For this, we should have some basic assumptions about the prevalence of TB, precision, level of significance, design effect and minimum coverage for the examination.

Eg: Sample size estimation for a TB prevalence survey: Sample size estimation is not an exclusively mathematical or statistical exercise. It should not be too low and too high. It should be adequate to give a precise estimate. Sample size is estimated based on certain assumptions which could be subjective. First, one should know roughly the prevalence (proportion 'p') of TB. To estimate the prevalence based on sample, one has to decide how precise the estimate should be i.e. the precision with which the prevalence is to be estimated. Precision is important based on the public health point of view and this is often expressed as % of p (= d). By normal approximation this is expected in 95% of the samples.

(1) Rough estimate for prevalence of TB= p

(2) Relative precision of the estimate = d (% of p)

(3) Significance level = 5% (normal value Z = 1.96)

Based on the above assumptions, the sample size is estimated using the binomial formula

$$n = \frac{Z^2 pq}{d^2}$$

Suppose, from previous experience it is known that prevalence of TB is about 4/ 1000. If an estimate of prevalence is required at an interval that extends 20% of prevalence ('d' units) on either side of the estimate and the population parameter is

$$n = \frac{z^2 pq}{d^2} = \frac{(1.96)^2 \text{ x } 0.004 \text{ x } 0.996}{\left(\frac{20 \text{ x } 0.004}{100}\right)^2} = 23914 \quad persons$$

expected to be in this interval except for 5% chance (significance level), then the sample size is estimated using the formula.

A simple random sample of about 24000 persons is required to be examined for TB. The increase in sample size needed to compensate for non-response is very costly. So, it is worthwhile to devote a substantial proportion of the resources to the reduction of non-response. The sample size estimated has to be adjusted for coverage. Suppose the coverage of 90% of the sample size is feasible for examination, then the required

sample size is $\frac{23914}{0.9} = 26571 \quad persons$

For operational and administrative convenience, usually, a cluster sampling method is adopted. To adjust for the cluster effect, the sample size is escalated by a factor called 'design effect' and this is usually estimated from the survey data by the ratio of the appropriate cluster-sample variance to the variance as if it were a simple random sample. Usually, the sample size is doubled. So, the adjusted sample size in our example after adjusting for cluster effect is 26571 x 2 =53142 ≈ 53000 persons.

After estimating the sample size for the study, the sample is selected by adopting a well defined sampling procedure for the conduct of the study. This is required to arrive at a valid conclusion after the study. The sampling procedure to be followed for the survey to select the sample is an important component of the survey. There are many sampling methods like simple random sample, stratified sampling, multi-stage cluster sampling, cluster sampling etc.

This part describes the data collection and analysis before writing up the report for dissemination. Before starting the survey, the questionnaire should be pilot tested for understanding the procedure, accessibility, feasibility and duration of the study.

## Data Collection

Reliable data is generated to answer the objective of the study. So, data collection is an important component of the study. There are two kinds of data namely, qualitative and quantitative. Usually quantitative type of data is collected from operational studies. An interview schedule is designed which contains the variables required to answer the objectives. The interview schedule should be semi-structured and pre-coded. It should be pilot tested before using it in the study to acquaint with the procedure. This is done by interviewing a few subjects not from the study population. The interview schedule is revised based on the experience in the pilot study. There is another document namely, survey manual which is followed in the field while generating data from a survey. It contains the details of the methodology for collecting reliable and valid data. Reliable data collection is an important prerequisite of the survey. Reliability refers to obtaining same results (response) in repeated measurements whereas validity refers to measurements that are not only reliable but also accurate and true. It has the provision

for coding and the instructions to record the information in individual questionnaire. Data should be collected keeping in view of the objective of the study. The success of the study to a great extent depends on reliable field work making provision for adequate supervisory staff. All efforts should be made to keep the procedures free from bias.

There are also two kinds of errors called sampling and non-sampling errors. The error involved in the conclusion drawn based on the evidence given by a sample is called sampling error and it is inherent and unavoidable in every sampling scheme. As distinct from sampling errors, the non-sampling errors arise at the stage of collection, compilation and analysis of the data and it can occur at any stage of the study. A good design should minimize the selection bias that occurs at every stages of the study namely; planning, selecting the study population and interviewing the participants. It should generate true and accurate data.

Work schedule is a small document, which gives the date the survey should be initiated concluded in a cluster. It also gives the order in which the clusters to be chosen for the survey. A small pilot survey to pretest the questionnaire is always advisable for improving data collection in the main survey. For the preparation of the work schedule, a planning visit to all the clusters is absolutely necessary to get an idea on the population of the cluster and to seek the cooperation of the participants.

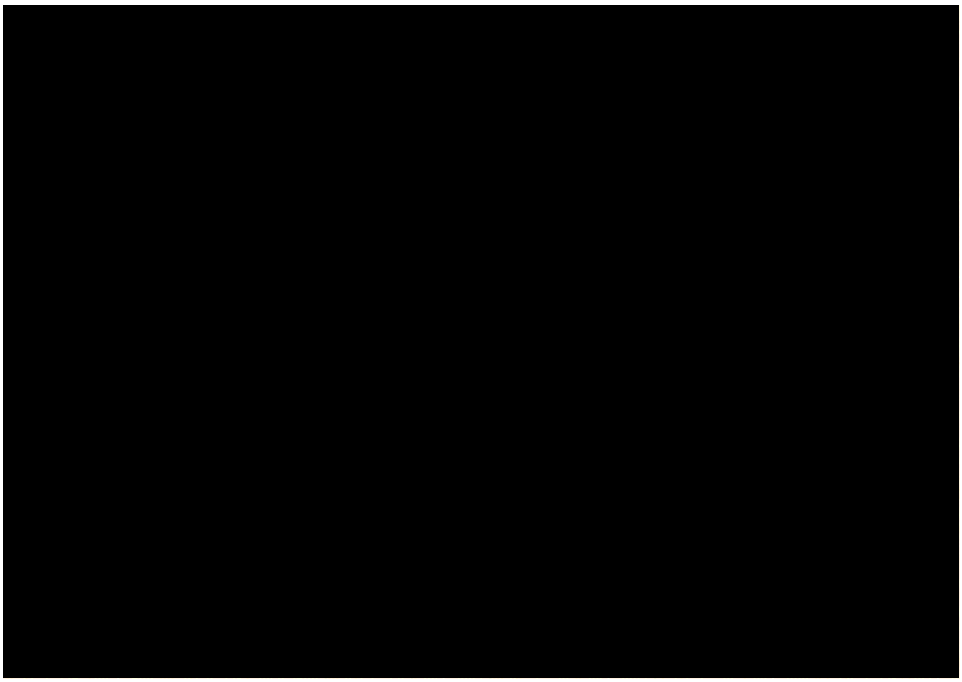
A planning visit to the population is essential before starting the survey. The map of the survey area is drawn during the visit. This map should contain the boundaries of the sampling unit (a village or an urban unit) and important landmarks for identification purposes. A meeting with the village leaders (politicians, important persons like village head or president, priest etc) is held and the purpose of the survey is explained to them to seek the cooperation from them and the participants of the survey as well. Ethical issues while conducting the surveys which include confidentiality, informed consent were taken from the participants of the survey.

We envisage a minimum coverage of 90% and provision is given in the sample population at the time of estimation of sample size. We should also aim to cover the all study participants in order to reduce the differences in the basic characteristics among those participated in the survey and not participated. It is quite reasonable to assume that the characteristics of the population in the covered and not covered population are same. So, the estimate obtained based on the covered population is valid and expected to be similar as if it would have been estimated from the entire sample. The estimate is, however, adjusted for non-coverage. Maximum efforts need to be taken to minimize the non-response from the participants. The increase in sample size needed to compensate the non-response is very costly. It is essential to devote a substantial proportion of resources to the reduction of non-response. If a participant is non-cooperative, he should be convinced of the importance of the survey and maximum efforts taken to include him in the survey. If the participant is absent for examination, repeat attempts should be made to contact him and collect the data.

The methodology should be standardized to collect consistent data. For this, adequate training is required to all the staff in the team including the supervisor. Adequate supervision is important at the time of data collection. Quality check is introduced to ensure good quality data collection. This is done at every stage of the data collection. A 5-10% sample is taken for this purpose for cross checking and in case of any deviation corrective steps taken to correct the data and avoid it subsequently.

## Data Analysis

It is the most important component of the survey. It consists of cleaning of data after computerization, tabulation of data, statistical analysis and interpretation of the results. After receipt of the documents the entire data set is scrutinized for computerization. The data is keyed in twice to avoid any mistakes. The computerized data is further edited and corrected for any missing information or discrepancy. The data is brought to a format which is amenable for analysis. The analysis is carried out using statistical packages like SPSS, EPI6, SAS, STATA etc. The data is tabulated accordingly to answer the question posed in the objective of the protocol. Appropriate statistical test of significance is performed and the results are interpreted correctly.



B/W

Note: The following portion gives a very preliminary, brief description on analysis and is incomplete.

## Type of analysis

### 1. Descriptive, association and inferential

Frequency distribution: Statistical data arranged in tables have some definite advantages over those descriptively stated. It is easily understood and facilitates quick comparisons.

Diagrammatic representation of the data has greater attraction and memorizing value than mere figures. There are different kinds of graphical representations namely; bar diagrams, histograms, frequency curve, pie diagram etc.

2. Measures of central tendency and dispersion

- Mean- average of the value
- Median- middle value when the values are arranged in ascending/ descending order
- Mode- most occurring value
- Standard deviation- closeness of the data to the mean value (it is defined on the mean of the sum of square of difference between the individual values from the mean value)
- Standard error- is the standard deviation of the sample.

The measures of central tendency alone are not enough to give a correct picture of a particular distribution. We need to have the information on the extent to which the items in a particular distribution are scattered around this central tendency (Dispersion), the direction of scatteredness; that is whether more items are attracted towards higher or lower values (Skewness) and the extent to which the distribution is more peaked or more flat-topped than the normal distribution (Kurtosis)

3. Proportion and Rate

- Proportion: a risk is essentially a proportion or equivalently a probability. Where numerator is the number of individuals experienced the event of interest and denominator is the total number of individuals followed-up for the defined period.
- Rate: a rate takes in to account both the number of persons at risk & the duration of observation for each person. Where numerator is same as the above and denominator is expressed on the number of person years at risk.

Calculating proportion/ rate: Eg: 120 persons were followed up for 3 years and 40 of them died at the some time during the period.

Proportion = 40/120 × 100 = 33.3%

Rate = 40/300 × 100 = 13.3%

(80 persons followed up for three years and the 40 were assumed to be died, on an average, at the mid-point ie. at 1.5 years. So, the total number of person years is (80×3) + (40×1.5) =300).

4. Confidence interval

An estimate is subject to sampling error because it is based on a sample not on the whole population. The methods of statistical inference allow drawing conclusions about the true value from the sample values.

Precision of the estimate is obtained by attaching a confidence interval to the estimate. It is the range of plausible values for the true value based on the observations in the study. A 95% confidence interval means there is a 95% chance that it includes the true value.

5. Test of significance

A procedure by which a decision regarding the acceptance or rejection of the null hypothesis (of equality) is taken based on the information supplied by the sample. The statistic calculated is referred to the appropriate set of statistical tables to determine the 'P' value. The 'P' value measures the probability of obtaining a value for the statistic as extreme as the one actually observed if the null hypothesis is true. The smaller the 'P' value it is unlikely that the null hypothesis seems an explanation of the observed data (when at a level of significance the null hypothesis is rejected) and the observed result is termed as 'significant' otherwise 'non significant'. There is an inherent uncertainty in any conclusion about the null hypothesis arrived at using a test of significance.

There are two kinds of errors. Type I error is the probability of rejecting a null hypothesis when it is true. Type I error occurs if an investigator fails to accept a null hypothesis that is actually true in the population. Type II error is the probability of accepting the null hypothesis when it is false. A type II error occurs if the investigator fails to reject the null hypothesis that is actually false in the population. Power is the probability of obtaining a significant difference. Thus power of 80% to detect a difference of a specified size means that if the study is repeated a statistically significant result will be obtained four out of five times if the true difference was really of the specified size. The power of the study depends on the true difference between the groups, the study size and the level at which a difference is regarded as statistically significant.

Confidence interval for a proportion: Eg: In a random sample of 200 children from a population number suffering from a disease is 80

$p = 80 / 200 = 0.4$

$V(p) = pq / n = 0.0012,$

$ó(p) = v(pq / n) = 0.035$

95% CI = $p \pm Zá\ ó(p) = 0.4 \pm 1.96 \times 0.035$ = 33% - 47% (Zá is the normal value)

There is a 95% chance that the true prevalence in the population from where the sample taken is in the interval 33-47%.

6. Association between exposure and disease

i. Relative risk (RR)

Cohort and case-control studies are designed to determine whether there is any association between exposure to a factor and development of a disease. A sample of individuals, some exposed to risk factor of interest and some not, are followed over time, and the rates of subsequently contracting the disease in the two groups are compared. The measure, Relative risk (RR) summarizes the strength of association between the risk factor and the disease. Thus, if there are n individuals distributed by disease status and exposure to a factor as shown in the following table

| | Disease | No Disease | Total |
|---|---|---|---|
| Exposed | a | b | a + b |
| Not exposed | c | d | c + d |
| Total | a + c | b + d | n=a + b + c + d |

Incidence among exposed $= \frac{a}{a+b}$

Incidence among non exposed $= \frac{c}{c+d}$

Relative Risk (RR) = (Incidence in exposed) / (Incidence in not exposed)

$$= \frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

Interpretation of Relative Risk (RR) of a disease

| | |
|---|---|
| If RR = 1 | Risk in exposed equal to risk in non-exposed (no association) |
| If RR > 1 | Risk in exposed greater than risk in non-exposed (positive association; possibly casual) |
| If RR < 1 | Risk in exposed less than risk in non-exposed (negative association; possibly protective) |

Its statistical significance can be tested using a 2 x 2 $\chi^2$ test.

ii. Odds ratio (OR)

In case-control studies, the incidence in the 'exposed' population or 'not exposed' population can not be estimated. In order to assess the association between a cause and an effect in such studies, another popular measure called 'Odds Ratio (OR)', the ratio of two odds i.e. the ratio of odds of developing disease in the exposed group to the odds of developing disease in the unexposed group. Suppose 'a' and 'b' are the number of cases and non-cases, respectively in the exposed group and 'c' and 'd' are the corresponding figures in the not exposed group, then

$$OR = \frac{\left(\frac{a}{b}\right)}{\left(\frac{c}{d}\right)} = \frac{ad}{bc}$$

where the numerator is the odds of developing disease in the exposed group (a/b) and the denominator (c/d), the odds of developing disease in the unexposed group.

The standard error of log (OR) is

$$\sigma_{\log(OR)} = \left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)^{1/2}$$

The test of significance of log (OR) is carried out by calculating

$z = $ [redacted] and the 95% CI is estimated as log(OR) ± Zα × σlog(OR). Confidence limits for OR may be calculated as: exp [log(OR) ± Zα × σlog(OR)].

For a rare disease, the odds ratio is numerically equivalent to relative risk.

Worked out example for OR:

| Smoking | Stroke | | Total |
|---|---|---|---|
| | Yes | No | |
| Yes | 171 | 3264 | 3435 |
| No | 117 | 4320 | 4437 |
| Total | 288 | 7584 | 7872 |

Odds of developing disease in the exposed group (a/b) to the odds of developing disease in the unexpected group (c/d).

$$OR = \frac{171 \times 4320}{117 \times 3264} = 1.93$$

Var log (OR) $= \frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$

$$= \frac{1}{171}+\frac{1}{3264}+\frac{1}{117}+\frac{1}{4320} = 0.0149$$

σ log (OR) = $\sqrt{(0.0149)}$=0.122

95% CI =log (OR) ± Zα ×σlog (OR)

=0.6575 ±1.96 × 0.122 = 0.4184-0.8966 (1.52- 2.45 by taking anti-logarithm)

The interpretation (OR=1.93) is that the chance of getting a stroke among smokers is about two times higher compared to non-smokers.

7. Correlation

A measure of the mutual relationship between two variables or it is the degree of association between two variables.

If $X_i$ (height) & $Y_i$ (weight) are two variables, the correlation coefficient

$r = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_1-\bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X})^2(Y_i-\bar{Y})^2}}$ =0.0149, where

$\bar{X}$ and $\bar{Y}$ are the means of height and weight respectively.

$X_i$ =71,68,66,67,70,71,70,73,72,65,66

$Y_i$ =69,64,65,63,65,62,65,64,66,59,62

r =0.55 means the extent of agreement between height and weight is 55%

8. Regression

It is the change in the dependent variable per unit change in the independent variable.

There is a variation in Blood pressure (BP) at different ages. It is logical increasing age may affect BP. Prediction of BP for a particular age is done using regression analysis. Age is the independent (y) & BP the dependent (x) and assuming linear relationship, y=a+bx where b is the regression co-efficient defined by

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})2}$$

Eg: age (35, 45, 55, 65, 75); BP (114, 124, 143, 158, 166)

y=65.1+1.38x, where b=1.38 and interpreted as BP increases by 1.38 units per year

9. Measures of association

Chi-square ($\chi^2$) is used to test the association between two variables. We can use this for testing the significance of difference in proportions from zero, and testing the odds ratios or relative risk from unity.

$\chi^2 = n \times [(ad-bc-n/2)]^2 / [(a+b)\ (c+d)\ (b+c)\ (b+d)]$

| Vit A deficiency | Diarrohea | | Total |
|---|---|---|---|
| | Yes | No | |
| Yes | 133 | 380 | 513 |
| No | 99 | 1692 | 1791 |
| Total | 232 | 2072 | 2304 |

$\chi^2 = 181$; P <0.001: The null hypothesis that there is no association between Vitamin A and diarrhea is rejected. This means occurrence of diarrhea is associated with Vitamin A deficiency.

10. Extent of agreement between two methods

The extent of agreement may be measured by the following measures

| Test result | Gold standard | | Total |
|---|---|---|---|
| | Positive | Negative | |
| Positive | a | b | a+b |
| Negative | c | d | c+d |
| Total | a+c | b+d | a+b+c+d(=n) |

i. Sensitivity = (a/ (a+c)

It is the proportion of positive cases rightly detected by the test as positive

ii. Specificity= d/ (b+d)

It is the proportion of negative cases rightly detected by the test as negative

iii. Over diagnosis= (b/ (a+b)

It is the proportion of the test positive cases that are really negative cases by gold standard.

iv. Under diagnosis=(c/ (c+d)

It is the proportion of the test negative cases that are really positive cases by the gold standard.

v. Kappa statistics= $(P_{0-} P_c) / (1-P_c)$

The extent of agreement between the two methods afteradjusting for chance expected agreement, where $P_0$ is the crude agreement and $P_c$ is the agreement corrected for chance agreement.

Po= (a+b)/n

Pc = $[\{(a+c)(a+b)\}+\{(c+d)(b+d)\}]/n^2$

Kappa is interpreted as follows.

| Kappa | Agreement |
|---|---|
| < 0.20 | Poor |
| 0.21-0.40 | Fair |
| 0.41-0.60 | Moderate agreement |
| 0.61-0.80 | Good |
| >0.81 | Very good |

Acknowledgements

I am grateful to Dr. P.R. Narayanan, Director, TB Research Centre for permitting to participate in the National Symposium on Tribal Health held at RMRCT, Jabalpur, M.P last year and present a paper on 'Operational studies in TB control programme- Data collection and analysis' and prepare this report to include in the proceeding. I thank the committee members of the symposium for having given a chance for both participation in the symposium and inclusion of this report in the proceedings.

Suggested books for reading

1. An introduction to Statistical Methods, C.B. Gupta, Ram Prasad and sons, New Delhi.
2. Biometry, The principles and practice of statistics in biological research, Robert R. Sokal, F. James Rohlf, W.H.Freeman and Company, San Francisco.
3. Epidemiology Second Edition, Leon Gordis, MD, MPH, DrPH, W.B. Saunders Company, A Harcourt Health Sciences Company, Philadelphia.
4. Nonparametric Statistics for the behavioral Sciences, International Student Second Edition, Sidney Siegel, Mcgraw-Hill Book Company, Inc, New York
5. Sampling Techiques, William G. Cochran, Asia Publishing House, Bombay.
6. Sampling Theory and Methods, M. N. Murthy, Statistical Publishing Society, Calcutta.

7. Sampling Theory of Surveys with Applications, Third Edition, Sukhatame P.V., Sukhatame S., Sukhatame B.V., Asok C., IOWA Society Press, AMES, IOWA (USA) and Indian Society of Agricultural Statistics, New Delhi.

8. Statistical Analysis of Epidemiological Data, Second Edition, Steve Selvin, Oxford University Press, New York.

9. Statistical Methods in Medical Research, Second Edition, P. Armitage, G. Berry, Blackwell Scientific Publications, Oxford.

10. Statistical Methods, Sixth Edition, George W. Snedecor, William G. Cochran, Oxford & IBH Publishing Co. Calcutta.