# Chapter 2

# Methodology

This chapter describes the methodology including an overview of the statistical methods for data analysis associated to the statistical models. Graphical and statistical analyses were carried out using R (R Development Core Team, 2010).

This presents statistical model used in the two studies contained in Chapter 3. These methods include logistic regression model, log-linear regression model, Poisson regression model, negative binomial regression model.

## 2.1 Data sources and data management

The first study, the data include hospital discharge database information routinely reported to the National Health Security Office (NHSO) in the Ministry of Public Health during the 8 fiscal years from October 1999 to September 2007. The focus was on 26 provinces in the whole Central Region, comprising 309 hospitals. We created a secondary data file, kept in MySQL database. Data cleaning was undertaken for correct coding and dealing with missing values by using phpMyAdmin and WebStat. Data were converted to a flat-file format as text file for calculating descriptive statistics and statistical modeling.

The second study used the data on terrorism and violence in Southern Thailand (Pattani, Yala, Narathiwat and four eastern districts of Songkla province) were recorded by the Deep South Coordination Centre (DSCC) from January 1, 2004 for six years, until December 31, 2009. To calculate the incidence rates, using the population were obtained from the 2000 population and housing census of Thailand.

**2.2 Study variables**

According to the first study, to consider principal diagnosis according to International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10), age, gender, the hospital size and geographic region as predictors for LOS. The diseases were classified into 9 groups: injuries, cerebrovascular disease (CVD), digestive system disease, infectious disease, respiratory disease, genito-urinary disease (GUD), respiratory infection, malignant neoplasms (cancer), and other diseases. Age was divided into three groups: < 60 years, 60-74 years and 75 years or more. Hospital size was classified into three groups (small: 60 or fewer, gender, dividing age into three groups: 0-59 years, 60-74 years and 75 and over. The 26 provinces of Central Thailand were grouped into 6 geographic regions compressing provinces as follows:

i. North : Nontaburi, Pathumtani, Ayuthaya, Aungthong, Lopburi and Singburi,
ii. Northwest: Chainat, Kanjanaburi, Suphanburi and Nachornpatom
iii. Centre: Bangkok
iv. East: Saraburi, Chachengtrao, Prachinburi, Nachornnayok and Srakaeo
v. Southeast: Samutprakarn, Chonburi, Rayong, Chuntaburi and Trad
vi. Southwest: Rajburi, Samutsakhorn, Samutsongkhram, Pechaburi and Prajubkerekun.

Except for cancer, which had longer LOS, the median LOS in each disease group varied from 2 - 6 days. We classified hospital LOS into binary outcomes: "less than 7 days" and "7 days or more (adverse outcome)".

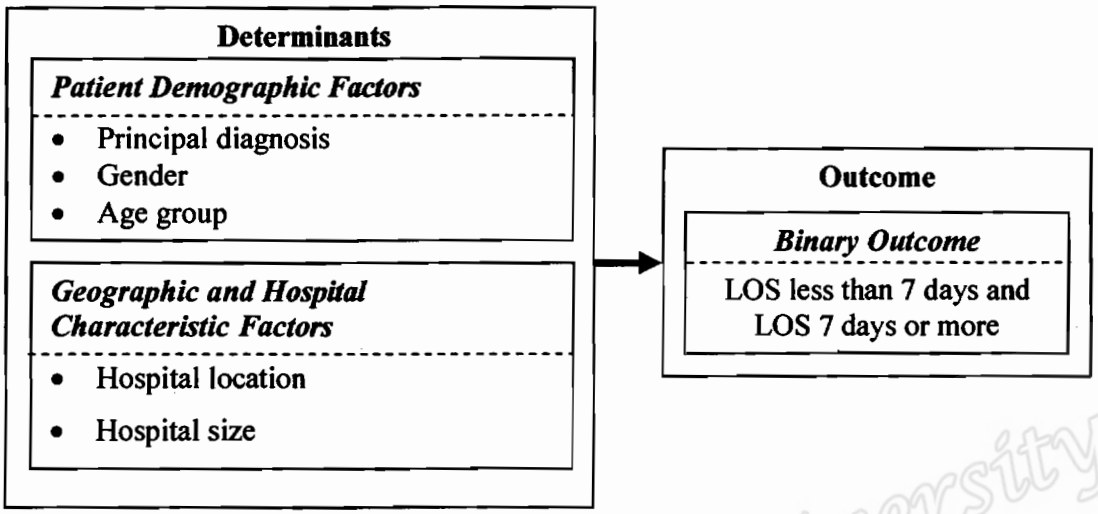
In the second study, incidence rates per 100,000 populations for Muslim resident civilian victims of terrorism events set as the outcome of interest. Factors classified by gender,

age group (<25, 25-44 and 45 or more), district of residence and year (six years from 2004 to 2009 inclusive). The target area comprises 37 districts in southern Thailand. Since the populations in these districts differ substantially, also initially created 23 regions with more equal populations by aggregating some adjoining districts in the same province as listed in Table 2.1. Two complex categorical factors were created: (1) gender and age group (6 levels); (2) region and year (138 levels), used to fit the models with two factors for Muslim victims.

| Province | RegionID: Districts | Population | |
|---|---|---|---|
| | | Muslim | Total |
| Songkla | 1: Chana/Thepha | 94,178 | 156,799 |
| | 2: SabaYoi/Na Thawi | 48,271 | 110,507 |
| Pattani | 3: Mueang Pattani | 67,149 | 104,145 |
| | 4: Kok Pho/Mae Lan | 40,816 | 75,628 |
| | 5: Nong Chik/Mayo/Kapho | 61,305 | 70,118 |
| | 6: Yaring | 79,051 | 81,495 |
| | 7: Panare/Sai Buri/Mai Kaen | 88,471 | 108,188 |
| | 8: ThungYang Dang | 69,745 | 73,545 |
| | 9: Yarang | 73,919 | 78,740 |
| Yala | 10: Mueang Yala | 79,343 | 154,634 |
| | 11: Betong/Than To | 31,487 | 68,193 |
| | 12: Raman | 54,451 | 62,756 |
| | 13: Yaha/Kabang/Krong Pinang | 50,522 | 56,546 |
| | 14: Bannang Sata | 69,892 | 73,408 |
| Narathiwat | 15: Mueang Narathiwat | 72,665 | 104,615 |
| | 16: Tak Bai | 45,781 | 61,157 |
| | 17: Bacho/Yi-ngo | 82,424 | 85,225 |
| | 18: Rueso | 53,333 | 59,108 |
| | 19: Rangae | 69,530 | 80,550 |
| | 20: SiSakon/Chanae | 50,075 | 54,039 |
| | 21: Sukirin/Waeng | 52,141 | 63,765 |
| | 22: Su-ngaiPadi/Cho-airong | 75,688 | 89,251 |
| | 23: Su-ngaiKolok | 41,317 | 64,640 |
| | Total Muslim Population | 1,451,554 | 1,937,052 |

*Table 2.1: Regions used in analysis of victims of terrorism in southern Thailand*

*Study 1*: Length of Stay of Patients Dying in Central Region Hospitals in Thailand.




*Study 2*: Muslim Victims of Terrorism Violence in Southern Thailand.

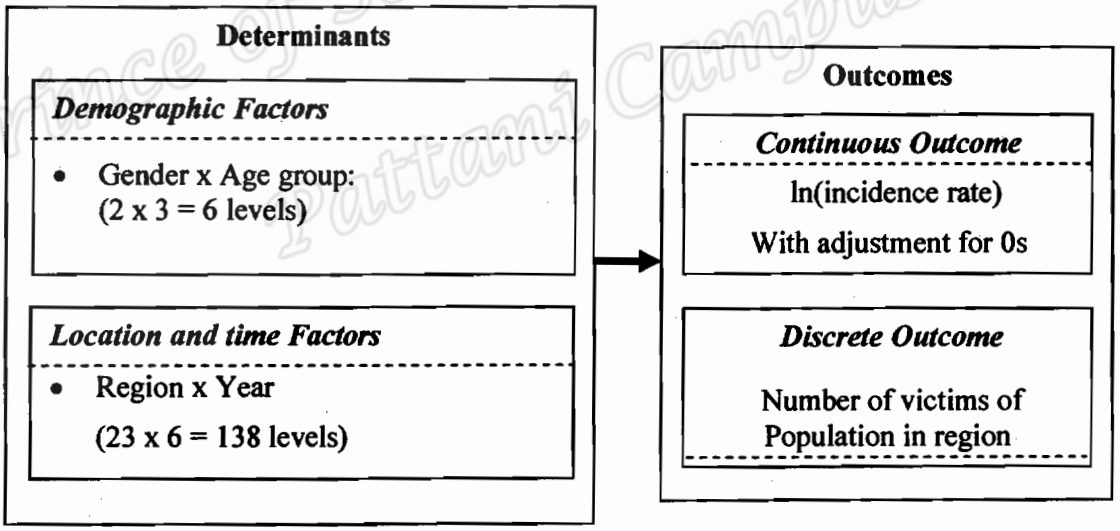



*Figure 2.1: Path diagrams for variables used in the two studies*

## 2.3 Statistical methods

*Univariate analysis*

Pearson's chi-square test is used to assess the association between the determinant variables and the outcome of this study.

*Pearson's chi-squared test*

Pearson's chi-squared statistic for independence (i.e., no association) is defined as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (2.3.1)$$

where $O_{ij}$ is the observed count in category $i$ of determinant and category j of the outcome, and $E_{ij}$ is the corresponding expected count, defined as before by dividing the product of the marginal totals by the overall total sample size, that is

$$E_{ij} = \frac{\sum_{j=1}^{c} O_{ij} \sum_{i=1}^{r} O_{ij}}{n} \qquad (2.3.2)$$

When the null hypothesis of independence is true, the right-hand side of Equation 2.1 has a chi-squared distribution with $(r-1)(c-1)$ degree of freedom (McNeil, 2006).

*Logistic regression model*

Logistic regression assumes binary outcomes. These refer to event that either happen or doesn't happen, so they comprise factor variables with two levels.

For the first study, while LOS was treated as a binary variable with patients staying in hospital at least 7 days as the outcome of interest. This outcome can be code as 0 or 1. Logistic regression analyses (Hosmer & Lemeshow, 2000; Kleinbaum & Klein, 2002) were performed to determine variables associated with LOS using the additive model (Chongsuvivatwong, 2008):

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \qquad (2.3.3)$$

In this model $\beta_0$ is the intercept and the terms $X_1$, $X_2$, $X_3$ , $X_4$ and $X_5$ are factors denoting gender, age group, disease, hospital size and geographic region, respectively. To avoid over-specification of parameters, each set of coefficients was constrained to have a mean equal to 0.

To calculate the proportion of LOS for each factor after adjusting for the effects of the other factors, equation (1) was used with the terms associated with the other factors replaced by constants, chosen to make the sum of the expected numbers of LOS as based on the model equal to the number observed (Kongchouy & Sampantarak, 2010).

*Weighted Sum Contrasts*

Sum contrasts (Venables and Ripley, 2002; Tongkumchum and McNeil, 2009) are used to obtain confidence intervals for comparing adjusted proportion within each factor with the overall mean. An advantage of these confidence intervals is that they provide a simple criterion for classifying levels of a factor into three groups according to whether each corresponding confidence interval exceeds, crosses, or is below the overall mean. These weighted sum contrasts provide standard errors for the differences between each factor level and their overall mean.

*Log-linear regression model*

Linear regression (see, for example, Cook and Weisberg, 1999) is a statistical method widely used to model the association between a continuous outcome and a set of fixed determinants. The model expresses the outcome variable as an additive function of the

determinants. For example, if there are two categorical determinants with levels indexed by subscripts $i$ and $j$, the model takes the form

$$Y_{ij} = \mu + \alpha_i + \beta_j \qquad (2.3.4)$$

In this case the number of parameters is $r+c-1$ where $r$ and $c$ are the number of levels of the factors $\alpha$ and $\beta$, respectively, thus requiring two constraints, taken as $\Sigma\alpha_i = 0$ and $\Sigma\beta_j = 0$ so that $\mu$ encapsulates the average of $Y$. We also assume that the errors are independent and normally distributed with mean 0 and constant standard deviation. The model may be fitted to the observations $y_{ij}$ by least squares, giving estimates and confidence intervals for the parameters. Equation (1) generalizes straightforwardly to any specified number of categorical determinants.

This method also applies to data that need to be transformed to satisfy the normality assumption, by first applying the method to the transformed data and then rescaling the result to ensure that the overall means of the untransformed data are the same before and after adjustment. It also extends straightforwardly to any number of covariate factors.

*Poisson regression and Negative binomial regression model*

The Poisson generalized linear model is widely used for modeling event counts in incidence rates (see, for example, Crawley, 2005). For two additive factors as in the linear model given by equation (1), if $P_{ij}$ is the population denominator, the expected value of the cell count $N_{ij}$ is expressed as

$$E[N_{ij}] = P_{ij} \exp(\mu + \alpha_i + \beta_j) \qquad (2.3.5)$$

However, the Poisson model often does not fit incidence data in practice because it assumes that the variance is equal to the mean, and in many situations the variance is

substantially greater than the mean (see, for example, Jansakul and Hinde, 2004; Kaewsompak et al, 2005; Paul and Saha, 2007; Kongchouy et al, 2010). The standard negative binomial GLM is a generalization of the Poisson model with the same mean $\lambda$, but the variance is $\lambda(1+\lambda/\theta)$ where $\theta > 0$ (see, for example, Chapter 7 of Venables and Ripley, 2002). This over-dispersion is often the result of clustering (see, for example, Demidenko, 2007).

By analogy with the method used for means based on the linear regression model, we define the adjusted incidence rate for level $j$ of factor $\beta$ as $\exp(\hat{\beta}_j + c)$, where the constant $c$ is chosen to ensure that the total number of adverse events based on the fitted model matches the number observed, that is,

$$\sum n_{ij} = \sum P_{ij} \exp(\hat{\beta}_j + c) \qquad (2.3.6)$$

## 2.4 Graphical display

Graphical displays are more useful for seeing and understanding such as geographic pattern by thematic map, model fitted by residual plot and comparing factors with adjusted for covariates by 95 % confidence interval plot.

*Thematic Map*

A thematic map is a type of map that uses different colors or shades to graphically display information about the underlying data representing estimated values of a variable at different locations on the map. The thematic map using data in regions might show one region in dark red to indicate that the region has high values, while showing another region in very pale red to indicate that the region has low values. A range map is a type of thematic map that displays data according to ranges set by the

users. The ranges are shaded using colors or patterns. These types of maps are used to show the geographical distribution of the adverse outcome and to identify areas of high risk.

*Histograms and bar charts*

A histogram shows the frequency distribution of a sample of numerical data. The data are grouped into intervals called bins, usually of equal width, and the histogram comprises a set of vertical bars stacked along a horizontal axis, where the height of a bar corresponds to the number of data values in the interval.

When graphing categorical data, we use bar charts instead of histograms to display their distribution.

*Residual plot*

A Poisson regression model can be fit to the data that is seen in a plot of deviance residuals, which measure the likelihood of the data when the fitted model is valid.

If the model is correct, these residuals should follow a straight line corresponding to normal quantiles, like residuals from linear regression models. Because of the outcome variable is a linear function of the parameters a, $b_1$, $b_2$, …, $b_p$. It is assumed that the errors from the model are independent and normally distributed with constant variance. This assumption is assessed by viewing a plot of the residuals against normal quantiles. If the normality assumption is valid, the points will follow a line corresponding to the normal quantiles.

The deviance itself should have a chi-squared distribution with degrees of freedom (*df*) equal to the number of residual *df* for the model. However, if the sample size is large the p-value for this chi-squared test will be very small even when the fitted

model is acceptable. In practice, the ratio of deviance to *df* should not be substantially greater than 1.

*Confidence interval plot*

A confidence interval is an interval, based on a sample of data from a population that contains the true value of a population parameter with specified probability (usually 0.95 = 95%).

A population parameter is a fixed value that is assumed not to vary, such as the speed of light, the average May temperature in Hat Yai, the proportion of Muslim people in Songkla in 2010, the concentration of $CO_2$ in the atmosphere in 2011, and the risk that a 70 year old man in Thailand will die within one year.

The range of a confidence interval decreases as the size of the sample increases: if the sample size is doubled, the confidence interval decreases by the factor $1/\sqrt{2}$ (0.71), and if the sample size is quadrupled, the confidence interval is halved.
Confidence intervals for parameters such as means, proportions and incidence rates (and differences between them) may be calculated using formulas based on statistical theory.

The 95% confidences intervals can be plot for comparing the proportions or percentage of an adverse outcome. When adjustments for covariates are required, the confidence interval with weighted sum contrasts can be use to show the pattern of the proportions or percentage for each factor. These confidence intervals are based on standard errors for the differences between each factor level and their overall mean.