# Chapter 4

# Statistical Modeling

The preliminary results in Chapter 3 show that liver cancer death reported were misclassification. The results of liver cancer deaths from the logistic regression model with province, gender-age group and VR cause location groups are presented in this chapter. Using the probabilities of liver cancer deaths from the model the numbers of liver cancer deaths in 2005 are estimated and they are used to correct liver cancer deaths in 2000 to 2009. The mortality rates are estimated from the Poisson model.

## 4.1 Logistic regression model

The logistic regression model as described in chapter 2 was fitted to the 2005 VA data. The coefficients, standard errors and p-value are presented in Table 4.1. The liver cancer deaths are high in Ubon Ratchathani in the Northeast and Phayao in the North. The liver cancer deaths are high in male aged 40-49, followed by male aged 60-69, male aged 50-59 and female aged 60-69. As expected, liver cancer deaths are more likely to be reported as liver cancer in hospital, liver cancer outside hospital and other cancer outside hospital.

Table 4.1 Coefficients, standard errors and p-value of logistic regression model

| determinant | Estimate | SE | p-value |
|---|---|---|---|
| **Constant** | 0.4460 | 0.5809 | 0.4426 |
| **Province** | | | |
| Bangkok | 0 | | |
| Nakhorn Nayok | 0.0135 | 0.4036 | 0.9734 |
| Ubon Ratchathani | 0.8073 | 0.2971 | 0.0066 |

Table 4.1: (Cont.)

| determinant | Estimate | SE | p-value |
|---|---|---|---|
| Loei | 0.5139 | 0.3371 | 0.1274 |
| Phayao | 0.9620 | 0.3438 | 0.0051 |
| Chiang Rai | 0.4197 | 0.3178 | 0.1867 |
| Suphan Buri | -0.3141 | 0.3469 | 0.3652 |
| Chumphon | -0.2090 | 0.5550 | 0.7065 |
| Songkhla | -0.0529 | 0.3910 | 0.8925 |
| **Gender: Age group** | | | |
| male:0-29 | 0 | | |
| male:30-39 | -0.7831 | 0.6184 | 0.2054 |
| male:40-49 | 1.1188 | 0.5096 | 0.0281 |
| male:50-59 | 0.9238 | 0.5027 | 0.0661 |
| male:60-69 | 0.9296 | 0.4946 | 0.0602 |
| male:70-79 | 0.6481 | 0.4989 | 0.1939 |
| male:80+ | 0.2897 | 0.5274 | 0.5827 |
| female:30-39 | -1.0873 | 0.7503 | 0.1473 |
| female:40-49 | 0.6175 | 0.5667 | 0.2759 |
| female:50-59 | 0.5531 | 0.5301 | 0.2967 |
| female:60-69 | 0.7936 | 0.5072 | 0.1177 |
| female:70-79 | 0.1088 | 0.5218 | 0.8349 |
| female:80+ | -0.1208 | 0.5340 | 0.8210 |
| **VR cause location group** | | | |
| liver cancer outside hospital | 0 | | |
| ill-defined+septicemia outside hospital | -2.7082 | 0.3050 | <0.0001 |
| digestive outside hospital | -3.4681 | 0.2544 | <0.0001 |
| other cancer outside hospital | -2.3235 | 0.2726 | <0.0001 |
| other digestivcancer outside hospital | -4.8355 | 0.2348 | <0.0001 |
| other cause outside hospital | -5.7123 | 0.2790 | <0.0001 |
| liver cancer in hospital | 0.4532 | 0.4170 | 0.2770 |
| ill-defined+septicemia in hospital | -3.0070 | 0.3736 | 0.0000 |
| digestive in hospital | -6.2592 | 1.0271 | 0.0000 |
| other cancer in hospital | -5.4846 | 0.6209 | <0.0001 |
| other digestivcancer in hospital | -5.9722 | 0.4087 | <0.0001 |
| other cause in hospital | -8.1173 | 0.7360 | <0.0001 |

## 4.2 P-values of estimated coefficients

The logistic regression model for estimating liver cancer deaths in the VA study gives the following p-values. We see that all the factors in the model are highly statistically significant as shown in Table 4.2.

Table 4.2: P-values of estimated coefficients

| factor | deviance reduction | df | p-value |
|---|---|---|---|
| VR cause-location | 1437.51 | 11 | <0.0000 |
| gender-age group | 57.89 | 12 | 0.0001 |
| Province | 43.41 | 8 | 0.0001 |
| error | 555.38 | 905 | |

## 4.3 The ROC curve

The ROC curve shows how well a model predicts a binary outcome. 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. Choosing $c = 0.216$ gives 500 predicted liver cancer deaths, in agreement with the VA study, for which the sensitivity is 0.64 and the false positive rate is 0.02. Table 4.3 shows cross classification between logistic model result and persons died from cancer based on VA assessed liver cancer deaths.

Note that just using the reported cause to predict the true cause has sensitivity 0.47 (236/500).

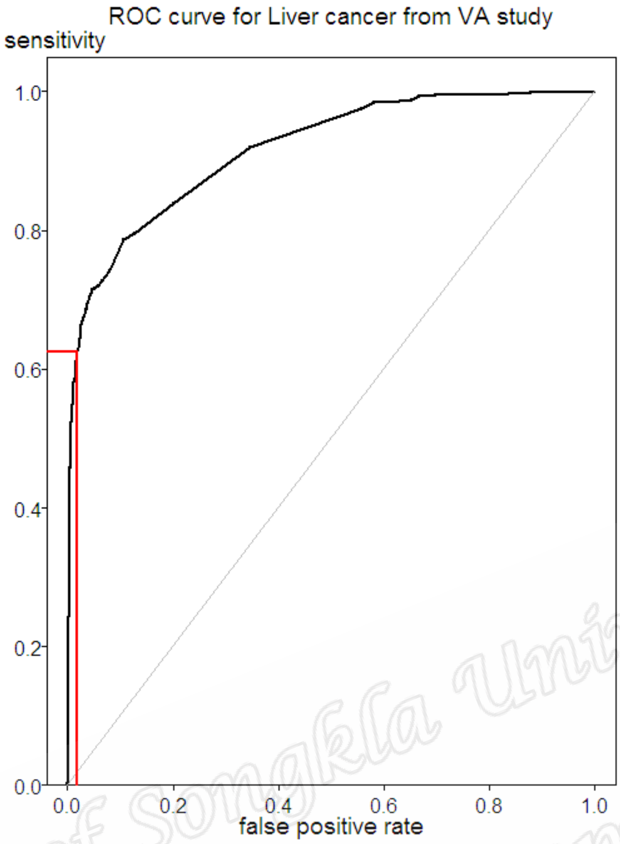


*Figure 4.1: Roc curve for liver cancer from the VA study*



Table 4.3: Cross classification of model results and VA assessed liver cancer deaths

| | | die from liver cancer | | |
|---|---|---|---|---|
| | | no | yes | total |
| model result | no | 8962 | 182 | 9144 |
| | yes | 182 | 318 | 500 |

## 4.4 Regional variation

The coefficients from the logistic regression model for the nine provinces are plotted (in black) in the map as shown in Figure 4.2. Values at other places (in blue) are averages of coefficients from nearby provinces.

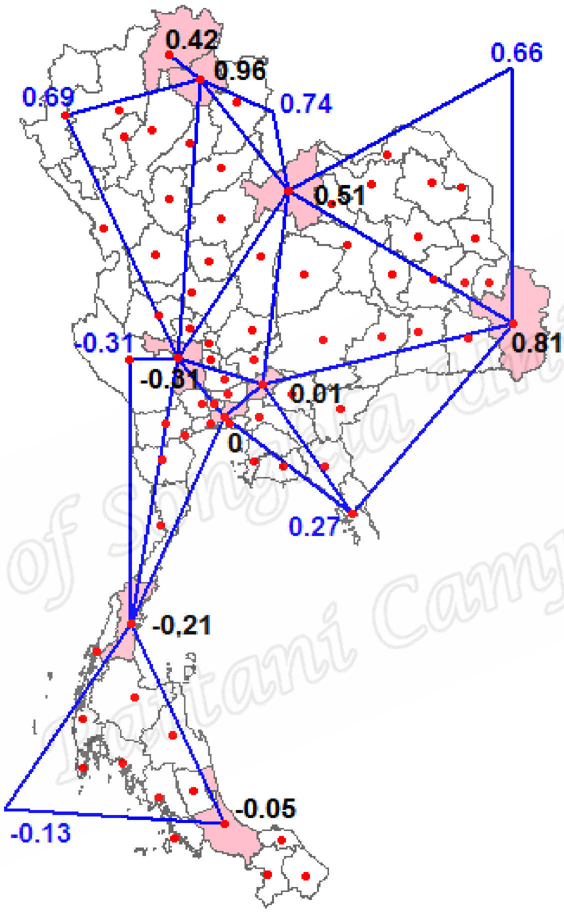


*Figure 4.2: Coefficients of province from logistic regression model*

The thematic map on Figure 4.3 interpolated values for all provinces. It shows that the highest liver cancer rates were in 12 provinces (Nong Khai, Udon Thani, Sakon Nakhon, Nakhon Phanom, Kalasin, Mukdahan, Maha Sarakham, Roi Et, Yasothon, Amnat Charoen, Si Sa Ket and Ubon Ratchathani) of the Northeast and 7 provinces

(Mae Hong Son, Chiang Mai, Lamphin, Lampang, Phayao, Phrae and Nan) of the Northern regions.

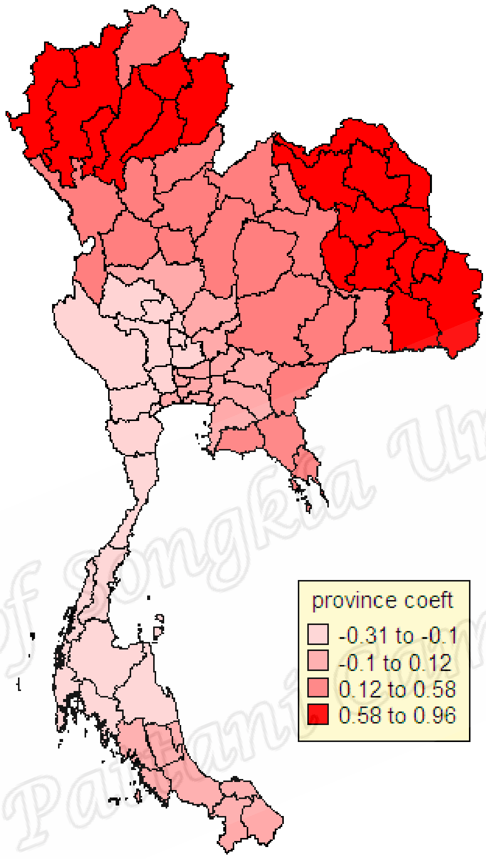


*Figure 4.3: The thematic map of province coefficients*



Next, we estimated percentage of liver cancer deaths. Figures 4.4-4.8 shows percent of liver cancer deaths by province from 2000 to 2009 using map of Thailand. The highest liver cancer mortality rates were in the Northeast and the Northern regions. The rates were remaining high from 2000-2009 in Phrae province of the North and the Northeast with the exception of Loei, Nakhon Ratchasima and Buri Ram.

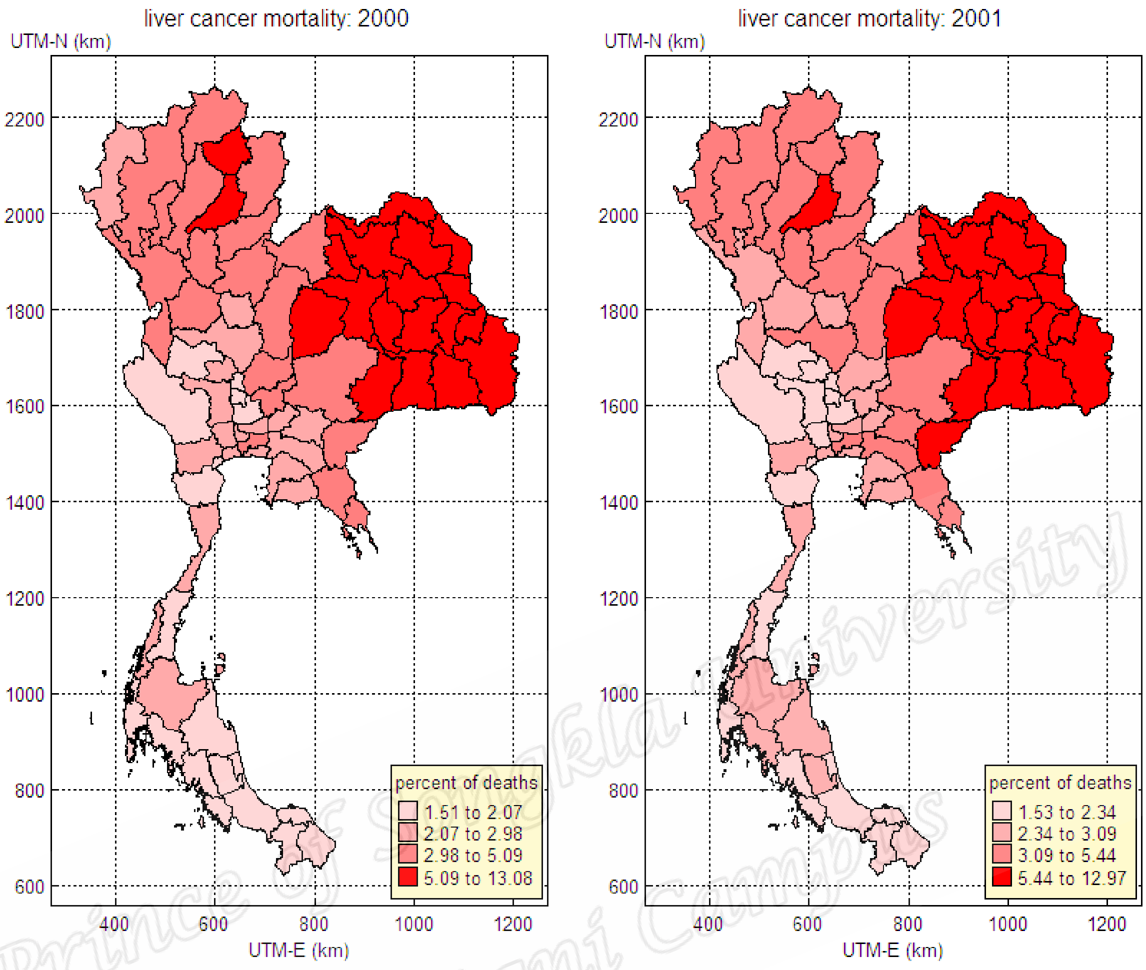


*Figure 4.4: Percent of liver cancer deaths in 2000-2001*

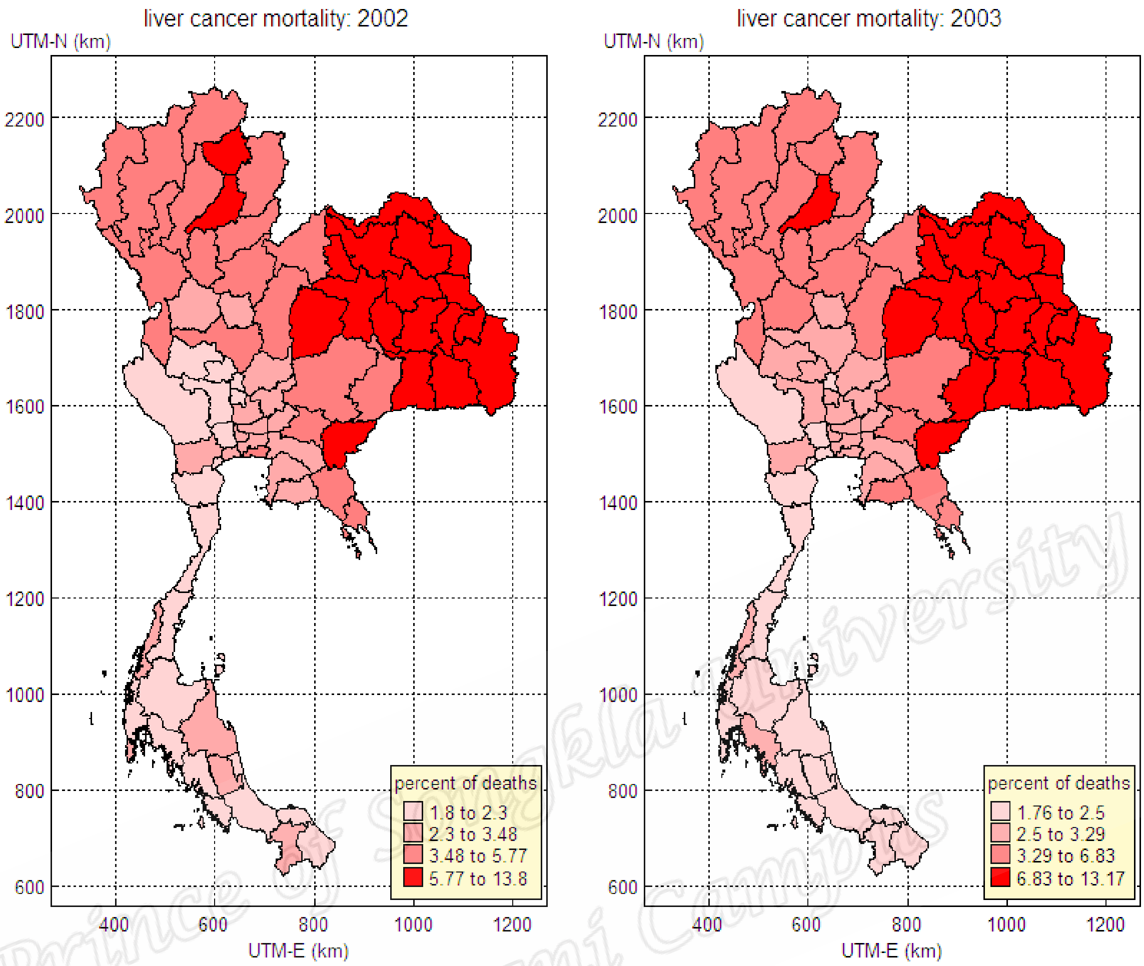


*Figure 4.5: Percent of liver cancer deaths in 2002-2003*

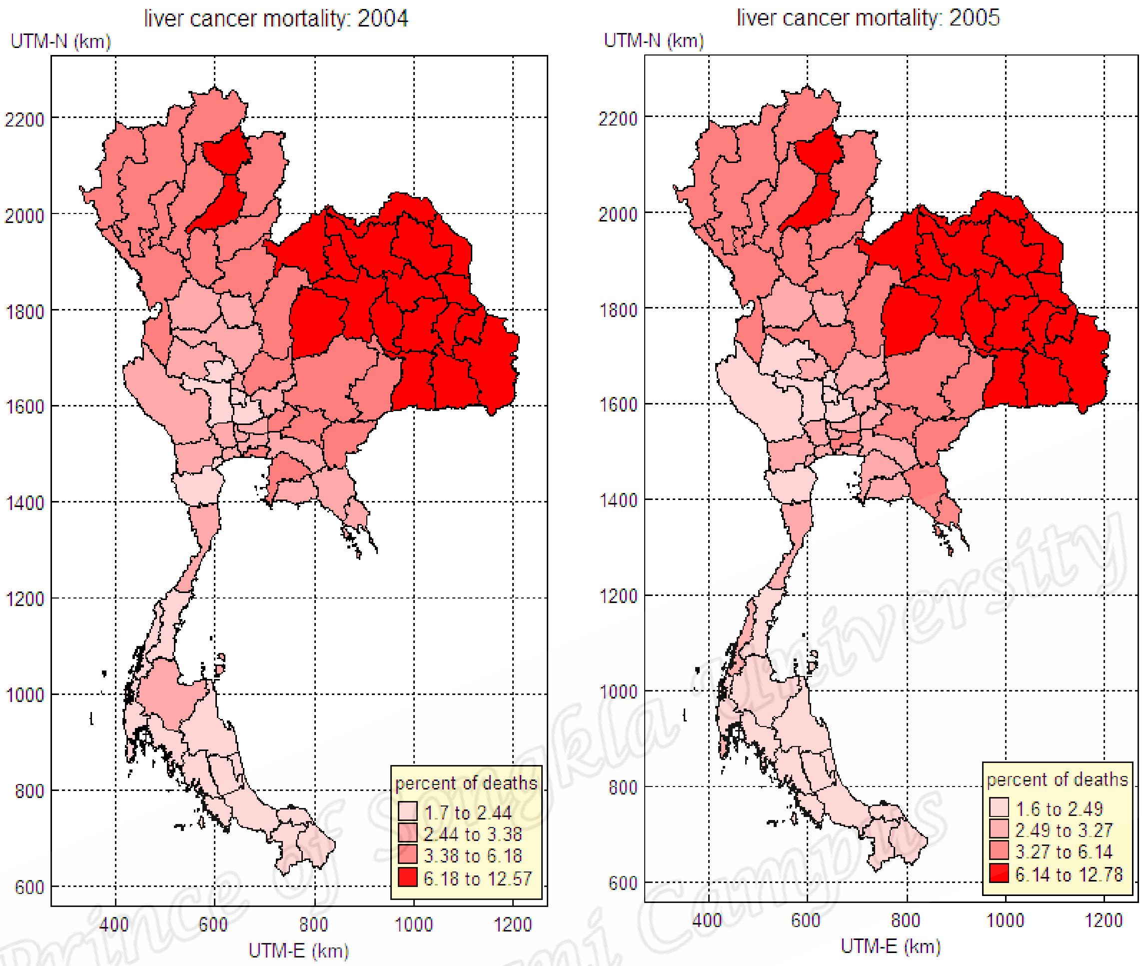


*Figure 4.6: Percent of liver cancer deaths in 2004-2005*

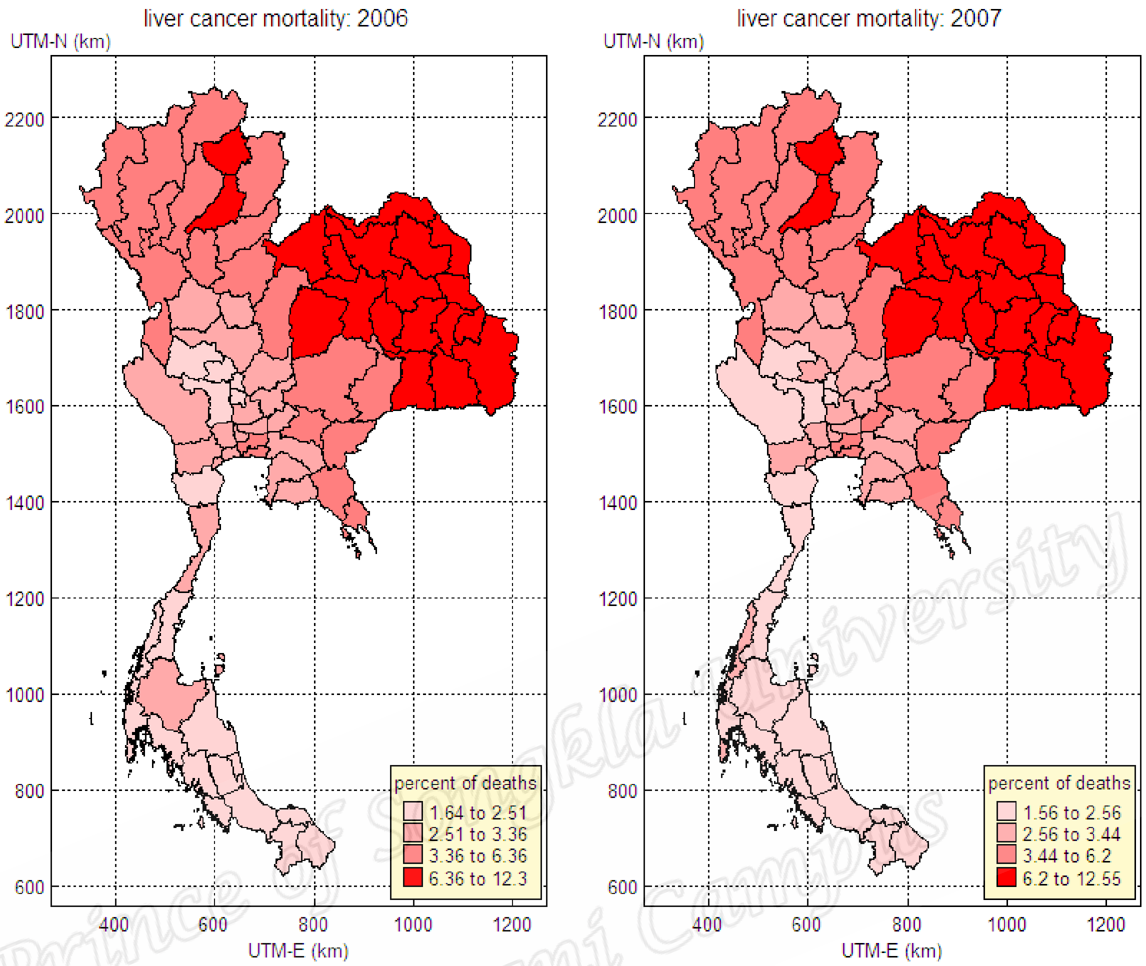


*Figure 4.7: Percent of liver cancer deaths in 2006-2007*

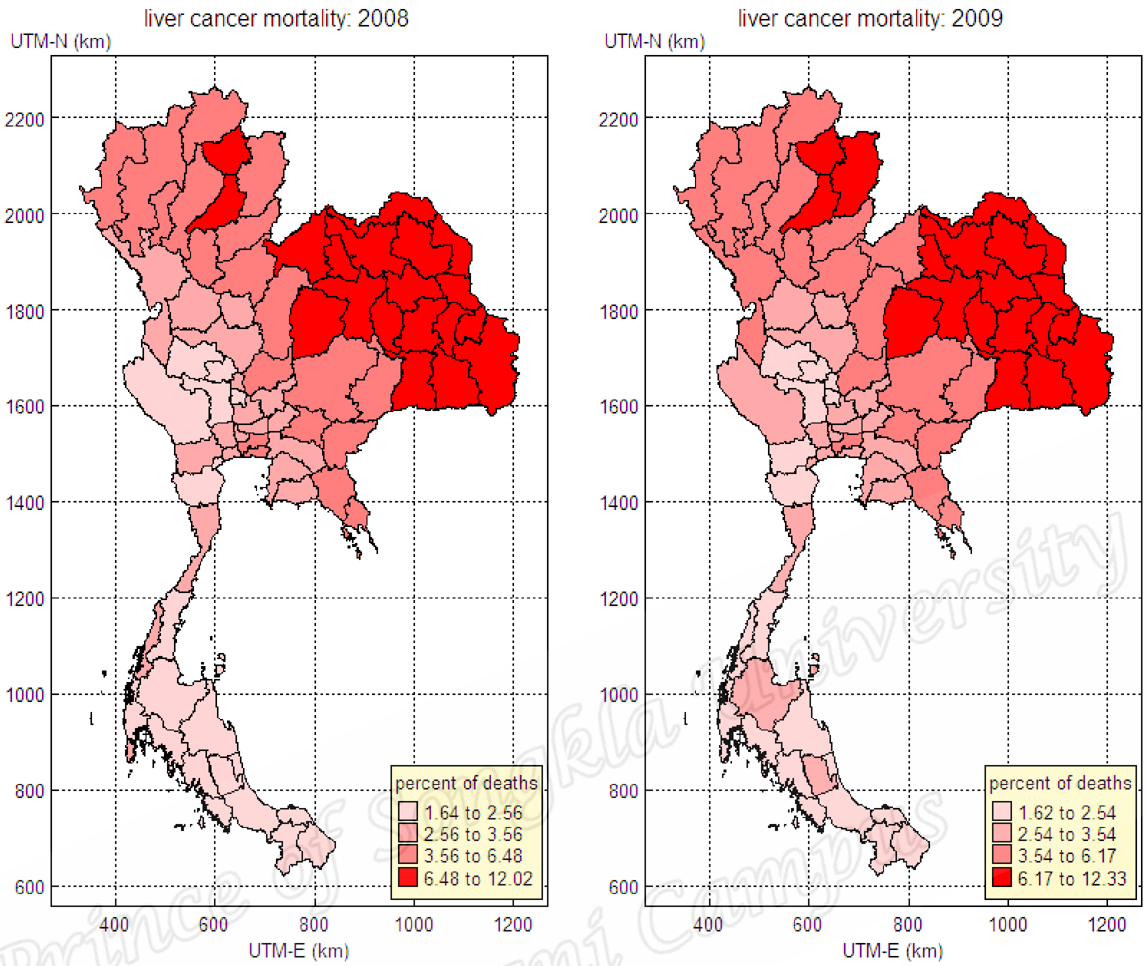




*Figure 4.8: Percent of liver cancer deaths in 2008-2009*

**4.5 Extending to the target population**

Finally, we apply the model to the target population. 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.

Over the decade 2000-2009, the estimated number of liver cancer deaths were 134,243.6 (males) and 61,964.5 (females). These are 56% and 74% higher than the reported totals of 85,873 and 35,643, respectively.

Figure 4.9 compares liver cancer death rates between VA estimated and VR reported deaths using area plot. The area plot shows increasing in number of deaths from 2000 to 2009. The VA estimated almost double the VR reported deaths.

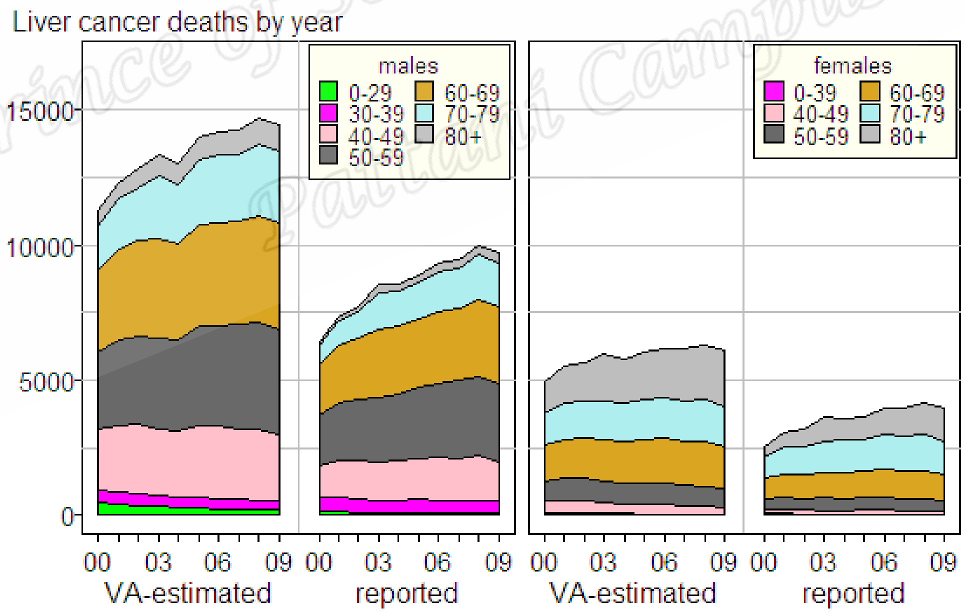


*Figure 4.9: Cumulative graphs from liver cancer deaths in 2000-2009*

### 4.6 Modeling death rates ("populating from the sample")

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. Figure 4.10 shows adjusted liver cancer death rates by gender-age groups, year and province. The liver cancer death rates are more pronounce among men. The overall death rate for male (42.16) is more than double of that for female (18.01). The adjusted rates increase with age for both male and female. The rates for male aged 60-69, 70-79 and 80+ are 198.20, 254.24 and 336.12 per 100,000 population, respectively. The rates at these ages for female are less than half of those for male. The year effect is relatively low. The rates range from 27.73 in 2009 to 31.57 in 2003.



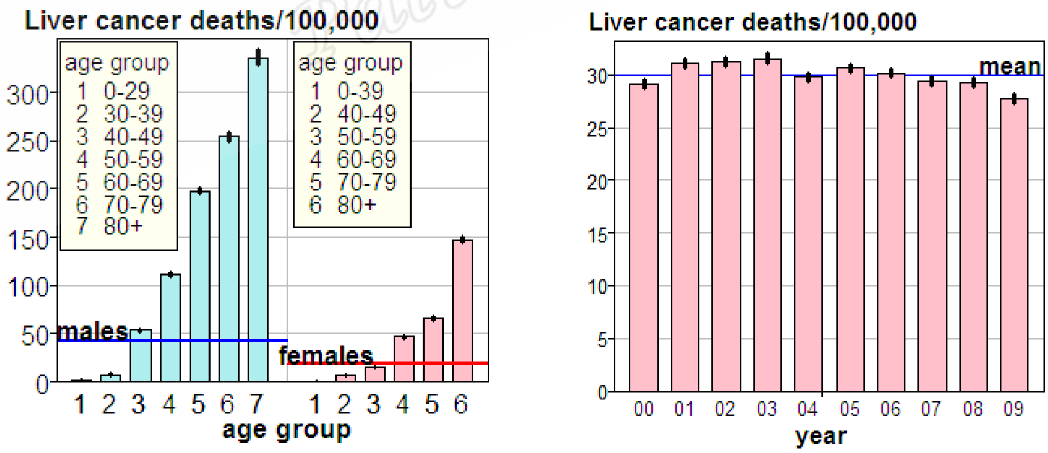


*Figure 4.10: Liver cancer deaths per 100,000 population by age group and year*

Figure 4.11 shows map of liver cancer mortality rates by province adjusted for gender-age group and year. Provinces with average mortality rates are Mae Hong Son and Prachin Buri. Provinces with above average mortality rates are in the Northeastern region except Nakhon Ratchasima and 10 provinces (Chiang Mai, Chiang Rai, Phayao, Nan, Phrae, Uttaradit, Sukhothai, Phetchabun, Lamphun and Lampang) of the Northern region. Mortality rates in Sa kaeo province is also above average. The rest are below average.

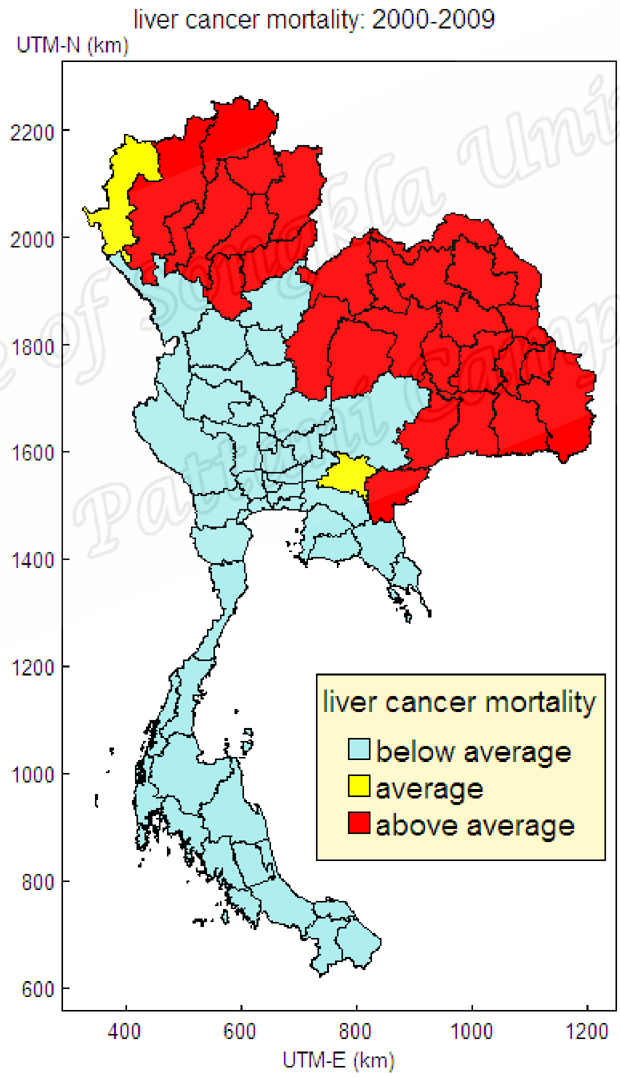


*Figure 4.11: Liver cancer deaths by province*