Chapter 2

Methodology

2.1 Study Design

We conducted a hospital-based retrospective study of twin births in Pattani Hospital for the period from 1 October 1996 to 30 September 2005.

Population: The target population comprises all women who delivered in Pattani and their newborn infants.

Sample: The study sample comprised all occasions of a delivery (single or multiple births) and all of twins at Pattani hospital during 1 October 1996 to 30 September 2005.

Inclusion Criteria

1. Mother who had twins.
2. Mother who had single births.

Exclusion Criteria

1. Mother who had triplet, quadruplets and quintuplets.
2. Referral mothers from other hospitals to Pattani hospital.

2.2 Data Collection

Data collection

Computerized data of the newborn infants delivered at Pattani hospital from 1 October 1996 to 30 September 2005 were obtained. These data were also kept in the form of the labour chart records report.

The study sample comprised all occasions of a delivery (single or multiple birth) at Pattani hospital during the nine-year period. Mothers delivered triplets (4 occasions) and referral mothers from other hospitals to Pattani hospital (3,433 occasions) were excluded, because such referral is an intervening variable associated with the twin outcome. This sample included 221 mothers who delivered twins and 22,464 who delivered singletons on separate occasions.

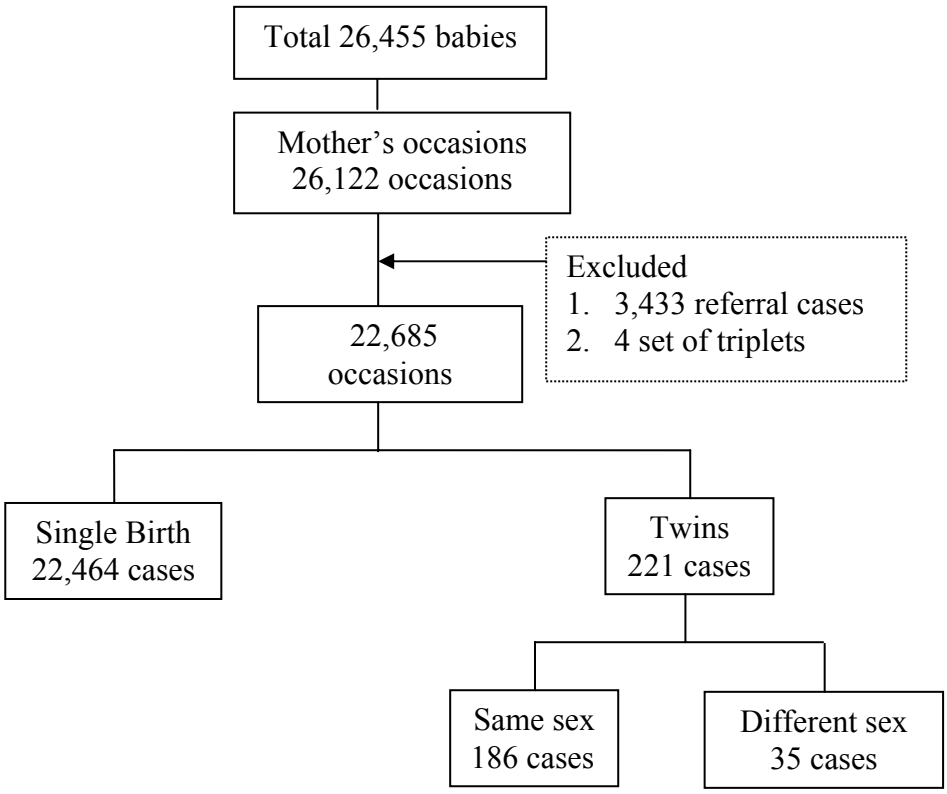

Figure 2.1: Diagram of mothers delivered in Pattani Hospital

The following diagram shows path diagram for the variables of interest.

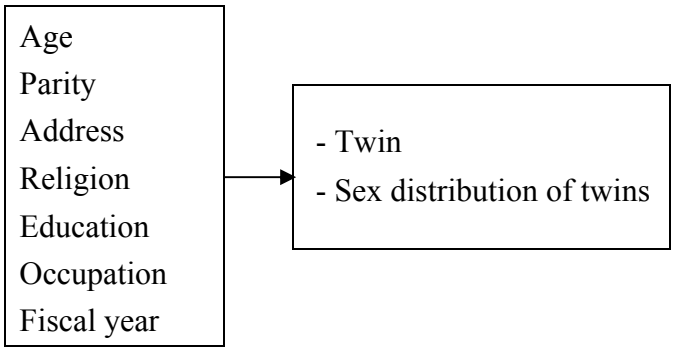

Figure 2.2: Path diagram showing variables in the study

The path diagram includes the variables used when twin status was taken as the outcome and sex distribution of the twins also was taken as the outcome.

2.3. Variables in the study

2.3.1 Outcome

We focus on the outcome variables: (a) the prevalence of twins among births, we divided twin status into 2 groups: twin and not twin (b) the sex distribution of twins comprises 2 groups: same-sex and opposite sex.

2.3.2 Determinants

The demographic factors of mother, namely age of mother, parity, address, religion, education, occupation and fiscal year.

The ethnicity of residents of Pattani province is mostly Jawi Malay but also includes Chinese and Thai and persons of mixed ethnicity. Since ethnicity is not recorded in the hospital records (and residents of mixed ethnic heritage are difficult to classify), the mother's sub-district of residence provides the most useful surrogate for ethnicity. Age was defined as the age of the mother at delivery at Pattani hospital during the

period, classified by 5-year group (less than 20, 20-24, 25-29, 30-34, 35+), and parity was similarly defined into three groups (0, 1 and 2+). Address was classified into 15 regions comprising aggregates of neighboring sub-districts of Pattani City or other districts of Pattani province or other provinces as follows: 1: Bana, 2: City East, 3: City South, 4: City Centre, 5: Sabarang, 6: Anuru, 7: Rusamilae, 8: Yaring, 9: Nong Chik, 10: Khok Pho and Mae Lan, 11: Pattani East, 12: Yarang, 13: Pattani South, 14: Songkla, 15: Narathiwat, Yala, not stated. Religion was classified as Islamic or other. Education was classified into four groups as 1: primary or less, 2: secondary, 3: tertiary or other, 4: not stated. Occupation was classified as 1: housewife, 2: gardener or farmer, 3: worker or in business, 4: government officer or other or not stated occupation. The fiscal year was defined as between 1 October in the preceding year and 30 September in the current year.

2.4 Data management

All ambiguous data such as overlapping, mislaid, wrong coding and duplicated record were cleaned. All statistical analysis was undertaken by using R program (R development core team, 2007).

2.5 Graphical Methods

The graphical methods are presented in the following steps.

Odds Ratio plot

Graphs of odds ratios and 95% confidence intervals can be used to present the associations between two nominal categorical variables. The association between an outcome variable and the determinants of interest were investigated by an odds ratio.

The graph of an odds ratio includes a 95% confidence interval. The confidence interval is graphed as a horizontal line containing a dot denoting the estimated odds ratio.

2.6 Statistical Methods

From the schematic diagram (figure2.2), the model specifies sex distribution of twins and twin status as outcome. This study focuses on the association between outcomes and the demographic of the mother. The prevalence of twins among all conceptions was calculated.

Descriptive Statistics

The variables of interest were summarized by using percentages. The determinants are nominal categorical variables which were described by using percentages

Statistical analysis about prevalence of twins

A. Prevalence of twins

The twinning rate was calculated by taking the number of twins divided by number of pregnancies and expressed per 1000 pregnancies. The prevalence of twins was calculated by

$$\text{Prevalence rate per } 1000 = \frac{\text{Number of twin in specified time}}{\text{Number of pregnancies at that specified time}} \times 1000 \qquad (2.1)$$

B. Estimated the number of dizygotic and monozygotic

The frequency of dizygotic (DZ) pairs and monozygotic (MZ) were estimated by Weinberg's method. Dizygotic twins are different sex and same sex. All monozygotic

twins are same sex. The Weinberg rule states that among dizygotic twins, there are equal numbers of same sex as of different sex. Thus the calculation of the frequency of monozygotic

Monozygotic pairs = total twins – estimated dizygotic (2.2)

And the calculation of the frequency of dizygotic

Dizygotic pairs = $2 \times$ (opposite sex pairs). (2.3)

Univariate Analysis

Pearson's Chi-squared test and 95% confidence intervals for odds ratio were used to assess the association between the determinant variables and the outcome of this study. The formulas of contingency tables (McNeil, 1998b) are as follows (X is the determinant of interest, Y is sex of twin and twin status)

A. 2 × 2 table

X is the determinant and Y is the outcome. The odds ratio is a measure of the strength of an association between two binary variables (i.e., in which both the outcome and the determinant are dichotomous) (McNeil, 1998b). To illustrate the definition of the odds ratio, assuming the variables takes values 1 and 0, a two-by-two table is constructed as follows.

		Y	
		1	0
X	1	a	b
	0	c	d

n = a + b + c + d

The ratio of these odds is referred to as the odds ratio (McNeil, 1998b). Thus the estimate the odds ratio is

$$\text{OR} = \frac{ad}{bc} \tag{2.4}$$

One method of testing the null hypothesis of no association between the determinant and the outcome is to use the z-statistics z=ln(*OR*) / *SE*, where *SE* is the standard error of the natural logarithm of the odds ratio (McNeil, 1996). An asymptotic formula for this standard error is given by

$$SE\ (ln\ OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{2.5}$$

A 95% confidence interval for the population odds ratio is thus

$$\text{OR} \times \exp\ (\pm\ 1.96\ \text{SE}\ [\ \ln\ \text{OR}]) \tag{2.6}$$

The more conventional method of testing the null hypothesis is based on Pearson's chi-square statistic, defined as

$$\chi^2 = \frac{(ad - bc)^2\, n}{(a+b)(c+d)(a+c)(b+d)} \tag{2.7}$$

The p-value is the probability that a chi-squared distribution with 1 degree of freedom exceeds this statistic.

Logistic Regression

Logistic regression analysis was used for modeling the association between several determinant variables and sex of twins, twin status. Logistic regression is a method of analysis that gives a particularly simple presentation for the logarithm of the odds ratio describing the association of a binary outcome with factors, and when fitted to

data involving a nominal outcome and multiple determinants, it automatically provides estimation of odds ratio and confidence intervals for specific combinations of the determinants (McNeil, 1996). For a set of predictor variables x_1, x_2,…, x_k and A binary outcome *Y* the logistic regression model takes the form:

$$\ln\left(\frac{p}{1-p}\right) = a + \sum_{i=1}^{k} \beta_i x_i \qquad (2.8)$$

where *p* denotes the probability of occurrence of the specified outcome. The probability of the outcome $Y = 1$ can be expressed as

$$P[Y=1] = \frac{\exp(a + \sum_{i=1}^{r} \beta_i x_i)}{1 + \exp(a + \sum_{i=1}^{p} \beta_i x_i)} \qquad (2.9)$$

Using the logistic regression model for the data arising from a two-by-two, we suppose $x_i = 1$ or 0, that is, the values of determinant *X* are taken to be 1 (exposure) and 0 (no exposure). Thus the logistic regression model can be written as

$$\ln\left\{\frac{P(Y=1/X=1)}{1-p(Y=1/X=1)}\right\} = \alpha + \beta x\,, \qquad (2.10)$$

$$\ln\left\{\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)}\right\} = \alpha \qquad (2.11)$$

The equations (2.10) and (2.11) actually are the (natural) logarithms of the odds for the outcome given the exposure ($x = 1$) and non-exposure (x=0), respectively. After exponentiation each equation, the odds for the exposed and non-exposed groups can be written as $\exp(\alpha + \beta)$ and $\exp(\alpha)$, respectively. The odds ratio is therefore obtained from the simple formula

$$OR = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = \exp(\beta) \qquad (2.12)$$

Comparing models

We have a logistic model, and we want to see if we can drop a subset of variables from the model. In other words, we want to see if a sub model of the original logistic model is adequate. Let L_{sub} and L_{full} represent the maximum values of the likelihood function under the sub model and under the full model respectively. Standard likelihood theory indicates that if the sub model is adequate, the difference $2\log L_{full} - 2\log L_{sub}$ will have a distribution that is approximately χ^2_d where *d* is number of variables dropped. This difference can be expressed as a difference of deviances:

$$2\log L_{full} - 2\log L_{sub} = (2\log L_{max} - 2\log L_{sub}) - (2\log L_{max} - 2\log L_{full}) \quad (2.13)$$

$$= \text{deviance of sub model} - \text{deviance of full model}$$

This difference represents the increase in the deviance when we drop the *d* terms from the model. While this difference will always be positive, if the increase is small then dropping the extra terms will not increase the deviance by very much and so the variables can be dropped in the interests of getting a simpler model. The difference in the deviance has approximately a χ^2_d distribution if the dropped variables are not needed in the model. Thus we can calculate a *p-value* to test the hypothesis that the dropped variables are not needed by comparing this difference in the deviance to a χ^2_d distribution. A small *p-value* provides evidence against the sub model-i.e. a small *p-value* indicates that we should not drop all *d* of the variables from the model.