CHAPTER 2

Methodology

This chapter describes the overall research methodology used for estimating transport accident deaths and external causes of deaths in Thailand based on VA study and the patterns of transport accident mortality rates in Thailand. The data sources and data management are described in detail. Study population, sample, variables and statistical analysis are described separately into two parts (estimating the number of deaths and exploring patterns of mortality rates). Path diagrams are used to illustrate the overall conceptual framework.

2.1 Study design

A retrospective data analysis was performed to estimate and explore transport accident mortality patterns in Thailand.

2.2 Data sources

The main sources of data used for this study consist of three databases. First, the Thai 2005 VA study data with a sample of 9,644 deaths was obtained from the Bureau of Health Policy and Strategy, Thai Ministry of Public Health. These data consist of five fields: (a) the deceased person's gender and age; (b) the deceased person's province of resident registration; (c) the ICD-10 code reported on the death certificate; (d) the location of death (in or outside hospital); and (e) the VA-assessed ICD-10 code.

Second, the Bureau of Health Policy and Strategy, the Thai Ministry of Public Health, provided the VR data with all reported Thai deaths based on death certification

without autopsy. The data were classified by gender, age, 76 provinces, the location of death (hospital or outside hospital), the ICD-10 code reported and year 1996 to 2009. Both datasets contain the ICD-10 codes for transport accidents (V01-V99) (World Health Organization, 2004). The codes for Thailand's 76 provinces are shown in Table 2.1.

Table 2.1 Codes for 76 provinces separate by four regions in Thailand

Code	Province	Code	Province	Code	Province
Central Region		**Northeast Region**		**North Region**	
10	Bangkok	30	Nakhon Ratchasima	50	Chiang Mai
11	Samut Prakan	31	Buriram	51	Lamphun
12	Nonthaburi	32	Surin	52	Lampang
13	Pathum Thani	33	Si Sa Ket	53	Uttaradit
14	Ayutthaya	34	Ubon Ratchathani	54	Phrae
15	Ang Thong	35	Yasothon	55	Nan
16	Lopburi	36	Chaiyaphum	56	Phayao
17	Sing Buri	37	Amnat Charoen	57	Chiang Rai
18	Chai Nat	39	Nong Bua Lam Phu	58	Mae Hong Son
19	Saraburi	40	Khon Kaen	60	Nakhon Sawan
20	Chonburi	41	Udon Thani	61	Uthai Thani
21	Rayong	42	Loei	62	Kamphaeng Phet
22	Chanthaburi	43	NongKhai	63	Tak
23	Trat	44	Maha Sarakham	64	Sukhothai
24	Chachoengsao	45	Roi Et	65	Phitsanulok
25	Prachinburi	46	Kalasin	66	Phichit
26	Nakhon Nayok	47	Sakon Nakhon	67	Phetchabun
27	Sa Kaeo	48	Nakhon Phanom	**South Region**	
70	Ratchaburi	49	Mukdahan	80	Nakhon Si Thammarat
71	Kanchanaburi			81	Krabi
72	Suphan Buri			82	Phang Nga
73	Nakhon Pathom			83	Phuket

Table 2.1 (cont.)

Code	Province	Code	Province	Code	Province
74	Samut Sakhon			84	Surat Thani
75	Samut Songkhram			85	Ranong
76	Phetchaburi			86	Chumphon
77	Prachuap Khiri Khan			90	Songkhla
				91	Satun
				92	Trang
				93	Phattalung
				94	Pattani
				95	Yala
				96	Narathiwat

Third, the population during the years 1996 to 2009 were obtained from the 2000 population and housing census by the National Statistical Office of Thailand. The details of these three data sources are shown in Figure 2.1.

2.3 Data management

For the 9,644 VA deaths, the deceased person's age was grouped contiguous ages into eight 10-year age groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69 and 70 years and over. The deceased person's gender was combined with the age groups to produce 16 gender-age groups (two genders each with eight age groups) for reducing gender and age group interaction. This process was performed for other accidents except for suicide. As no children aged below 5 years died from suicide only 12 gender-age groups were analyzed 5-29, 30-39, 40-49, 50-59, 60-69 and 70 years and over.

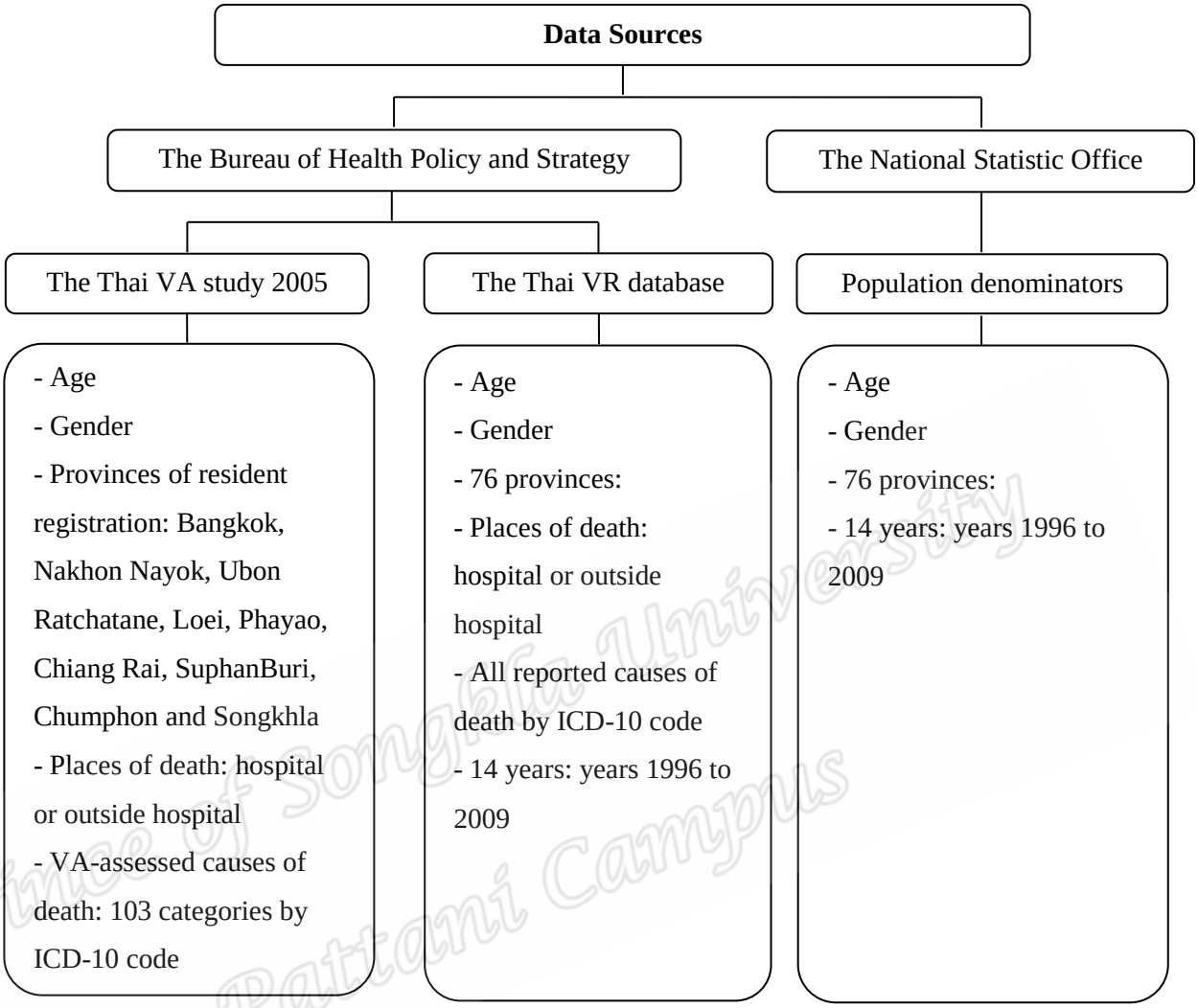

Figure 2.1 List of data sources for this study

In the 2005 VA study, causes of death were aggregated, reducing the ICD-10 Mortality Tabulation List 1 consisting of 103 cause categories into the 21 leading causes of death (Rao et al., 2010) according to mortality distribution. This thesis used the chapter-block classification of the ICD-10 codes to group causes of deaths for 22 major cause groups after consultation with the Epidemiology Unit, Faculty of Medicine, Prince of Songkla University. Each cause group had to be large enough for statistical analysis (at least 200 deaths in each group), except for septicemia (over-report), perinatal and congenital child death (confined to aged 0-4) and suicide as

groups deserving special attention. The size of these groups based on the VA counts ranged from 77 for septicemia (A40-41) to 1,076 for stroke (I60-69), and 546 were from transport accident (V01-V99), 341 were from other accidents (W00-W99, X00-X59) and 158 were from suicide (X60-X84) as shown in Table 2.2.

Table 2.2 VA counts for 22 major causes of deaths in Thailand, 2005: The VA study

No	Cause groups	ICD codes	VA data (cases)
1	Tuberculosis (TB)	A15-A19	195
2	Septicaemia	A40-A41	77
3	Human immunodeficiency virus disease (HIV/AIDS)	B20-B24	514
4	All other infectious and parasitic diseases (Other Infectious)	A00-A09, A20-A28, A30-A39, A42-A49, A50-A99, B00-B19, B25-B99	222
5	Malignant neoplasm of liver and intrahepatic bile ducts (Liver cancer)	C22	500
6	Malignant neoplasms of respiratory and intrathoracic organs (Lung Cancer)	C30-C39	320
7	Malignant neoplasms of digestive organs (Other Digestive)	C15-C21, C23-C26	290
8	Other neoplasms (Other Cancer)	C00-C14, C40-C97, D00-D48	704
9	Endocrine, nutritional and metabolic diseases (Endocrine)	E00-E90	649
10	Mental and behavioural disorders, and Diseases of the nervous system (Mental and Nervous)	F00-F99, G00-G99	228
11	Ischaemic heart diseases (Ischemic)	I20-I25	617
12	Cerebrovascular diseases (Stroke)	I60-I69	1,076
13	Other diseases of the circulatory system (Other CVD)	I00-I15, I26-I52, I70-I99	544
14	Diseases of the respiratory system (Respiratory)	J00-J99	802
15	Diseases of the digestive system (Digestive)	K00-K93	490
16	Diseases of the genitourinary system (GU)	N00-N99	412

Table 2.2 (cont.)

No	Cause groups	ICD codes	VA data (cases)
17	Pregnancy, Childbirth and the puerperium, Certain conditions originating in the perinatal period, and Congenital malformations, deformations and chromosomal abnormalities (Child Mortality)	P00-P96, Q00-Q99 (aged ≤ 5)	97
18	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified (Ill-defined)	R00-R99	502
19	Transport accidents	V01-V99	546
20	Other external causes of accidental injury (Other Accidents)	W00-W99,X00-X59	341
21	Intentional self-harm (Suicide)	X60-X84	158
22	All other causes (All Other)*	All other codes	360
	Total deaths		**9,644**

* All other: D50-D89, H00-H95, L00-L99, M00-M99, O00-O99, P00-Q99 age>5, X85-Y09, Y10-Y89

From 22 major cause groups, a cross-tabulation of the number of VR reported causes of deaths and corresponding VA-assessed causes of deaths found VR reported causes of transport accident deaths attributed to other VA-assessed causes of deaths (11 groups): “all other” (156 deaths), “other injuries” (64 deaths), “ill-defined” (33 deaths), “stroke” (13 deaths), “respiratory” (8 deaths), “mental and nervous” (5 deaths), “septicemia” (3 deaths), “other CVD” (2 deaths), “endocrine” (1 death), “digestive” (1 death) and “suicide” (1 death) as shown in Figure 2.2.

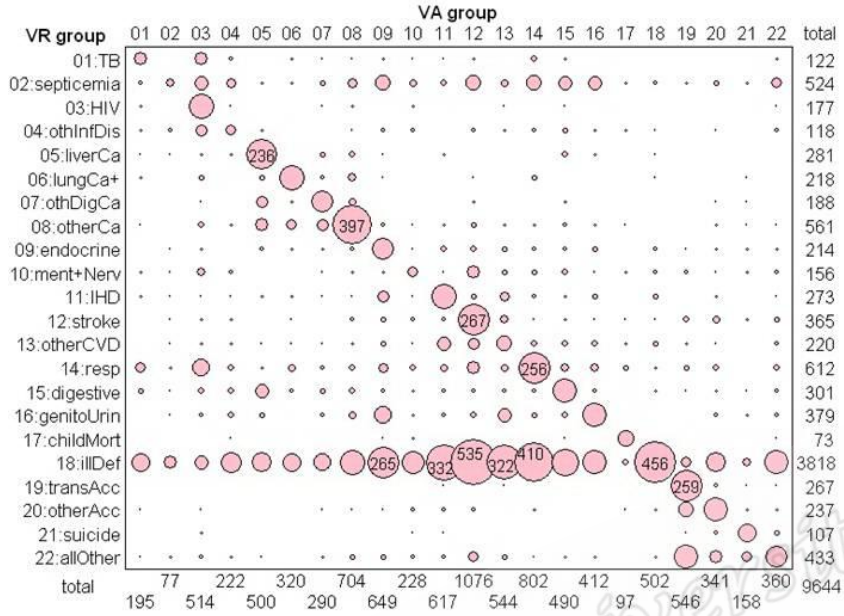

Figure 2.2 A cross-tabulation of the number of VR counts and corresponding VA counts from 22 major cause groups. Area of bubbles denote numbers of death

To select the best subset of predictors, each group should have sufficient cell counts to avoid zeros. Thus, the VR reported causes were arranged in seven groups consisting of the six most likely VR cause groups and a combined "other" group: "transport accident", "all other", "other accidents", "ill-defined", "stroke", "respiratory" and "other groups". These seven cause of death groups were then combined with the two locations (in or outside hospital) to produce 14 VR cause-location groups (two locations each with VR cause groups) as shown in Table 2.3.

Table 2.3 VR reported causes of transport accident deaths attributed to other VA-assessed causes of deaths

VR cause groups	VA cause groups							
	Transport accident	All other	Other accidents	Ill-defined	Stroke	Respi-ratory	Other groups*	Sum VR (cases)
Transport accident	259	1	4	0	0	0	3	267
All other	156	119	41	2	31	1	83	433
Other accidents	64	3	143	2	6	3	16	237
Ill-defined	33	143	97	456	535	410	2,144	3,818
Stroke	13	9	14	1	267	1	60	365
Respiratory	8	10	11	3	45	256	279	612
Other groups	13	75	31	38	192	131	3,432	3,912
Sum VA (cases)	546	360	341	502	1,076	802	6,017	9,644

*Other groups: septicaemia, endocrine, mental and nervous, other CVD, digestive, suicide

For "other accidents", VR reported causes of other deaths were attributed to other VA-assessed causes of deaths (16 groups): "ill-defined" (97 deaths), "all other" (41 deaths), "stroke" (14 deaths), "respiratory" (11 deaths), "septicemia" (8 deaths), "genitourinary" (5 deaths), "suicide" (5 deaths), "transport accident" (4 deaths), "digestive" (3 deaths), "other cancer" (2 deaths), "endocrine" (2 deaths), "IHD" (2 deaths), "other infectious diseases" (1 death), "other digestive cancer" (1 death), "mental and nervous" (1 death) and "other CVD" (1 death). The VR reported causes

were grouped into eight groups consist of the seven most likely VR cause groups and the combined "other" group: "other accidents", "ill-defined", "stroke", "respiratory", "septicemia", "genitourinary", "all other" and "other groups", and combined with the two locations (in or outside hospital) to produce 16 VR cause-location groups (two locations each with VR cause groups) as shown in Table 2.4.

Table 2.4 VR reported causes of other accident deaths attributed to other VA-assessed causes of deaths

VR cause groups	VA cause groups								
	Other Accidents	Ill-defined	Stroke	Respi-ratory	Septicemia	GU	All Other	Other Groups*	Sum VR (cases)
Other Accidents	143	2	6	3	0	0	3	80	237
Ill-defined	97	456	535	410	41	146	143	1,990	3,818
Stroke	14	1	267	1	1	1	9	71	365
Respiratory	11	3	45	256	2	20	10	265	612
Septicemia	8	6	60	63	20	52	32	283	524
GU	5	0	12	8	1	151	6	196	379
All other	41	2	31	1	2	2	119	235	433
Other Groups	22	32	120	60	10	40	38	3,590	3,912
Sum VA (cases)	341	502	1,076	802	77	412	360	6,074	9,644

*Other groups: other infectious, other digestive, other cancer, endocrine, mental and nervous, ischemic, other CVD, digestive, transport accident, suicide

For suicide, VR reported causes of other deaths were attributed to other VA-assessed causes of deaths (13 groups): “all other” (32 deaths), “ill-defined” (18 deaths), “endocrine” (3 deaths), “septicemia” (2 deaths), “mental and nervous” (2 deaths), “stroke” (2 deaths), “genitourinary” (2 deaths), “other accidents” (2 deaths), “lung cancer” (1 death), “other digestive cancer” (1 death), “other cancer” (1 death), “digestive” (1 death) and “transport accident” (1 death). The VR reported causes were grouped into four groups consisting of the three most likely VR cause groups and the combined “other” group: “suicide”, “ill-defined”, “all other” and “other groups”, and combined with the two locations (in or outside hospital) to produce eight VR cause-location groups (two locations each with VR cause groups) as shown in Table 2.5.

Table 2.5 VR reported causes of suicide deaths attributed to other VA-assessed causes of deaths

VR cause groups	VA cause groups				
	Suicide	Ill-defined	All other	Other groups	Sum VR (cases)
Suicide	90	0	7	10	107
Ill-defined	18	456	143	3,201	3,818
All other	32	2	119	280	433
Other groups	18	44	91	5,133	5,286
Sum VA (cases)	158	502	360	8,624	9,644

*Other groups: septicaemia, lung cancer, other digestives, other cancers, endocrine, mental and nervous, stroke, digestive, GU, transport accident, other accidents

For exploring trend pattern of transport accident mortality rates, the estimated number of transport accident deaths for the years 2004 to 2009 were used as they had the least errors in the VR data. The province factor was combined with the six years (2004,

2005, 2006, 2007, 2008 and 2009) to produce 456 province-year groups (six years each with provinces) for reducing province and year group interaction.

2.4 Statistical analyses for estimating transport accident deaths and external causes of deaths in Thailand based on VA study (part I and part II)

According to the VA data, the causes of deaths were estimated for identifying the misclassification of causes of deaths in VR data which were not estimated for undercounted of the target population. Thus, the capture-recapture method was not used to estimate the number of deaths in this thesis. For estimating the number of death from causes of death of interest (i.e. transport accident death, other accident death and suicide death) a model was created for each cause. The outcome is a binary outcome.

For the binary outcome, first, the outcome "the interesting cause of death" was coded 1 and "the non-interesting cause of death" was coded 0. Therefore, the outcomes of interest for three models were created including (1) transport accident death coded as 1 and other causes of deaths coded as 0 (2) other accident deaths coded as 1 and other causes of deaths coded as 0 and (3) suicide coded as 1 and other causes of deaths coded as 0. The determinants were province, gender-age group and VR cause-location groups. VR cause-location groups were grouped according to the most likely VR reported causes of deaths corresponding to the outcome. So, the VR cause-location group determinant was made up of different groups for each outcome. The logistic regression model was used to identify the association between the outcome and determinants.

Each model formulates the logit of the probability *P* that a person died from the interesting cause of death (i.e. transport accident, other accident and suicide) as an additive linear function of the three determinant factors as follows:

$$ln\left[\frac{P_{ijk}}{1-P_{ijk}}\right] = \mu + \alpha_i + \beta_j + \gamma_k \qquad (2.1)$$

In this model μ is a constant and the terms α_i, β_j and γ_k refer to province, gender-age group and VR cause-location group, respectively. The province factor has nine provinces in The VA study. The gender-age group factor depends on the age distribution of deaths from the selected outcome (i.e. 16 levels in transport accidents and other accidents, and 12 levels in suicide). Similarly, the VR cause-location factor depends on the number of such reported cause groups that affect the outcome cause group (i.e. 14 levels in transport accidents, 16 levels in other accidents and 8 levels in suicide).

The p-value for a factor in a regression model is Prob[χ^2>*D*], the tail area of a chi-squared distribution with *k*-1 degrees of freedom (*df*), where *k* is the number of levels and *D* is the reduction in deviance (a measure of lack of fit of the model) achieved by the factor. Thus, each model was assigned an appropriate constant to yield 546 transport accident deaths, 341 other accident deaths and 158 suicide deaths by each factor which was the same as the VA study.

Next, the 95% confidence intervals (CI) of adjusted percentages for each determinant are used for adjusting the outcome to reduce the effects of confounding bias arising from covariates associated with both the binary outcome and the factors (McNeil, 1996). These confidence intervals provide standard errors for the differences between

each factor level and their overall mean. Sum contrasts was used to obtain CI for comparing means/proportions percentage with the overall mean/proportion. A method for performing this strategy is described by Tongkumchum and McNeil (2009), Kongchouy and Sampantarak (2010).

Then, the ROC curve was used to validate the logistic regression model, how well a model predicts a binary outcome (Hosmer and Lemeshow, 2000; Westin, 2001). The interpretation of how well a model predicts a binary outcome is considered the area under the ROC curve (AUC), sensitivity (proportion of positive outcomes correctly predicted by the model) and false positive rate (proportion of all outcomes incorrectly predicted). This method is an important step in ensuring that the model is most likely to estimate the same number for the cause of death of interest, in agreement with the VA study number. Indicating the predicted outcome as 1 (the interesting cause of death) if $P \geq c$ (cutoff point) or 0 (other death) if $P < c$, it plots sensitivity against the false positive rate, as c varies. Then, this study chose an optimal cut-off point for predicting the same number of deaths as the VA study (546 transport accident deaths, 341 other accident deaths and 158 suicide deaths).

According to the VA study, the province factor comprises nine of Thailand's 76 provinces. Thus, a method is needed to interpolate coefficients for the other 67 provinces outside the VA study. The coefficients of the nine provinces in the logistic regression model with treatment contrasts were used for estimating them based on the latitude *(lat)* and longitude *(long)* of their central points. Triangles were used to line the nine VA provinces, where co-ordinates of Bangkok were used as a reference point. Thus triangle created areas covered each province, where its coefficient value

was obtained at the vertices of the triangle. For each triangle, values *(a, b, c)* are obtained by solving three equations as follows:

$$a + longP_1 \times b + latP_1 \times c = \beta_1 P_1 \qquad (2.2)$$

$$a + longP_2 \times b + latP_2 \times c = \beta_2 P_2 \qquad (2.3)$$

$$a + longP_3 \times b + latP_3 \times c = \beta_3 P_3 \qquad (2.4)$$

The coefficient for any province *j* within a triangle is now given by

$$a + longP_j \times b + latP_j \times c = \beta_j P_j \qquad (2.5)$$

In each triangle *P* is province and β is coefficient. Coefficients for provinces outside triangles were obtained similarly by extrapolation.

According to the triangulation method, three provinces of Southern Thailand were not covered. Then, we assumed each province coefficient value was the same as the nearby province (Songkhla province).

After creating the appropriate model, we assumed that the patterns of misreporting of deaths before and after year 2005 were the same as in 2005 when the VA study was undertaken as shown in Table 2.6.

Table 2.6 Percentage of VR reported of ill-defined deaths years 1996 to 2009

Years	VR-reported		
	Total deaths	Number of ill-defined deaths	Percent (%)
1996	341,136	121,414	35.6
1997	296,713	103,244	34.8
1998	305,559	116,351	38.1

Table 2.6 (cont.)

Years	VR-reported		
	Total deaths	Number of ill-defined deaths	Percent (%)
1999	358,420	150,014	41.9
2000	358,325	147,755	41.2
2001	368,233	140,181	38.1
2002	366,658	138,814	37.9
2003	367,705	123,068	33.5
2004	366,712	141,351	38.5
2005	393,354	150,524	38.3
2006	389,583	149,377	38.3
2007	393,116	150,309	38.2
2008	397,256	150,575	37.9
2009	393,877	149,394	37.9
Total	5,096,647	1,932,371	37.9

Thus, the logistic regression models were applied to estimate the number of deaths for each of the three interesting cause groups: transport accidents, other accidents and suicide from the VR database years 1996 to 2009.

2.5 Statistical analyses for exploring patterns of transport accident mortality rates in Thailand. (part III)

The patterns of transport accident mortality in each province in Thailand were explored for both severity levels and trend directions of transport accident mortality rates.

Poisson regression is an appropriate method for fitting models with count data (non-negative integer-values). It has been commonly used for modeling mortality rates and number of deaths in a specific population within a certain time (Rhodes and Freitas, 2004). Moreover, the data used here show no evidence of over-dispersion that followed the assumption of Poisson regression model. Thus, it was used to fit the transport accident mortality rates in Thailand for all gender-age groups and province-year groups during 2004 to 2009.The model is a simple generalized linear model based on the Poisson distribution, taking the additive form as follows:

$$log(\lambda_{jt}/P_{jt}) = \mu + \beta_j + \gamma_t \qquad (2.6)$$

For this model, λ_{jt} is the mean of the Poisson distribution giving the estimated number of transport accident deaths for 16 gender-age groups (j) (j=1, 2, 3,…, 16) and 456 province-year groups (t) (t=102004, 112004,…, 962009). P_{jt} is the corresponding population at risk in 100,000s and the terms β_j and γ_t represent gender-age group and province-year group, respectively. μ is a constant encapsulating the overall incidence. The model thus has 7,296 cells (16 x 456) corresponding to 16 gender-age groups combinations and 456 province-year groups.

The province-year coefficients from the Poisson regression model were used for dividing the transport accident mortality rate into three groups (low, medium and high).

Next, the trend directions of transport accident mortality rates were considered for classifying levels of each transport accident mortality rate group by the slope or regression coefficient. A simple linear regression model was used for fitting the

association between the province-year coefficients from the model and years as follows:

$$y = a + bx \qquad (2.7)$$

Here y is the coefficient of the province-year determinant of each province, a is the intercept, b is the slope or regression coefficient and *x* is years from 2004 to 2009. The slopes of the simple linear regression model were used for dividing trend direction of the transport accident mortality rates into nine region-year groups: (1) high mortality rate and fast decreasing trend (2) high mortality rate and slow decreasing trend (3) high mortality rate and slow increasing trend (4) medium mortality rate and fast decreasing trend (5) medium mortality rate and slow decreasing trend (6) low mortality rate and flat trend (7) low mortality rate and fast decreasing trend (8) low mortality rate and fast increasing trend and (9) low mortality rate and slow decreasing trend. A thematic map was used to compare differences in region-year group rates between geographical areas.

After that, the transport accident mortality rates of each region-year groups by gender-age group and year were modeled using the Poisson regression model as follows:

$$log(\lambda_{jt}/P_{jt}) = \mu + \beta_j + \gamma_t \qquad (2.8)$$

Here λ_{jt} is the mean of the Poisson distribution giving the estimated number of transport accident deaths for 16 gender-age groups (j) (j=1, 2, 3, …, 16) and six years (t) (t=2004, 2005, 2006, …,2009). P_{jt} is the corresponding population at risk in 100,000s and the terms β_j and γ_t represent gender-age group and year, respectively. μ is a constant encapsulating the overall incidence. The model of each region-year

groups thus has 96 cells (16 x 6) corresponding to 16 gender-age groups combinations and six years.

Finally, the confidence intervals from the sum contrasts was provided as a criterion for classifying levels of transport accident mortality rates of two factors (i.e. gender-age and region-year) into three groups, according to whether the confidence interval exceeds, crosses, or is below the overall mean.

2.6 Statistical software

All maps, graphs, data processing and manipulation, and statistical analysis were performed using R statistical software (R Core Team, 2013).