Chapter 2

Methodology

This chapter describes methods used in the study including study design, data collection and management, path diagram, variables, and statistical methods.

2.1 Study design

The retrospective 10-years analysis death rates for children under-five years old from 2000 to 2009 were carried out according to gender and public health area.

2.2 Study groups

The study group comprises all under-five year death records in the Thai DR database from 2000 to 2009. Since the DR data are misclassified causes of deaths, a supplementary data from the VA survey in 2005 were also used.

2.3 Data collection and management

Death registration data

The numbers of children under-five deaths are based on deaths certificates obtained from the Bureau of Policy and Strategy, Ministry of Public Health, Thailand from January 2000 to December 2009. These data are classified by age, gender, place of death (province), year, and cause of death for principal diagnosis, which based on the International Classification of Disease (ICD-10).

The total number of all-cause deaths was 3,794,821 deaths. The deaths data were recorded as a text file. Error checking was performed in order to find wrong codes, missing values, and extreme values. All of the errors were fixed before analyzing the data. Out of 3,794,821 all-cause record, missing age was found for 0.24%. All of these missing ages were grouped into aged 80 years. Unknown province of death was found for 69,498 (1.80%) mostly in the age less than 1 year and they were omitted. The DR reported under-five deaths with 84,227 records were used for analysis.

Causes of death in DR database were judged as low quality due to high percentage of ill-defined and misclassified cause of deaths. To further investigate under-five deaths by cause, the DR reported cause of deaths were reallocated into more credible categories based on the analysis of the 2005 verbal autopsy (VA) data.

Verbal autopsy data

The VA study assessed cause of death from a sample of 9,644 cases from DR database in 28 districts of nine provinces. The nine provinces are Bangkok, Supan Buri, Nakon Nayok, Ubon Ratchathani, Loei, ChiangRai, Chumpon, and Songkhla (Rao et al 2010, Pattaraarchachai et al. 2010, Polprasert et al. 2010, Porapakkham et al. 2010). The VA study gave a data table with 5 fields: (a) the deceased person's province; (b) the person's gender and age; (c) the ICD-10 code reported on the death certificate; (d) the location of death (in hospital or outside hospital); (e) the VA-assessed ICD-10 code.

The VA study team separated results by field (d), grouped fields (c) and (e) into the 22 leading causes of death including a single group of perinatal and congenital (ICD-10 block P and Q) for each location, and thus found inflation factors for determining percentages of deaths in specific cause groups.

For under-five deaths, there were 149 cases with 59 due to perinatal originating conditions, 38 due to congenital malformation, and 52 due to other causes.

Estimating number of death by cause

Numbers of under-five deaths by cause were adjusted for misclassification using logistic regression model based on the VA data. It is simpler to separately fit logistic regression models to each of the three outcome cause groups.

This model formulates the logit of the probability that a person died from the selected cause (perinatal, congenital or all other causes) as an additive linear function of the three determinant factors as follows:

$$\log\left(\frac{p_{ijk}}{1-p_{ijk}}\right) = \mu + \alpha_i + \beta_j + \gamma_k \tag{2.1}$$

In this model μ is a constant and the terms α_i, β_j, and γ_k, refer to province, gender, and DR cause-location, respectively.

The province factor has nine levels corresponding to the nine provinces in the VA sample. The gender factor has two levels, boy and girl. The DR cause-location factor has four levels, corresponding to the two DR cause groups, where one group is the cause of interest and the other group aggregates death from all other causes and the two locations (in or outside hospital).

A receiver operating characteristic (ROC) curve was used to diagnose how well a model predicts a binary outcome. The logistic regression model gave nine province coefficients. Those coefficients were used to extrapolate the other province coefficients of the remaining provinces outside the VA survey using spatial triangulation method. The estimated probability of perinatal originating conditions can be obtained using equation 2.1. The estimated probabilities of conginental and all other causes were obtained using the same procedures. Then, the estimated probabilities were applied to the DR data from 2000-2009. The estimated number of deaths were thus scaled to ensure that the total number of deaths estimated for each group matches all reported from 2000-2009. Finally, the numbers of deaths by cause were estimated for each gender and year. The area plot was used to show estimated perinatal originating conditions, congenital malformations, and other causes for each gender for each year during 2000-2009.

Population data

Projected populations at risk classified by province, gender, 5-year age group, and year from 2000 to 2009 were obtained from the Institute for Population and Social Science Research at Mahidol University.

2.4 Path diagram and variables

Path diagram

Path diagram shows relationship between the determinants and outcomes. The determinants are gender, year, and Public Health Area. The outcome is under-five death rate (per 100,000 population) for all-cause and perinatal originating conditions. The path diagram for this study is shown in Figure 2.1.

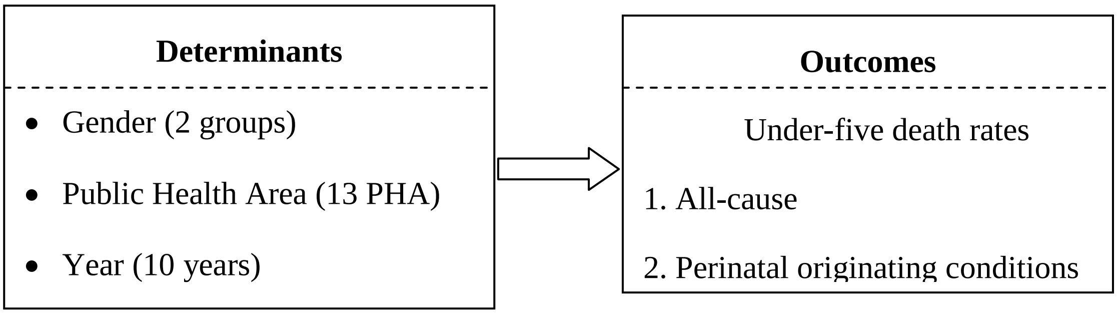

Figure 2.1: Path diagram

The map of Thailand's Public Health Areas (PHA) is shown in Figure 2.2 and provinces in each PHA are shown in Table 2.1.

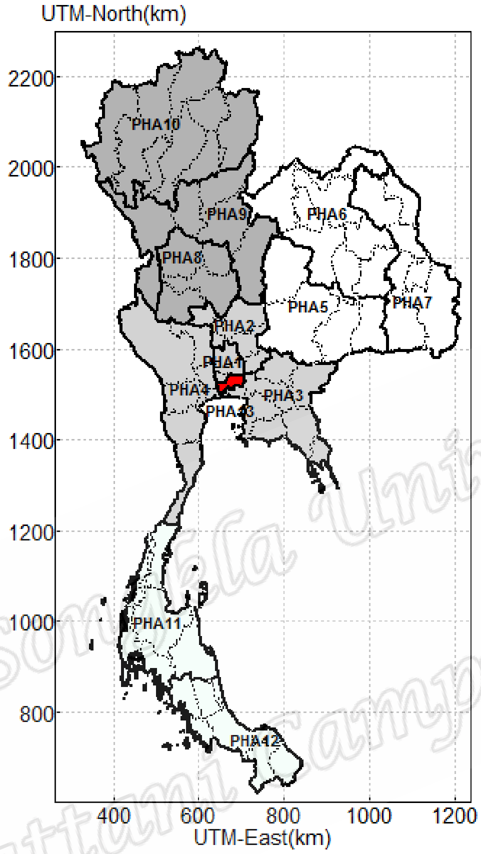

Figure 2.2: Public Health Area (PHA) map of Thailand

Table 2.1 shows provinces in each PHA. Each PHA comprises 3-8 provinces except PHA13.

Table 2.1: Labels for public health area

PHA	Province	PHA	Province
1	Nonthaburi	7	Si Sa Ket
	Patum Thani		Ubon Ratchathani
	Phra Nakhon Si Ayutthaya		Yasothon
2	AngThong		Amnat Charoen
	Lop Buri		Sakon Nakhon
	Sing Buri		Nakhon Phanom
	ChaiNat		Mudahan
	Saraburi	8	Nakhon Sawan
	Nakhon Nayok		Uthai Thani
3	Samut Prakan		Kamphaeng Phet
	Chon Buri		Phichit
	Rayong	9	Uttaradit
	Chanthaburi		Tak
	Trat		Sukhothai
	Chachoengsao		Phisanulok
	Prachin Buri		Phechabun
	SaKaeo	10	Chiang Mai
4	Ratchaburi		Lampang
	Kanchanaburi		Phrae
	Suphan Buri		Nan
	Nakhon Pathom		Phayao
	Samut Sakhon		Chiang Rai
	Samut Songkhram		Mae Hong Son
	Phetchaburi	11	Nakhon Si Thammarat
	Prachuap Khiri Khan		Krabi
5	Nakhon Ratchasima		Phangnga
	Buri Ram		Phuket
	Surin		Surat Thani
	Chaiyaphum		Ranong
6	Nong Bua Lam Phu		Chumphon
	Khon Kaen	12	Songkhla
	Udon Thani		Satun
	Loei		Trang
	Nong Khai		Phattalung
	Maha Sarakham		Pattani
	Roi Et		Yala
	Kalasin		Narathiwat
		13	Bangkok

2.5 Statistical methods

Death rate

Death rates (y_{ijt}) was computed as the number of deaths for children under-five years divided by the number of mid-year projected population and multiply by 100,000 population, given by

$$y_{ijt} = \frac{D_{ijt}}{P_{ijt}} \times K \qquad (2.1)$$

where D_{ijt} is the number of deaths for gender (i) (i= 1, 2), PHA (j) (j= 1, 2, 3, …,13), and year (t) (t= 2000, 2001, 2002, …, 2009). P_{ijt} is the corresponding population at middle year and K is a specified constant, here equal to 100,000.

Multiple linear regression

Multiple linear regression analysis was the appropriate statistical model for continuous outcome. The model for three factor determinants takes the form

$$y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_t \qquad (2.2)$$

where y_{ijt} is the death rates, μ is the constant, α_{i} is the effect of gender i, β_j is the effect of PHA j, and γ_t is the effect of years t. The model is fitted to the data using least squares, which minimize the sum of squares of the residuals. Multiple linear regression analysis based on three assumptions including the association is linear, the variability of the errors (in the outcome variable) is constant and these errors are normally distributed. When these assumptions are not met, the data may need to be transformed.

The rates generally have positively skewed distribution, so it is conventional to transform them by applying logarithms. The estimated additive model for logarithm of death rates is similar to model (2.2) and takes the form

$$\ln(y_{ijt}) = \mu + \alpha_i + \beta_j + \gamma_t \qquad (2.3)$$

Goodness of Fit

Coefficient of determination or R^2 is a common statistic tool for evaluating goodness of fit. It relates to the correlation coefficient (the coefficient of determination is the square of the correlation coefficient). It gives the percentage of total variation in the outcome variable explained by the regression line (Faraway, 2005). Other method for evaluating the model is the normal quantile plots. It is the plot of the quantiles of the residuals against the theorized quantiles. If the residuals come from a normal distribution the plot should be similar a straight line.

Sum Contrasts

Sum contrast (Venables and Ripley, 2002; Tongkumchum and McNeil, 2009) was used to obtain confidence intervals for comparing means within each factor with the overall mean. An advantage of these confidence intervals is that they provide a simple criterion for classifying levels of the factor into three groups according to whether each corresponding confidence interval exceeds, crosses, or is below the overall mean.

All statistical modeling and graphical displays were produced using R statistical software (R Development Core Team, 2011).