Chapter 2

Methodology

This chapter describes the methods used in the study. These methods include the following components.

1 Study design

2 Variables

3 Data collection, management and data analysis

4 Graphical and statistical methods

2.1 Study Design

Our study has a quantitative component and a qualitative component. The quantitative component is a cross-sectional design involving the collection of data retrospectively based on records of patients visiting the Thai Traditional Medicine Clinic in Khokpho Hospital of Pattani Province. The qualitative component involves interviewing a small number of both these patients and the staff working in the clinic.

The study population comprises 327 patients who visited the Thai Traditional Medicine Clinic in Khokpho Hospital of Pattani Province between 1 January 2003 and 31 December 2003, together with six patients and six staff members.

2.2 Variables

The schematic diagram for this study is shown in Figure 2.1.

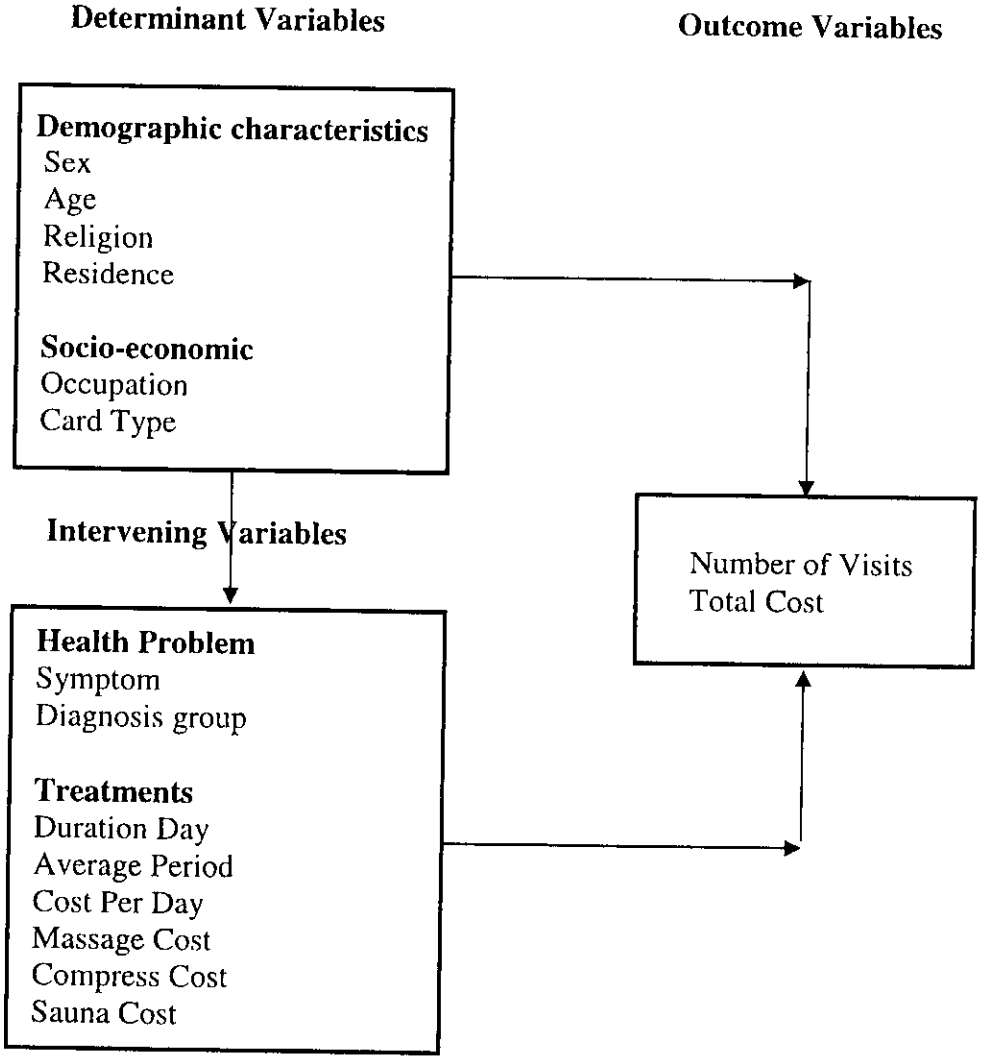

Figure 2.1: Putative relationships between the variables.

Determinant variables

The main determinants of interest are (1) demographic characteristics (sex, religion, age and residence, (2) socio-economic characteristics (occupation and card type), and (3) health problem symptoms, diagnosis group, patient type and treatment.

Intervening variables

Patients visiting the Thai Traditional Massage have many symptoms when they use

massage, and many diagnostic groups, as well as three treatments (massage, sauna and compress), so it is important to classify them according to intervening variables. These consist of health problems (patient type, symptom and diagnosis group) and treatment components (duration, average period, cost per day, massage cost, compress cost and sauna cost).

Outcome variables

Total cost and number of visits are the outcome variables of interest.

2.3 Data Collection, Management and Data Analysis

Data collection

Secondary data were collected from the Thai Traditional Medicine Clinic at Khokpho Hospital in Pattani province from 1 January 2003 to 31 December 2003, and retrospective data concerning visits by these patients going back to October 1999. These data are available in medical records kept at the clinic. There were 327 patients and their total number of visits was 1319.

Qualitative data were collected by unstructured, opportunistic interviews for six patients who were treated with Thai Traditional Massage in 2003. All patients selected for interview agreed to participate. Six patients who were treated with Thai Traditional Massage in 2003 and 2004 were interviewed twice by using an unstructured questionnaire for at least 30 minutes for each patient. The interview included information about demographic characteristics, the patient's attitudes, belief, satisfaction, health problems, accessibility to clinic, and reasons for visiting the clinic (see Appendix A). Six staff members were also interviewed using an unstructured questionnaire, which also took approximate 30 minutes to complete for each staff (see

Appendix B). The interview included information about demographic characteristics, work experience, attitudes towards the occupation, and reasons why they became staff members.

Data management

The data are stored in Excel, imported to Microsoft SQL Server and analyzed using *Webstat*, the new statistical expert system technology written in HTML and VBScript. The Excel add-in *Ecstat* is also used for statistical data analysis. This is a suite of functions for graphing and analysis of statistical data.

Data analysis

Statistics for descriptive analysis include percentages, means, and standard deviations. One way analysis of variance is used to test that the population means of the outcome variable corresponding to the different categories of the determinant are the same. Pearson's chi-squared test is used to assess associations between categorical variables.

2.4 Graphical and Statistical Methods

Graphical Methods

Histograms are used to graph the frequency distributions of continuous variables that could be determinants, intervening and outcomes.

Histograms and bar charts with statistical summaries of raw data for all variables are used to represent the distribution. These summaries include the size, mean, standard deviation, minimum and maximum of a set of data. A bar chart presents the data as bars extending away from the axis representing the frequencies.

One way analysis of variance is described by using box plots and 95% confidence intervals for means. Box plots are abbreviated histograms invented by Tukey (1977).

The box covers the central 50% of the distribution of the data between the lower quartile and the upper quartile. It could be denoted by a rectangular box, with the median given special attention. Unusually high or low measurements (outliers) are plotted separately, with the bulk of the data represented simply by continuous lines. An outlier is defined as an observation that is more than distance D away from a quartile, where D is the inter-quartile range, that is

$$\text{D} = \text{Upper Quartile} - \text{Lower Quartile}$$

Statistical Methods

This study focuses on the association between the outcome and the demographic and socio-economic determinants.

One-way analysis of variance

Considering the analysis of data in which the outcome is continuous and the determinant is categorical, this leads to a procedure called the (one-way) analysis of variance (anova). The null hypothesis is that the population means of the outcome variable corresponding to the different categories of the determinant are the same, and this hypothesis is tested by computing a statistic called the *F-statistic* and comparing it with an appropriate distribution to get a *p-value*. Suppose that there are n_j observations in sample j, denoted by y_{ij} for i = 1, 2, …, n_j. The F-statistic is

$$F = \frac{(S_0 - S_1)/(c-1)}{S_1/(n-c)}$$

where $$S_0 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2,\ S_1 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$$

and $$\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n_j} y_{ij},\ \bar{y} = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij},\ n = \sum_{j=1}^{c} n_j$$

S_0 is the sum of squares of the data after subtracting their overall mean, while S_1 is the sum of squares of the residuals obtained by subtracting each sample mean. If the population means are the same, the numerator and the denominator in the F-statistic are independent estimates of the square of the population standard deviation (assumed the same for each population). The p-value is the area in the tail of the F-distribution with c-1 and n-c degrees of freedom (McNeil, 1996).

Pearson's test of independence

Pearson's chi-squared test is used to assess the association between categorical determinants and outcome variables.

In this study, some of variables are multi-categorical. For example, X is category of occupation and Y is category of number of visits group. We use an $r \times c$ contingency table to compare them. This takes the form

	$y = 1$	$y = 2$	...	$y = c$
$x = 1$	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 for independence (i.e., no association) in the $r \times c$ table is defined as

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(a_{ij} - \hat{a}_{ij})^2}{\hat{a}_{ij}} \quad ,$$

where $\hat{a}_{ij}$ is the expected value of a_{ij}. When the null hypothesis of the independence is true, this has a chi-squared distribution with (r-1)(c-1) degrees of freedom (McNeil, 1998).

Multiple linear regression

Regression used to analyse data in which both the determinants and the outcome are continuous variables. It can summarise the data in the scatter plot by fitting a straight line. In conventional statistical analysis the line fitted is the *least squares line*, which minimises the distances of the points to the line, measured in the vertical direction. If there is more than one determinant, the method generalises to multiple linear regression, in which the regression line extends to the multiple linear relation represented as.

$$Y = \beta_0 + \sum \beta_i x_i + \varepsilon,$$

where Y is the outcome variable, β_0 is a constant, $\{\beta_i\}$ is a set of parameters (i = 1 to p), and $\{x_i\}$ is a set of determinants (i = 1 to p) (McNeil, 1998).

The model is fitted to data using least squares, which minimises the sum of squares of the residuals.

After the usual assumption of independent observations is made, linear regression analysis rests on three assumptions as follows.

(1) The association is linear.

(2) The variability of the errors (in the outcome variable) is uniform.

(3) These errors are normally distributed.

If these assumptions are not met, a transformation of the data may be appropriate. Linear regression analysis may also be used when one or more of the determinants are categorical. In this case each categorical determinant is broken down into c-1 separate binary determinants, where c is the number of categories. The omitted category is taken as the baseline or referent category.