Chapter 2

Methodology

The methodology for the study includes the following components.

(1) Study Design

(2) Data Collection and Management

(3) Statistical Analysis

(4) Graphical Methods

Study Design

The association between a suspected determinant and an outcome of interest in the pulmonary TB patients is investigated. A cross-sectional study is chosen for this investigation, as the determinants do not vary with time, enabling such an association to be measured.

Population : The target population comprises all patients diagnosed with pulmonary tuberculosis with HIV serosurveillance by unlinked anonymous testing twice a year within Zonal TB Center.

Sample : The sample comprises the patients diagnosed with pulmonary tuberculosis with HIV serosurveillance by unlinked anonymous testing twice a year, within Zonal TB Center 11, Nakhon Si Thammarat from November 1, 1994 to September 20, 1998, a total of 1,080 cases. The samples were divided into following two groups.

Study group : Pulmonary TB patients with HIV infection (124 cases).

Comparison group : Pulmonary TB patients without HIV infection (956 cases).

1. Inclusion Criteria

1.1 New patients with pulmonary tuberculosis who were diagnosed by a positive result of acid fast bacilli with direct smears and / or abnormal chest x-ray findings and had not been treated for over 1 month.

1.2 They had been investigated for HIV infection by Elisa method and confirmed with Western Blot.

2. Exclusion Criteria

2.1 New TB patients with no intrathoracic tuberculosis.

2.2 New TB patients with HIV serosurveillance with an indeterminate result.

3. Variables in the Study

3.1 Determinant variables :

- HIV status : negative or positive

3.2 Outcome variables :

- Characteristics of disease : cough, haemoptysis, chest pain, dyspnea, fever, weakness, weight loss, other symptoms, chest x-ray finding and degree of sputum.

3.3 Intervening variables :

- Health status and behaviours : receiving BCG, other diseases, smoking and drinking.

3.4 Stratification variables :

- Demographic factors : gender, age, marital status, occupation, religion and family history with TB.

Data Collection and Management

The secondary data were collected from the medical records (treatment card and x-ray card) and HIV serosurveillance records of new patients diagnosed as having pulmonary tuberculosis by the physicians at Zonal TB Center 11, Nakhon Si Thammarat and from the community hospitals in Nakhon Si Thammarat where the cases were referred to.

HIV serosurveillance data were collected by the unlinked anonymous testing method. In this study, it is necessary to link the data of HIV serosurveillance with the medical records as permitted by the director of Communicable Disease Control Regional 11, Nakhon Si Thammarat. These data are kept confidentially by the researcher.

The data were put into a database system using the Microsoft Access program as shown in Figure 2.1. Microsoft Excel was used to load the data from Microsoft Access and replace missing values. Programmer's file editor (PFE) was used to develop programs for analysis using the student version of Matlab 5. The data were analysed using ASP *(A Statistical Package)* (McNeil, 1998a). This is a suite of functions for graphing and analysing statistical data. These programs are also used with the student Matlab version 5 program (software that runs under Microsoft Windows, Macintosh, and Unix operating systems) which has limited storage capability. Thus, the data were separated into two files for analysis by ASP. The data structure is listed in the appendix.

1. Descriptive Statistics

The variables of interest are summarized by histograms and by means, standard deviations, and minimum, and maximum values. The demographic factors are described by percentages.

2. Univariate Analysis

Pearson's chi-squared test and 95% confidence intervals for odds ratios are used to assess the associations between the determinant variables and the outcome of this study. The Mantel-Haenszel chi-squared test and the Mantel-Haenszel adjusted odds ratio are used to adjust for confounding. The formulas of contingency tables (McNeil, 1998b) are as follows (X is the determinant of interest, Y is the HIV status, Z is a stratification variable)

(a) 2 x 2 tables

X is the determinant and Y is the outcome. Each variable is binary (0 or 1).

	Y = 1	Y = 0
X = 1	a	b
X = 0	c	d

n = a + b + c + d

Pearson's chi-squared statistic is defined as

$$X^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \quad (1)$$

The odds ratio is $$w = \frac{ad}{bc} \quad (2)$$

and its asymptotic standard error is given by

$$SE(\ln w) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad (3)$$

a 95% confidence interval is thus

$$95\%CI = w \times \exp(\pm 1.96\ SE[\ln w]) \quad (4)$$

(b) Stratified 2 x 2 tables

X is the determinant, Y is the outcome, Z a stratification variable. X and Y are binary (0 or 1), Z has s levels.

	Y = 1	Y = 0
X = 1	a_k	b_k
X = 0	c_k	d_k

Stratum k : $n_k = a_k + b_k + c_k + d_k$

Mantel-Haenszel (1959) chi-squared statistic is defined as

$$X^2 = \frac{(\sum(a_k d_k - b_k c_k)/n_k)^2}{\sum(a_k + b_k)(c_k + d_k)(a_k + c_k)(b_k + d_k)/(n_k - 1)n_k^2} \quad (5)$$

and also the odds ratio, adjusted for confounding is given by

$$w = \frac{\sum a_k d_k / n_k}{\sum b_k c_k / n_k} \quad (6)$$

Breslow and Day (1980) gave a chi-squared test for homogeneity and Robins et al (1986) gave a formula for the standard error of ln(w), (McNeil, 1996 : 105), that is

$$SE[\ln w] = \sqrt{\frac{\sum P_k R_k}{2R_+^2} + \frac{\sum(P_k S_k + Q_k R_k)}{2R_+ S_+} + \frac{\sum Q_k S_k}{2S_+^2}} \quad (7)$$

where $R_+ = \sum R_k$, $S_+ = \sum S_k$

and $P_k = \frac{a_k + d_k}{n_k}, Q_k = \frac{b_k + c_k}{n_k}, R_k = \frac{a_k d_k}{n_k}, S_k = \frac{b_k c_k}{n_k}$

(c) Non-stratified r x c tables

X is nominal (1, 2, …, r), Y is nominal (1, 2, …, c).

	Y = 1	Y = 2	.....	Y = c
X = 1	a_{11}	a_{12}	.....	a_{1c}
X = 2	a_{21}	a_{22}	.....	a_{2c}
.	.	.	.....	.
X = r	a_{r1}	a_{r2}	.....	a_{rc}

Pearson's chi-squared statistic is given by

$$X^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(a_{ij} - \hat{a}_{ij})^2}{\hat{a}_{ij}} \tag{8}$$

The odds ratios may be defined as (McNeil, 1998b)

$$w_{ij} = \frac{a_{ij}d_{ij}}{b_{ij}c_{ij}} \tag{9}$$

with standard errors given by

$$SE(\ln w_{ij}) = \sqrt{\frac{1}{a_{ij}} + \frac{1}{b_{ij}} + \frac{1}{c_{ij}} + \frac{1}{d_{ij}}} \tag{10}$$

where, $b_{ij} = \sum_{j=1}^{c} a_{ij} - a_{ij}, c_{ij} = \sum_{i=1}^{r} a_{ij} - a_{ij}, d_{ij} = n - a_{ij} - b_{ij} - c_{ij}$

(d) Stratified r x c tables

X is nominal (1, 2, …, r), Y is nominal (1, 2, …, c), Z has s levels.

	Y = 1	Y = 2	…..	Y = c
X = 1	a_{11k}	a_{12k}	…..	a_{1ck}
X = 2	a_{21k}	a_{22k}	…..	a_{2ck}
.	.	.	…..	.
X = r	a_{r1k}	a_{r2k}	…..	a_{rck}

Stratum k : $n_k = a_{ijk} + b_{ijk} + c_{ijk} + d_{ijk}$

Birch (1965) gave a matrix formula for a chi-squared statistic, generalising both Pearson's & Cochran's test. The odds ratios may be defined as (McNeil, 1998b)

$$w_{ij} = \frac{\sum a_{ijk} d_{ijk} / n_k}{\sum b_{ijk} c_{ijk} / n_k} \tag{11}$$

where, $b_{ijk} = \sum_{j=1}^{c} a_{ijk} - a_{ijk}, c_{ijk} = \sum_{i=1}^{r} a_{ijk} - a_{ijk}, d_{ijk} = n_k - a_{ijk} - b_{ijk} - c_{ijk}$

$$SE(\ln w) = \sqrt{\frac{1}{a_{ijk}} + \frac{1}{b_{ijk}} + \frac{1}{c_{ijk}} + \frac{1}{d_{ijk}}} \quad (12)$$

3. Multivariate Analysis

A statistical model used for analysis of epidemiological data is the logistic regression model that requires the response variable be dichotomous. It is well suited for the analysis of the binary outcome data and can handle general exposure variables, not just dichotomous ones. For this study, the outcome is the HIV status (negative and positive), thus it is reasonable to use the logistic regression model. This model takes the form (McNeil, 1996)

$$\ln\left(\frac{p}{1-p}\right) = a + b_1 x_1 + b_2 x_2 + \ldots + b_j x_j \quad (13)$$

where p denotes the probability of occurrence of the outcome and x_j represents the j^{th} determinant. This equation may be inverted to give an expression for the probability p as

$$P = \frac{1}{1 + \exp(-a - \sum_{j=1}^{p} b_j x_j)} \quad (14)$$

The functional form of the right-hand side ensures that its values are always between 0 and 1, which is reasonable given that they are probabilities. Thus the odds ratio for comparing two levels of a determinant x_j is given by

$$OR = \exp\left(\sum_{j=1}^{p} b_j (x_j^{(1)} - x_j^{(0)})\right) \quad (15)$$

The parameter b in the model may be interpreted directly as the (natural) logarithm of the odds ratio.

Variable selection and steps in model building

1. The selection process begins with a careful univariate analysis of each variable. It is also a good idea to estimate the individual odds ratios with confidence intervals to assess associations between the outcome and determinants.

2. Then, select variables for the multivariable analysis. Any variable whose univariate test has a p-value < 0.25 should be considered as a candidate for the multivariable model along with all variables of known biologic importance or of particular interest to the investigator. Other useful approaches to variable selection are the 'best subsets' technique and stepwise methods. Then a model containing all of the selected variables (p-value < 0.05) is established.

3. Following the fit of the multivariable model, the importance of each variable included in the model should be verified.

4. After obtaining a model that contains the essential variables, the need for including interaction terms among the variables in the model is considered.

Specification of the model

The logistic regression model, which is being increasingly applied in epidemiologic research, may be used by defining the dependent variable (Y) to be the dichotomous results of a screening test, where Y=1 if the disease is presumed to be present, and Y=0 otherwise. The presence or absence of disease, as defined by the "gold standard" in logistic regression model, is included as a binary explanatory variable (x_1), along with variables used to define the subgroups of interest. Thus, the log odds of presumptive disease is modeled as a linear function of (1, ..., j) explanatory variables, one of which corresponds to the binary results of the "gold standard", along with their β-coefficients.

$$\text{logit}\,\Pr(Y=1|X)=\alpha+\sum_{j=1}^{J}\beta_j X_j \qquad (16)$$

The logistic coefficients β_j do not have a simple relationship to differences in risk. We have to use the estimated coefficients to calculate predicted probabilities p, and then compare these probabilities or risk to each other.

Graphical Methods

The data are graphed by using ASP with the student version of Matlab 5. The data are graphed by histograms and statistical summaries of a set of variables that are determinants, outcomes, or other variables. The association between the outcome variables and the determinants of interest are graphed by odds ratio plots. The logistic regression in the multivariate analysis is graphed by logistic regression print outs, which may include standardized residuals plots.

1. Histogram

A histogram (or bar graph) presents the data as bar extending away from the axis representing independent variable. The length of each bar is determined by the value of the dependent variable.

2. Odds Ratio Plot

The associations between the outcome variables and the determinants of interest are investigated by an odds ratio which provides a useful measure. The graph of an odds ratio also includes a 95% confidence interval. Confidence intervals may be graphed using line intervals. The dot on the line interval is the estimated odds ratio. For an odds ratio, the null value is conventionally taken to be 1, corresponding to equal risks of an outcome in two comparison groups. This corresponds to a null value of 0 for the difference between two population proportions under the null hypothesis represented by the dotted line. The p-value shown at the top is the overall (Pearson's) chi-squared test of no relationship between the determinant and outcome. Additionally, the p-value shown in the horizontal panels of the graph may be used to assess the associations between the outcome and a set of binary determinants obtained by aggregating the counts for the unspecified levels of the determinant. The homogeneity test is used to tell if the association could be the same in the different strata, a small p-values providing evidence to the contrary.

3. Logistic Regression Print Out

The odds ratio measures the association between the determinant and outcome. If a third variable (a possible confounder) needs to be considered, a stratified analysis may be carried out. For statistical modeling, logistic regression analysis provides an alternative approach to Mantel-Haenszel methods for analysing contingency tables. The results should agree with those obtained from the Mantel-Haenszel analysis. This is based on using the odds ratio as a test statistic for the null hypothesis that the odds ratio in the target population is 1. In the printout from the logistic regression analysis, the numbers labelled *coeff* and *St.Error* are the estimates of the parameter b (the natural logarithm of the odds ratio) and its standard error, respectively. The second row coefficient gives the value of the parameter a. The combination of determinants commencing with 0 coefficients (each indicated with the symbol (0) in the second column of the print out) is referred to as the referent category. The numbers in the bottom line, the first (df) is the number of degrees of freedom remaining in the table after fitting the logistic model. The second number (deviance) is a measure of the error after fitting the model. Finally, the number of iterations is the number of steps in the model fitting procedure.