Chapter 2

Methodology

The present study used liver cancer mortality data from two sources. There were national vital registration (VR) data from Bureau of Health Policy and Strategy from 2000 to 2009 and the verbal autopsy (VA) study in 2005. The population denominator is from Institute of Population Studies, Mahidol University.

This chapter describes data comprising the national VR and the 2005 VA data in Sections 2.1 and 2.2. Section 2.3 describes statistical analysis.

2.1 The National Vital Registration Data

The numbers of deaths from liver cancer over the period 2000-2009 are from the national vital registration database. The vital data processing after 1996-present (United Nation 2011) are shown in Figure 2.1.

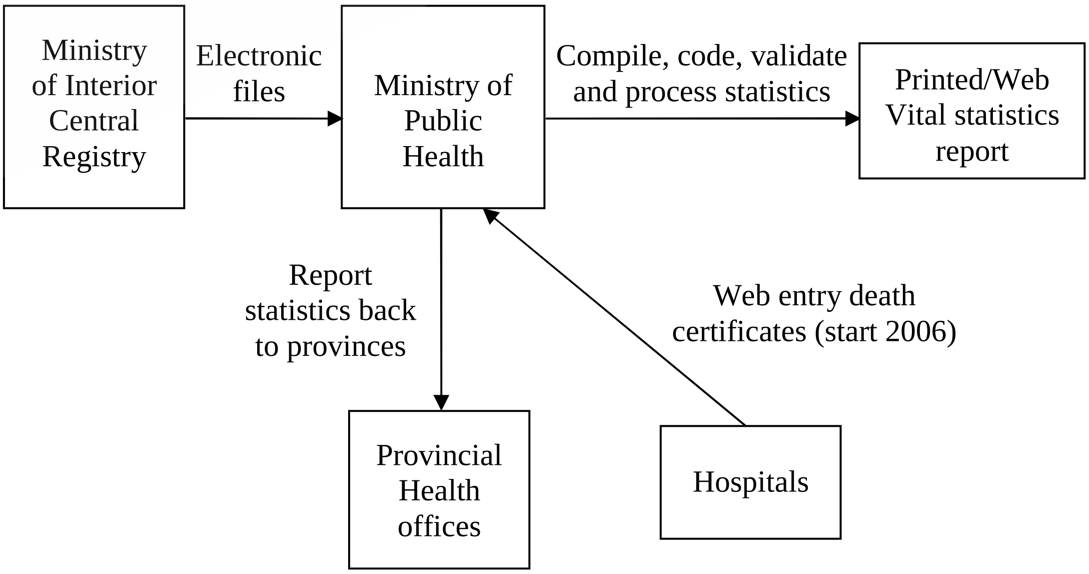
Ministry of Interior Central Registry
Electronic files
Ministry of Public Health
Compile, code, validate and process statistics
Printed/Web Vital statistics report
Report statistics back to provinces
Web entry death certificates (start 2006)
Provincial Health offices
Hospitals

Figure 2.1: Thai vital registration data processing

Cause of death classification used was coded according to the Tenth Revision of the International Classification of Diseases (ICD-10). The classification of liver cancer is shown in Table 2.1.

Table 2.1: The ICD-10 for liver cancer deaths

ICD-10	Description	ICD-10	Description
C22	Malignant neoplasm of liver and intrahepaticbile ducts Excludes: biliary tract, unspecified (C24.9) secondary malignant neoplasms of liver (C78.7)	C22.0	Liver cell Hepatocellular carcinoma Hepatoma
C22.1	Intrahepatic bile duct Cholangiocarcinoma	C22.2	Hepatoblastoma
C22.3	Angiosarcoma of liver Kupffer cell sarcoma	C22.4	Other sarcomas of the Liver
C22.7	Other specific types	C22.9	Liver, unspecified

The data set comprises the ICD-10 codes for principal diagnosis, other variables are gender, age, year, district and in/out hospital place of death.

Liver cancer is one of the 22 major cause groups. The 22 major cause groups were created from the chapter-block classification of ICD-10 codes based on the distribution of the VR reported and VA-assessed deaths. The cause group will be use as a factor determinant in the model to verify the reported cause of deaths in 2005. The 22 cause groups are shown in Table 2.2.

Table 2.2: Major cause groups

Cause group major categories	Cause group major categories
1: TB (A15-19)	12: Stroke (I60-69)
2: Septicemia (A40-41)	13: Other CVD (I)
3: HIV (B20-24)	14: Respiratory (J)
4: Other Infectious (A,B)	15: Digestive (K)
5: Liver Cancer (C22)	16: GenitoUrinary (N)
6: Lung Cancer+(C30-39)	17: Childbirth (OPQ)
7: Other Digest. (C15-26)	18: Ill-defined (R)
8: Other Cancer (C, D0-48)	19: Transport Acc. (V)
9: Endocrine (E)	20: Other Injury (W, X0-59)
10: Mental, Nervous (F, G)	21: Suicide (X60-84)
11: Ischemic (I20-25)	22: All other

2.2 The 2005 Verbal Autopsy Study

The VA study assessed cause of death from a sample of 9,644 cases (Rao et al 2010, Pattaraarchachai et al. 2010, Polprasert et al. 2010, Porapakkham et al. 2010), giving a data table with 5 fields: (a) the deceased person’s province: the nine provinces with sample sizes shown in Figure 2.2; (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.

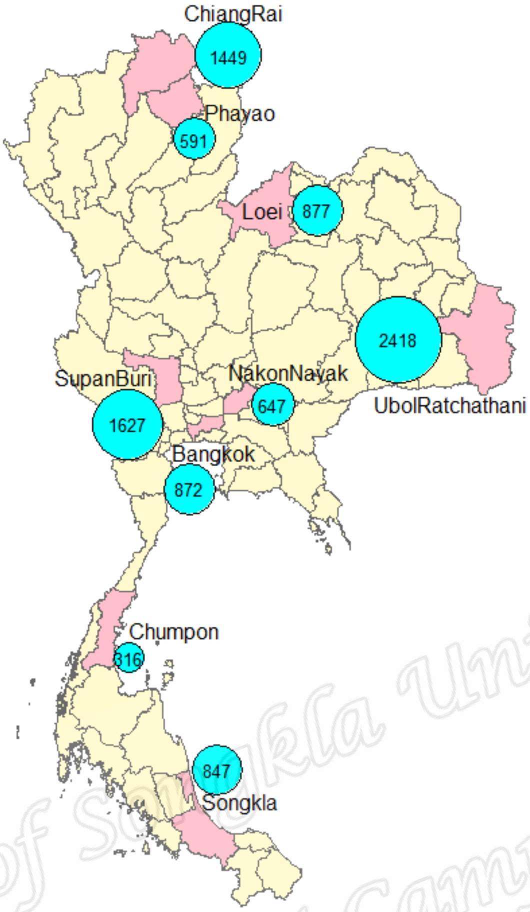

Figure 2.2: Nine sample provinces with sample size from the 2005 VA study

The VA study team separated results by field (d), grouped fields (c) and (e) into the 20 leading causes of death for each location, and thus found inflation factors for determining percentages of deaths in specific cause groups.

2.3 Statistical analysis

Target population

Target population is all reported Thai deaths from January 2000 to December 2009.

Sample

Sample is the verbal autopsies assessing true cause of death for 9,644 selected residents from nine provinces who died in 2005.

Path diagram

Path diagram for our study of liver cancer mortality is shown in Figure 2.3.

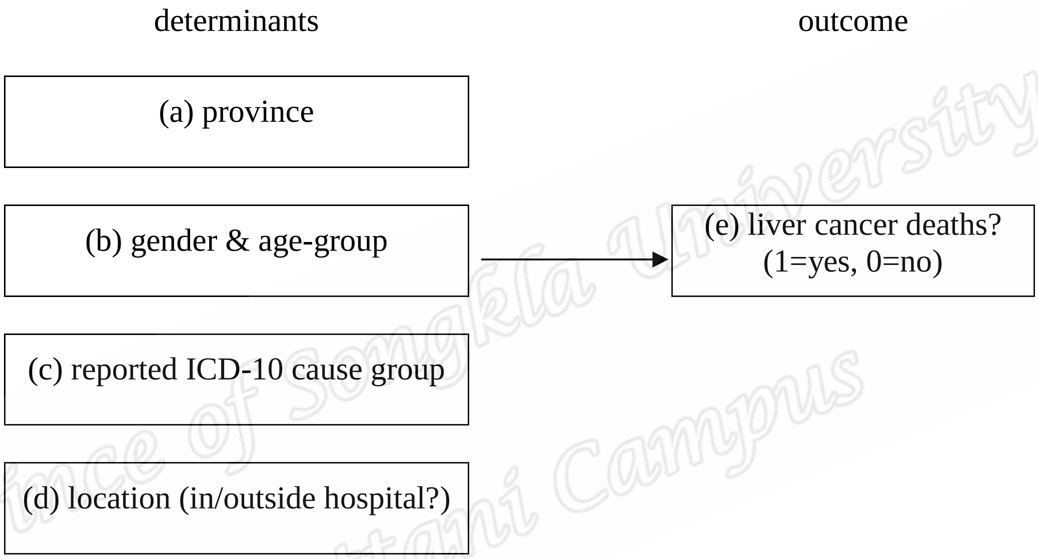

Figure 2.3: Path diagram

The determinants are thus separated naturally into regional, demographic, and medical components.

Figure 2.4 shows analysis process of our study. The analysis process comprised two main steps. The first step is to estimate number of liver cancer death based on the VA study using logistic regression. The predicted outcome values from logistic model are used to correct number of reported deaths. As a result, the estimated numbers of liver cancer deaths from 2000 to 2009 are obtained. The second step is to estimate liver cancer mortality rates based on the estimated deaths using Poisson regression model.

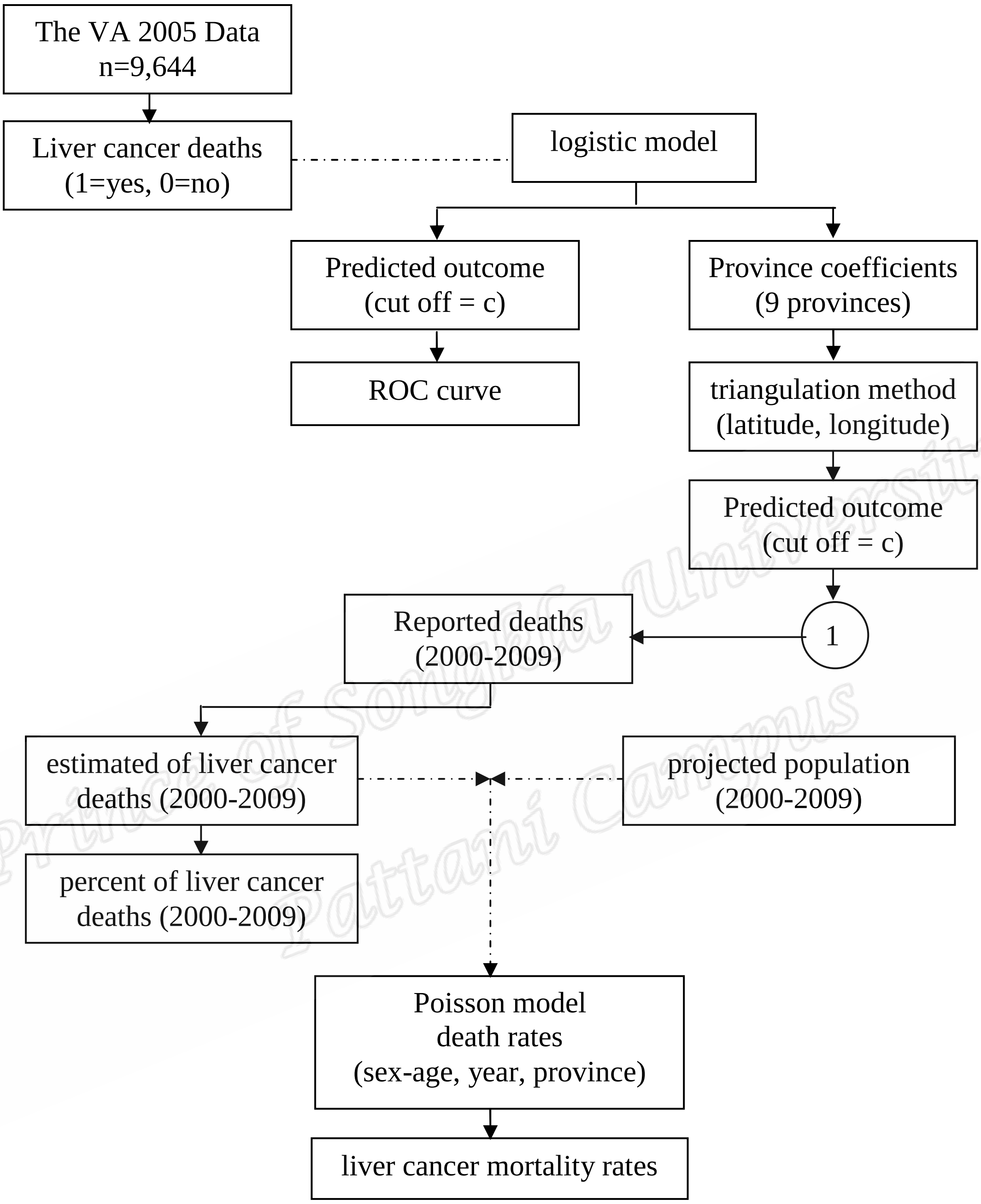

Figure 2.4: Analysis process

Logistic regression model

This model formulates the logit of the probability *p* that a person died from liver cancer as an additive linear function of the three determinant factors as follows:

$$\log\left(\frac{p}{1-p}\right) = \mu + \alpha_i + \beta_j + \gamma_k \qquad (2.1)$$

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, respectively.

The province factor has 9 levels corresponding to the 9 provinces in the VA sample. The gender-age group factor has 13 levels, by classifying age into 7 groups (0-29, 30-39, 40-49,…,70-79, 80+) for males and 6 groups for females (no females aged below 30 died from liver cancer). The VR cause-location factor has 12 levels, corresponding to the 6 most likely VR cause groups (liver cancer, other digestive cancer, other cancer, digestive, ill-defined and septicaemia, and other cause) and the two locations (in or outside hospital).

P-values of estimated coefficients

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.

ROC curve

The Receiver Operating Characteristic (ROC) curve is used to show how well a model predicts a liver cancer binary outcome. The ROC curve is defined as a plot of sensitivity on the y axis against false positive rate (1-specificity) on the x axis for the different possible cut points. It is an effective method of evaluating the performance of the model. It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

Denoting the predicted outcome as 1 (liver cancer) if $P \geq c$, or 0 (other death) if $P < c$, it plots sensitivity (proportion of positive outcomes correctly predicted by the model) against the false positive rate (proportion of all outcomes incorrectly predicted), as c varies.

Triangulation Method

To predict results for provinces outside the VA study, we estimated provinces' coefficients based on latitude and longitude of their central points. Triangles were drawn linking the nine VA provinces. These triangles were set at planes, like roofs on poles with heights corresponding to their model coefficients value at the vertices of the triangles.

For each triangle, values (a, b, c) are obtained by solving three equations as follows:

$$\mathrm{a}+(\mathrm{longitude}(\mathrm{prov1})\times \mathrm{b})+(\mathrm{latitude}(\mathrm{prov1})\times \mathrm{c}) = \mathrm{coef}(\mathrm{prov1})$$

$$\mathrm{a}+(\mathrm{longitude}(\mathrm{prov2})\times \mathrm{b})+(\mathrm{latitude}(\mathrm{prov2})\times \mathrm{c}) = \mathrm{coef}(\mathrm{prov2})$$

$$\mathrm{a}+(\mathrm{longitude}(\mathrm{prov3})\times \mathrm{b})+(\mathrm{latitude}(\mathrm{prov3})\times \mathrm{c}) = \mathrm{coef}(\mathrm{prov3})$$

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

$$\mathrm{coef}(\mathrm{prov}\boldsymbol{j}) = \mathrm{a}+(\mathrm{longitude}(\mathrm{prov}\boldsymbol{j})\times \mathrm{b})+(\mathrm{latitude}(\mathrm{prov}\boldsymbol{j})\times \mathrm{c})$$

Coefficients for provinces *outside* triangles are obtained similarly by extrapolation. The interpolated values for all 76 provinces reflect regional variation of liver cancer mortality compared to the reference province (Bangkok). The results are presented using thematic map.

Finally, we apply the model to the target population (all reported Thai deaths 2000-2009). To do this, we use the interpolated values for the province effects, and we assume that the model is valid for years before and after 2005. By doing this, the numbers of deaths were estimated for each gender-age groups and year. The area plot was used to show estimated liver cancer deaths for each gender-age groups for each year during 2000-2009.

2.4 Model of death rates

Estimated liver cancer death rates per 100,000 populations by province, gender, age group and year are now obtained by summing the fitted proportions given by the model over the 12 combinations of VR cause group and location, and multiplying by 100,000/*P*, where *P* is the corresponding population.

We then fit a Poisson generalized linear model and graph the adjusted death rates. The model takes the form.

$$\log\left(\frac{\lambda}{P}\right) = \alpha_i + \beta_j + \gamma_k \qquad (2.2)$$

In this model, λ is the mean of the Poisson distribution giving the number of liver cancer deaths for a specified province, gender-age group and year, and *P* is the corresponding population at risk in 100,000s.

Using sum contrasts (Venable and Ripley 2002), we obtained adjusted mortality rates and corresponding confidence intervals for comparing them with the overall average. The adjusted liver cancer mortality by region and year were presented using barchart, confidence interval and the map of Thailand.