Pertanika J. Sci. & Technol. 25 (2): 469 - 478 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Estimating Lung Cancer Deaths in Thailand Based on Verbal Autopsy Study in 2005

Nattakit Pipatjaturon[1,2], Phattrawan Tongkumchum[2]* and Attachai Ueranantasun[2]

[1]*The office of Diseases Prevention and Control 2nd Phitsanulok, Phitsanulok 65000, Thailand*

[2]*Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, 94000, Thailand*

ABSTRACT

Information on the causes of death obtained from death certificates in Thailand is incomplete and inaccurate. Therefore, mortality statistics from death registrations (DR) remains unreliable. Accurate mortality statistics is essential for national policies on intervention and care and resource allocation. Verbal Autopsy (VA) is a more reliable source for cause of deaths than the DR. In this study, the classification of lung cancer deaths in Thailand from 1996 to 2009 was investigated based on a logistic regression model of lung cancer deaths with demographic and medical factors from the 2005 VA data. The estimated proportions of lung cancer deaths from the model were applied to the DR data. The goodness of fit of the model was assessed using the ROC curve. The resulting estimates of lung cancer deaths were higher than those reported with inflation factors 1.54 for males and 1.44 for females. Meanwhile, misclassified cases were reported mainly as other cancer types. There is no evidence of regional variation for lung cancer. The methods enable health professionals to estimate specific cause of deaths in countries where low quality of causes of death in the DR database and reliable data such as the VA data is available. The findings provide useful information on death statistics for policy interventions related to lung cancer prevention and treatment.

Keywords: Adjusted percentage, lung cancer deaths, logistic regression model, ROC

INTRODUCTION

An accurate statistics of causes of death is essential for monitoring the health of a nation and identifying priorities. Data on the causes of death obtained from death registration (DR) in Thailand are of low quality (Mathers et al., 2005) because 35-40% of deaths are ill-defined (Pataraachachai et al., 2010; Rao et al.,

Article history:
Received: 06 April 2016
Accepted: 09 August 2016

E-mail addresses:
nattakit@hotmail.com (Nattakit Pipatjaturon),
phattrawan.t@psu.ac.th; phattrawan@gmail.com (Phattrawan Tongkumchum),
attachai@gmail.com (Attachai Ueranantasun)
*Corresponding Author

ISSN: 0128-7680

2010). Extensive misclassification of causes of death (Tangcharoensathien et al., 2006) makes it necessary for mortality studies in Thailand to estimate the number of deaths using other data sources.

The VA in Thailand was conducted by Setting Priorities using Information on Cost-Effectiveness analysis (SPICE) project in 2005 to verify registered causes of death. This was the first national application of this WHO methodology to Thailand. Mortality estimates derived from making adjustments to the DR data in 2005 based on the VA using the simple cross-referencing method have been published (Porapakkham et al., 2010; Rao et al., 2010; Pattaraarchachai et al., 2010; Polprasert et al., 2010). However, this simple cross-referencing method ignored the effects of gender-age groups and location of the decease. That could give incorrect estimates due to confounding. This study offered an alternative approach based on statistical methods applied to a large-scale VA study focusing on lung cancer death.

The reasons for misclassification of the causes of death include a lack of properly trained physician to identify the causes of death for chain of illnesses and a lack of medical knowledge by head of the village (Kijsanayotin et al., 2013). Extensive misclassification of causes of death (Tangcharoensathien et al., 2006) makes it necessary for mortality studies in Thailand to estimate the valid number of deaths for improved DR database and thus vital statistics system in Thailand.

VA is a research method initiated by World Health Organization (WHO) to determine probable causes of death in case that there is no medical record or formal medical attention given. When DR cause of death is misclassified, the VA survey can be used to determine individuals' cause of death.

In 2005, the causes of death in Thailand were re-identified and reviewed using a VA questionnaire and a survey conducted by a physician with a training certificate in specifying causes of death based on International Classification of Diseases (ICD) (Polprasert et al., 2010). This was the first national application of this WHO methodology to find a solution for the low quality causes of death in Thailand.

The mortality estimations derived from making adjustments to the DR data in 2005 based on the VA have been published (Porapakkham et al., 2010; Rao et al., 2010; Pattaraarchachai et al., 2010; Polprasert et al., 2010). In order to reduce costs from conducting VA study for the whole country, an analysis of the VA data using appropriate statistical methods is an alternative approach to a large-scale VA survey such as in the case of HIV (Chutinantakul et al., 2014).

In Thailand, lung cancer contributed to 3.7% of all deaths for males in 2005, whereas it was 3.3% in 1999 (Porapakkham et al., 2010). The rising number of lung cancer deaths and rates for both sexes have been observed (Kamnerdsupaphon et al., 2008). The lung cancer incidence rates among Thai women exceeded those of women from many European countries such as Germany and Finland (Jemal et al., 2010).

Thus, this study aimed to estimate number of lung cancer deaths obtained from the DR during 1996-2009 using the VA data from 2005 with a statistical model of lung cancer deaths taking into account demographic and medical factors. Thus, after correction for the misclassified lung cancer deaths, a more accurate estimate of the proportion (or percentage or number) of deaths due to lung cancer could be obtained.

MATERIALS AND METHODS

Data source and management

This study used secondary data from a 2005 VA survey, which assessed the causes of death based on a sample of 9,644 cases (3,316 in-hospital deaths and 6,328 outside-hospital deaths) from 28 districts in nine provinces (Rao et al., 2010). The nine provinces selected were Bangkok and two provinces from each of the four regions in Thailand. The selected provinces were those whose numbers of reported deaths were above (one province) and below (one province) the median. Similarly, twenty-eight districts were selected from the provinces. Approximately 50% of the death certificates were selected from all the villages and urban areas within the 28 selected districts using simple random sampling.

Since no lung cancer deaths occurred in those aged below five years, the study sample was reduced to 9,495 cases aged five years and older (3,212 in-hospital deaths and 6,283 outside-hospital deaths). Data obtained from each case were the province, gender, age, location of death (in or outside hospital), DR-reported International Statistical Classification of Diseases (ICD-10) code reported on the death certificate and VA-assessed ICD-10 code.

Data Analysis

We analysed the VA data in this study using the chapter-block classification for ICD-10 codes based on mortality tabulation (World Health Organization, 2004), creating 21 major cause groups of deaths. The groups had to be large enough for statistical analysis. The 21 groups are described elsewhere (Chutinantakul et al., 2014; Waeto et al., 2014).

The outcomes of interest were the VA-assessed ICD-10 codes for lung cancer deaths (C30-C39) or others. The determinants were province, gender, age, location of death and DR-reported ICD-10 code. The VA-assessed ICD-10 codes and the DR-reported ICD-10 codes were cross tabulated to give five cause groups (namely, lung cancer, ill-defined, other cancer, respiratory disease and other), where lung cancer deaths are often misreported. The location of death and DR-reported ICD-10 codes were categorised into 10 groups: 5 DR-reported ICD-10 code groups each for the two locations (in and outside the hospitals). Gender and age were classified into 7 groups by gender: ages 5-29, 30-39, 40-49, 50-59, 60-69, 70-79 and 80+years for each sex. Nine provinces (namely, Bangkok, Nakhon Nayok, Suphan Buri, Ubon Ratchathani, Loei, Phayao, Chiang Rai, Chumphon, and Songkhla) were included in the VA study.

Logistic Regression Model

We estimated the logit of the probability that a person died from lung cancer as a linear function of the determinant factors using logistic regression (McNeil, 1996; Venables & Ripley, 2002; Hosmer & Lemshow, 2004). The simple model is formulated as,

$$\log\left[\frac{p_i}{1-p_i}\right] = \mu + \alpha_i \qquad [1]$$

Where, p_i is the probability of death due to lung cancer, μ is a constant, and α_i is the parameter of DR cause location *i*. The simple model was compared with the full model (2), which included an additive linear function of further determinant factors. The full model is formulated as,

$$\log\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = \mu + \alpha_i + \beta_j + \gamma_k \qquad [2]$$

Where, p_{ijk} is the probability of death due to lung cancer, μ is a constant, and α_i, β_j, and γ_k are parameters specifying DR cause location *i*, gender-age group *j*, and province *k*, respectively. This equation may be inverted to give an expression for the probability p_{ijk} as,

$$p_{ijk} = 1/\left(1+\exp\left(-\left(\mu+\alpha_i+\beta_j+\gamma_k\right)\right)\right) \qquad [3]$$

We fitted logistic regression model using sum contrasts (Venables & Ripley, 2002; Tongkumchum & McNeil, 2009; Kongchouy & Sampantarak, 2010; Sampantarak et al., 2011) instead of conventional treatment contrast, where the first level is left out from the model to be the reference. This model allows us to compute the 95% confidence intervals of lung cancer deaths for each of the covariate levels in the VA data.

Goodness of fit of the model

We used the Receiver Operating Characteristic (ROC) curve (Chongsuvivatwong, 2007) to show how well the simple and full model predicts a binary outcome. It plots sensitivity (proportion of positive outcomes correctly predicted by the model) against the false positive rate (proportion of all outcomes incorrectly predicted). Sensitivity and specificity of the model is a cut-off point in the curve, where the predicted number of lung cancer death is in agreement with the observed value in the VA data. Area under the curve (AUC) represents model accuracy (Sarkar & Midi, 2010).

Spatial triangulation method

The full model gave 10 coefficients for DR-cause location, 14 coefficients for gender-age group and 9 coefficients for province. The province coefficients were used to interpolate coefficients for remaining 67 provinces outside the VA study using a spatial triangulation method based on the latitude and longitude of their central point. The spatial triangulation method is described elsewhere (Chutinantakul et al., 2014).

Extension to DR data

Coefficients for province, gender-age group and DR cause-location were applied to all deaths in the DR data for each year. Assuming the models were correct for 1996-2009, the VA-estimated lung cancer deaths from 1996 to 2009 were thus obtained. Graphical displays and statistical analyses were performed using the R programme version 3.0.1 (R Core Team, 2013).

RESULTS

Of the 9,495 deaths, the VA-assessment gave 320 lung cancer deaths (117 in-hospital deaths and 203 outside-hospital deaths). Only 164 lung cancer deaths were correctly DR-reported. The rest were reported as ill-defined (89), other cancer (32), respiratory (17) and others (18).

The DR-cause-location factor was found to be highly statistically significant in the simple model. Table 1 shows all p-values from the full model. The DR-cause location and gender-age-group factors were highly statistically significant, but there was no significant evidence of a province effect. Although province was not significant, it was retained in the model as a basis for estimating lung cancer deaths for every province in the country.

Table 1
P-values of the estimated coefficients

Factor	Deviance reduction	df	p-value
DR cause-location	1005.23	9	<0.0000001
gender-age group	61.60	13	<0.0000001
province	10.81	8	0.2124
error	468.98	903	

Figure 1 shows ROC curves for both the simple and the full models. The cut-off point in the ROC curve gives the predicted number of lung cancer deaths (319) agreement of the observed value in the VA data set (320). The red lines, drawn from the cut-off point to the *x*-axis and *y*-axis, show the model sensitivity and specificity (1-false positive rate). The full model gives 55.3% sensitivity, 98.5% specificity and AUC 0.80, whereas the simple model gives 51.2% sensitivity, 99.4% specificity and AUC of 0.70.

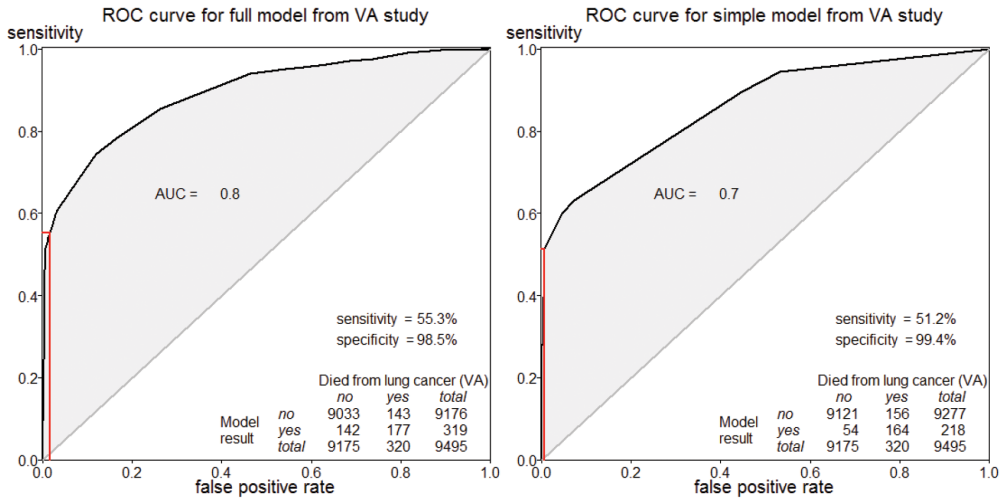

Figure 1. ROC curve for the full fitted model and simple fitted model from the VA study

Adjusted Percentages of Lung Cancer Deaths

The results derived for the model are presented as a graph of adjusted percentages and their corresponding 95% confidence intervals. Figure 2 shows the crude percentages of lung cancer deaths superimposed with adjusted percentages and their corresponding 95% confidence intervals. The horizontal red line is the average percentage of lung cancer deaths (3.4%). In order to distinguish the bar chart and 95% confidence interval, a non-linear vertical axis scale was used.

There is no evidence of province effects. The 95% confidence intervals above average were found in the age groups 60-79 for males. These age groups were more likely to have high levels of under-reporting. Meanwhile, the 95% confidence intervals for lung cancer and other cancer (outside-hospital) are above average. Other cancer outside hospital is the group in which lung cancer deaths were often misclassified.

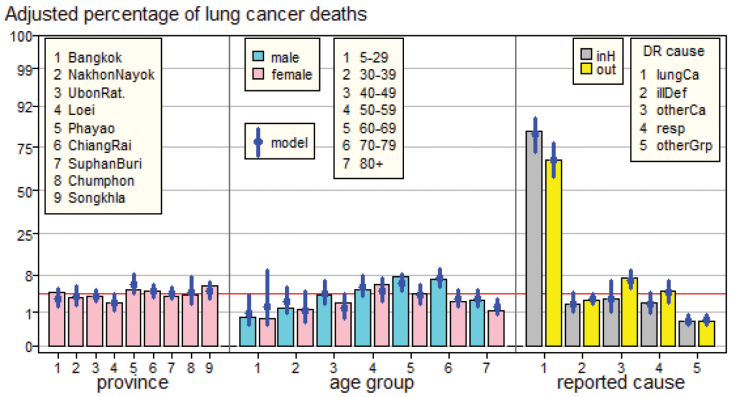

Figure 2. Adjusted percentages of lung cancer death by province, gender-age group and DR

Figure 3 shows the DR estimate of lung cancer deaths by gender-age group in 2005. The numbers of lung cancer deaths from DR reports were 5,887 and 2,549 cases for males and females, respectively. The simple model estimated the number of lung cancer deaths of 7,549 for males and 4,796 for females. The full model estimated the number of lung cancer deaths to be 8,503 in males and 3,433 in females, respectively. These were 44.4% and 34.7% higher than the corresponding DR reports.

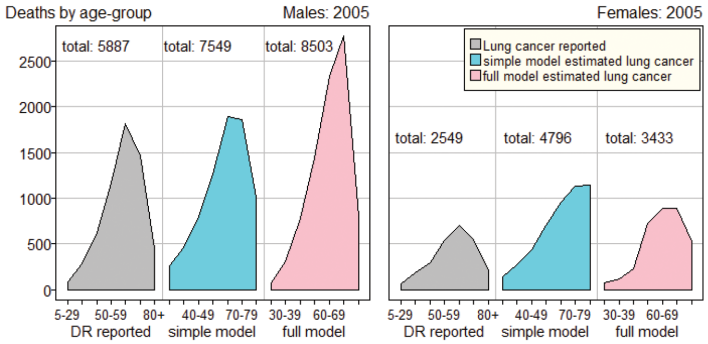

Figure 3. DR reports of lung cancer deaths and estimates from the simple and full models in 2005 by age groups

Figure 4 shows DR reported, the simple model estimated and the full model estimated the numbers of lung cancer deaths by age group in the years from 1996 to 2009. Apart from the drop in 1997-1998, when data are known to be incomplete in the DR database and a correction for temporally lost data from 2004 to 2005, the curves based on the statistical models are quite smooth and thus provide a credible basis for forecasting.

The numbers of lung cancer deaths rose rapidly with year, especially in males. Lung cancer deaths at ages 40+ years tended to increase in both sexes over the 14-year period, whereas deaths at ages 5-39 years tended to decrease. The total numbers of lung cancer deaths reported for 14 years were 64,819 in males and 28,491 in females. The estimated total numbers of lung cancer deaths from the simple model were 89,877 in males and 58,152 in females, whereas the estimates from the full model were 99,671 in males and 40,980 in females. The resulting estimates of lung cancer deaths from the full model were higher than those reported with inflation factors 1.54 for males and 1.44 for females.

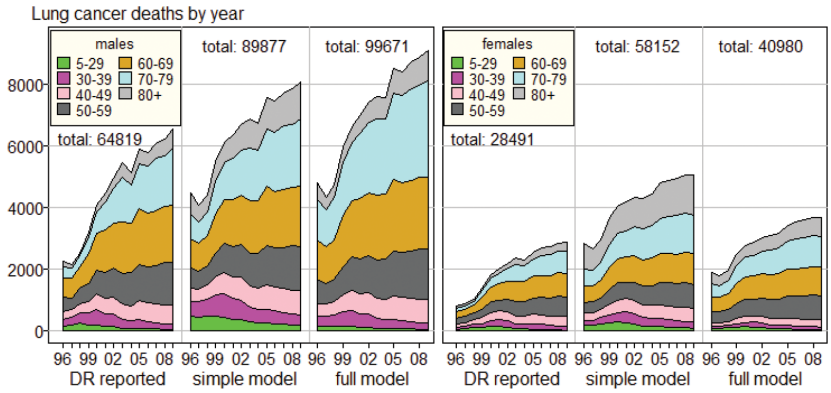

Figure 4. DR reports of lung cancer deaths and estimates from the simple and full models of lung cancer deaths by age groups and years

DISCUSSION

This study has shown that a logistic regression model of lung cancer deaths, with gender-age group, DR cause location and province from the VA data, can be used to adjust the number of deaths in the DR database. However, the accurate cause of death for individuals is uncertain, particularly for causes being reported as ill-defined or unknown cause. Meanwhile, goodness of fit of the model, as assessed by the ROC curve, indicates that the model adequately separates lung cancer deaths from other causes.

The finding from this study is that the medical and demographic factors in the model are highly statistically significant, but there is no evidence of any regional effect. The adjusted percentages of lung cancer deaths were high among the elderly males. Most lung cancer deaths were correctly reported. Misreported cases were mostly observed for deaths outside hospitals, and they were often reported as other cancer groups (C^-, D00-D48). The estimated numbers of lung cancer deaths from the models were higher than those being reported.

Logistic regression is commonly used in health studies. There are advantages in using the logistic regression model. The method gives confidence intervals for percentages of lung cancer deaths for levels of each risk factor adjusted for other risk factors using methods developed by Tongkumchum and McNeil (2009) and Kongchouy and Sampantarak (2010). These confidence intervals, when compared with bar charts of sample percentages, provide evidence of confounding bias. Moreover, the model can be extended to the larger target population comprising all deaths in Thailand for longer periods of time. Additionally, it can also be used to forecast lung cancer deaths and other specific causes.

However, the method also has some limitations. First, bias may have arisen in the sampling design. The VA study used a clustered sample design, but this sample did not include many subjects from rural places, and none at all from the many Muslim majority districts. Moreover, our model assumes that the patterns of misreporting of deaths in 1996-2009 are the same as in 2005 when the VA study was undertaken. This assumption is questionable, particularly for the years before 2005 when reporting practices were distorted by the HIV epidemic.

Our findings of high lung cancer deaths in elderly males are not surprising. It is well known that lung cancer is common among the elderly, and it more pronounced in males. A previous study in Thailand reported that lung cancer was common in patients aged 50 years or more (Deesomchok et al., 2005).

Although no evidence of regional effect was found in this study, a study on cancer control in Thailand using cancer registration data found high incidence rates of lung cancer in the northern region (Vatanasapt et al., 2002). Geographical variation on lung cancer deaths in 2000 have been observed with high rates in Bangkok (Faramnuayphol et al., 2008). This inconsistency is difficult to explain and there are not many studies on lung cancer deaths in Thailand. The findings on having no evidence of regional effects reported in this study will be useful for research in lung cancer mortality meta analyses.

This study found high percentages of lung cancer deaths, especially deaths in hospitals that were correctly reported and some misclassifications due to other cancers. This finding agrees with a previous study, where lung cancer deaths were observed to not contributing significantly to ill-defined cancer coding (Porapakkham et al., 2010).

CONCLUSION

This method enables public health researchers to estimate the percentage of specific causes of death in countries where there is low quality for recorded causes of deaths, but reliable sample data such as a VA study are available.

ACKNOWLEDGEMENTS

We are grateful to Prof. Don McNeil for his guidance, support and assistance. We are also thankful to Dr. Kanitta Bundhamcharoen from the Bureau of Policy and Strategy, Ministry of Public Health, Thailand, for providing us the data. Finally, we thank to the Graduate School of the Prince of Songkla University for the supported scholarship for Nattakit Pipatjaturon.

REFERENCES

Chutinantakul, A., Tongkumchum, P., Bundhamcharoen, K., & Chongsuvivatwong, V. (2014). Correcting and estimating HIV mortality in Thailand based on 2005 verbal autopsy data focusing on demographic factors, 1996-2009. *Population Health Metrics, 12*(1), 25-32. http://dx.doi.org/10.1186/s12963-014-0025-x.

Chongsuvivatwong, V. (2007). *Graphs, Tables and Equations for Health Research*. Bangkok: Chulalongkorn University Press.

Core R Team. (2012). *A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing*. Retrieved from http://www.R-project.org [Cited 2013-05-16].

Deesomchok, A., Dechayonbancha, N., & Thongprasert, S. (2005). Lung cancer in Maharaj Nakorn Chiang Mai Hospital: Comparison of the clinical manifestations between the young and old age groups. *Journal of the Medical Association of Thailand, 88*(9), 1236-1241.

Faramnuayphol, P., Chongsuvivatwong, V., & Panarunothai, S. (2008). Geographical variation of mortality in Thailand. *Journal of the Medical Association of Thailand, 91*(9), 1455-1460.

Hosmer, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. New York: John Wiley & Sons.

Jemal, A., Center, M. M., DeSantis, C., & Ward, E. M. (2010). Global Patterns of Cancer Incidence and Mortality Rates and Trends. *Cancer Epidemiology Biomarkers & Prevention, 19*(8), OF1-15.

Kamnerdsupahon, P., Srisukho, S., Sumitsawan, Y., Lorvidhaya, V., & Sukthomya, V. (2008). Cancer in Northern Thailand. *Biomedical Imaging and Intervention Journal*, *4*(3), e46-e52. Retrieved from http://dx.doi.org/10.2349/biij.4.3.e46.

Kijsanayotin, B., Ingun, P., & Sumputtanon, K. (2013). *Rapid Assessment of National Civil Registration and Vital Statistics Systems: A case study of Thailand*. Nonthaburi, Thailand: Thai Health Information Standards Development Center, Health Systems Research Institute.

Kongchouy, N., *& Sampantarak*, U. (2010). Confidence intervals for adjusted proportions using logistic regression. *Modern Applied Science, 4*(6), 2-6.

Mathers, C. D., Fat, D. M., Inoue, M., Rao, C., & Lopez, A. D. (2005). Counting the dead and what they died from: and assessment of the global status of cause of death data. *Bulletin of the World Health Organization*, *83*(3), 171-177.

McNeil, D. (1996). *Epidemiological Research Methods*. New York: John Wiley & Sons.

Pattaraarchachai, J., Rao, C., Polprasert, W., Porapakkham, Y., Pao-in, W., Singwerathum, N., & Lopez, A. D. (2010). Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. *Population Health Metrics, 8*(1), 1-12. http://dx.doi.org/ 10.1186/1478-7954-8-12.

Polprasert, W., Rao, C., Adair, T., Pattaraarchachai, J., Porapakkham, Y., & Lopez, A. D. (2010). Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. *Population Health Metrics, 8*(1), 13-27. http://dx.doi.org/ 10.1186/1478-7954-8-13.

Porapakkham, Y., Rao, C., Pattaraarchachai, J., Polprasert, W., Vos, T., Adair, T., & Lopez, A. D. (2010). Estimated causes of death in Thailand, 2005: implications for health policy. *Population Health Metrics, 8*(1), 14-24. Retrieved from http://dx.doi.org/ 10.1186/1478-7954-8-13.

Rao, C., Porapakkham, Y., Pattaraarchachai, J., Polprasert, W., Swampunyalert, N., & Lopez, A. D. (2010). Verifying causes of death in Thailand: rationale and methods for empirical investigation. *Population Health Metrics, 8*(1), 11-23. Retrieved from http://dx.doi.org/ 10.1186/1478-7954-8-11.

Sampantarak, U., Kongchouy, N., & Kuning, M. (2011). Democratic confidence intervals for adjusted means and incidence rates. *American International Journal of Contemporary Research, 1*(3), 38-43.

Sarkar, S. K., & Midi, H. (2010). Importance of Assessing the Model Adequacy of Binary Logistic Regression. *Journal of Applied Sciences*, *10*(6), 479-486.

Tangcharoensathien, V., Faramnuayphol, P., Teokul, W., Bundhamcharoen, K., & Wibulpholprasert, S. (2006). A critical assessment of mortality statistics in Thailand: potential for improvements. *Bulletin of the World Health Organization*, *84*(3), 233-239.

Tongkumchum, P., & McNeil, D. (2009). Confidence interval using contrasts for regression model. *Songklanakarin Journal of Science and Technology*, *31*(2), 151-156.

Vatanasapt, V., Sriamporn, S., & Vatanasapt, P. (2002). Cancer control in Thailand. *Japanese Journal of Clinical Oncology, 32*(suppl 1), S82-S91.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th Ed.). New York: Springer-Verlag.

Waeto, S., Pipatjaturon, N., Tongkumchum, P., Choonpradub, C., Saelim, R., & Makaje, N. (2014). Estimating liver cancer deaths in Thailand based on verbal autopsy study. *Journal of Research in Health Sciences*, *14*(1), 18-22.

WHO. (2004). *ICD-10 International Statistical Classification of Diseases and Related Health Problems*. Geneva: World Health Organization.