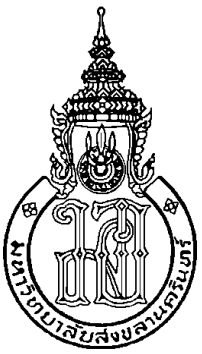


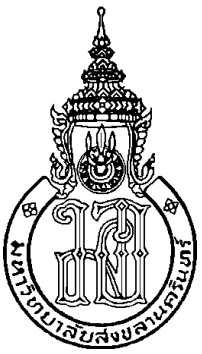
A Statistical Method for Correcting Misreported Death Registration Data with Application to Mortality in Thailand in 1996-2009

Nattakit Pipatjaturon

A Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Research Methodology

Prince of Songkla University

2016



A Statistical Method for Correcting Misreported Death Registration Data with Application to Mortality in Thailand in 1996-2009

Nattakit Pipatjaturon

A Thesis Submitted in Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Research Methodology

Prince of Songkla University

2016

Thesis Title A Statistical Method for Correcting Misreported Death Registration Data with Application to Mortality in Thailand in 1996-2009

Author Mr. Nattakit Pipatjaturon

Major Program Research Methodology

Major Advisor:

……..……..
(Asst. Prof. Dr. Phattrawan Tongkumchum)

Co-advisor

…….……
(Emeritus Prof. Dr. Don McNeil)

Examining Committee:

…..…….……Chairperson
(Asst. Prof. Dr. Apiradee Lim)

…….……..
(Asst. Prof. Dr. Phattrawan Tongkumchum)

…….……...
(Emeritus Prof. Dr. Don McNeil)

……..…….…….
(Asst. Prof. Dr. Chamnein Choonpradub)

……..…….……..
(Assoc. Prof. Dr. Yothin Sawangdee)

The Graduate School, Prince of Songkla University, has approved this thesis as fulfillment of the requirements for the Doctor of Philosophy Degree in Research Methodology.

……..….
(Assoc. Prof. Dr. Teerapol Srichana)
Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgements have been made of any assistance received.

……..……... Signature

(Asst. Prof. Dr. Phattrawan Tongkumchum)

Major Advisor

……..……... Signature

(Mr. Nattakit Pipatjaturon)

Candidate

I hereby certify that this work has not been accepted in substance for any degree, and is not being currently submitted in candidature for any degree.

……..……... Signature

(Mr. Nattakit Pipatjaturon)

Candidate

ชื่อวิทยานิพนธ์ วิธีการทางสถิติสำหรับการปรับแก้ข้อมูลรายงานสาเหตุการตายที่ผิดพลาดในประเทศไทย ปี 2539-2552

ผู้เขียน นายณัฐกิจ พิพัฒน์จาตุรนต์

สาขาวิชา วิธีวิทยาการวิจัย

ปีการศึกษา 2558

บทคัดย่อ

วิทยานิพนธ์นี้ใช้วิธีการทางสถิติในการปรับแก้ข้อมูลรายงานสาเหตุการตายของประเทศไทย (Death Registration: DR) ที่ให้สาเหตุการตายไม่ถูกต้อง โดยประยุกต์ใช้กับสาเหตุการตายจากมะเร็งตับ มะเร็งปอด และอัมพฤตอัมพาต ตั้งแต่ปี 2539-2552

การศึกษาทั้งสามโรคนี้ มีวัตถุประสงค์เพื่อปรับแก้ข้อมูลรายงานสาเหตุการตายที่ผิดพลาด โดยใช้วิธีการทางสถิติประมาณจำนวนการตายของมะเร็งตับ มะเร็งปอด และอัมพฤตอัมพาต ในกลุ่มอายุตั้งแต่ 5 ปีขึ้นไป ขั้นตอนแรก เริ่มจากสร้างตัวแบบการถดถอยโลจิสติก (Multiple logistic regression) โดยใช้ข้อมูลจากการสอบสวนสาเหตุการตายด้วยการสัมภาษณ์ (Verbal Autopsy: VA) ปี 2548 ของมะเร็งตับกับ 3 ตัวแปรอิสระคือ จังหวัด 9 จังหวัด กลุ่มเพศ-อายุ 12 กลุ่ม และกลุ่มสาเหตุการตายตามสถานที่ตาย 16 กลุ่ม ความเหมาะสมของตัวแบบที่ได้ตรวจสอบได้จากพื้นที่ใต้โค้งของ AUC (Area under an ROC curve) ซึ่งเป็นตัววัดความสอดคล้องระหว่างการตายด้วยมะเร็งตับจากการสัมภาษณ์กับค่าความน่าจะเป็นของการพยากรณ์การตายด้วยโรคมะเร็งตับ โดยหาจุดตัดบนเส้นโค้ง ROC curve ที่ทำให้การตายด้วยโรคมะเร็งตับจากตัวแบบเท่ากับหรือใกล้เคียงกับจำนวนการตายด้วยโรคมะเร็งตับจากข้อมูลการสอบสวนสาเหตุการตายด้วยการสัมภาษณ์ ขั้นตอนที่สอง คำนวณค่าสัมประสิทธิ์ถดถอยของ 67 จังหวัดที่เหลือโดยใช้วิธี Spatial triangulation method เพื่อนำค่าสัมประสิทธิ์ถดถอยมาแทนค่าในสมการของตัวแบบการถดถอยโลจิสติก เพื่อคำนวณหาค่าความน่าจะเป็นของการตายด้วยโรคมะเร็งตับในทุกจังหวัดตามกลุ่มเพศ-อายุ และกลุ่มสาเหตุการตายตามสถานที่ตาย ต่อไป ขั้นตอนที่สาม ประมาณค่าจำนวนการตายด้วยโรคมะเร็งตับโดยนำข้อมูลการตายในระบบทะเบียนราษฎร์ประเทศไทยปี 2548 ซึ่งประกอบด้วยจำนวนตายจากโรคมะเร็งตับ จำนวนตายทั้งหมด จำแนกตามจังหวัด กลุ่มเพศ-อายุ และกลุ่มสาเหตุการ

ตายตามสถานที่ตาย โดยแต่ละกลุ่มจะมีค่าสัมประสิทธิ์ถดถอยของเพศ-อายุ สัมประสิทธิ์ถอดถอยของสาเหตุการตายตามสถานที่ตายที่ได้จากตัวแบบ และสัมประสิทธิ์ถดถอยจังหวัดที่ได้จากตัวแบบและได้จากวิธี Spatial triangulation method จากนั้นคำนวณหาค่าความน่าจะเป็นจากการตายด้วยโรคมะเร็งตับโดยแทนค่าในสมการของตัวแบบเมื่อได้ค่าความน่าจะเป็นของการตายด้วยโรคมะเร็งตับในแต่ละกลุ่มทุกกลุ่มแล้ว นำค่าที่ได้คูณกับจำนวนการตายทั้งหมดในแต่ละกลุ่ม จะได้จำนวนการตายด้วยโรคมะเร็งตับในแต่ละกลุ่ม จากนั้นนำตัวแบบการถดถอยโลจิสติกการตายจากมะเร็งตับ ปี 2548 และค่าสัมประสิทธิ์ถดถอยจังหวัดที่ได้จากวิธี Spatial triangulation method ไปใช้ในการคำนวณหาค่าความน่าจะเป็นจากการตายด้วยโรคมะเร็งตับ และจำนวนการตายด้วยโรคมะเร็งตับกับข้อมูลปี 2539-2552 ต่อไป วิธีการเดียวกันนี้นำไปประมาณจำนวนการตายจากมะเร็งปอด และอัมพฤตอัมพาต

ผลการศึกษาพบว่า ตัวแบบการถดถอยโลจิสติกที่ใช้ประมาณค่าความน่าจะเป็นของการตายด้วยโรคมะเร็งตับ มะเร็งปอด และอัมพฤตอัมพาต มีความไว (Sensitivity) 62.6%, 55.3% และ 41.4% ตามลำดับ และให้ผลบวกลวง (False positive) 2.1%, 1.5% และ 7.4% ตามลำดับ

ข้อมูลการตายจากระบบทะเบียนราษฎร์ของประเทศไทยปี 2539-2552 มีจำนวนการตายด้วยโรคมะเร็งตับ มะเร็งปอด และอัมพฤตอัมพาตต่ำกว่าความเป็นจริงเมื่อเปรียบเทียบกับจำนวนการตายที่ประมาณได้จากตัวแบบการถดถอยโลจิสติก

การตายด้วยมะเร็งตับส่วนใหญ่ถูกบันทึกสาเหตุการตายเป็น มะเร็งทางเดินอาหาร โรคทางเดินอาหาร และมะเร็งอื่น ๆ การตายจากสาเหตุมะเร็งตับพบมากในเพศชายวัยทำงานในกลุ่มอายุ 40-59 ปี ในภาคเหนือตอนบนและภาคตะวันออกเฉียงเหนือ

การตายด้วยมะเร็งปอดส่วนใหญ่ถูกบันทึกสาเหตุการตายเป็นมะเร็งอื่น ๆ การตายจากสาเหตุมะเร็งปอดพบมากในเพศชายวัยทำงานอายุ 60-79 ปี ไม่มีความแตกต่างระหว่างพื้นที่

การตายด้วยอัมพฤตอัมพาตส่วนใหญ่ถูกบันทึกสาเหตุการตายเป็น ความผิดปกติทางจิตใจและระบบประสาท โรคหลอดเลือดหัวใจ และสาเหตุอื่น ๆ การตายจากสาเหตุอัมพฤตอัมพาตพบมากในเพศชายวัยทำงานในกลุ่มอายุ 60-79 ปี ในภาคกลางและภาคใต้

ร้อยละของการตายทั้ง 3 สาเหตุมีค่าสูงในทุกจังหวัด และทุกกลุ่มเพศ-อายุ จำนวนการตายของโรคมะเร็งตับ และมะเร็งปอดที่ได้จากการประมาณค่ามีแนวโน้มสูงขึ้นทุกปีอย่างต่อเนื่องยกเว้นปี 2540 และ 2547 ในขณะที่โรคอัมพฤตอัมพาตมีแนวโน้มสูงขึ้นเล็กน้อยทุกปี

Thesis Title A Statistical Method for Correcting Misreported Death Registration Data with Application to Mortality in Thailand in 1996-2009

Author Mr. Nattakit Pipatjaturon

Major Program Research Methodology

Academic Year 2015

ABSTRACT

In this thesis, we attempt to explain the methods necessary to estimate number of deaths based on verbal autopsy (VA) data because death registration (DR) data are misclassification cause of death. The methods were applied to liver cancer, lung cancer and stroke deaths in 1996-2009.

The analysis of the three causes of death aims to correct misclassification causes of death using statistical methods resulted in adjusted numbers of deaths for each cause using 2005 VA data of death at age five year and older. Firstly, logistic regression was used to model deaths from liver cancer with province (9 provinces), gender-age group (12 categories) and reported cause combined with location of death (16 categories). A receiver operating characteristic (ROC) curve was used to assess how well a model predicts a binary outcome. Secondly, interpolate province coefficients of the remaining 67 provinces outside the VA survey using spatial triangulation method. The province, gender age group, and cause location coefficients were used to estimate proportions of death from liver cancer in each combination category of determinants. Finally, estimate number of death in the DR database in 2005 using numbers of reported deaths for a particular gender-age group and cause location of death in the province multiply by their corresponding proportions of death. The model was extended to years 1996-2009. The same methods were also applied to lung cancer and stroke deaths.

It was found that the models of liver cancer, lung cancer, and stroke deaths give 62.6%, 55.3%, and 41.4% sensitivity, respectively. They give 2.1%, 1.5%, and 7.4% false positive rates.

Liver cancer deaths were most common among working age groups of males in upper north and northeast of the country. Most misclassifications of liver cancer deaths were classified as other digestive cancer or digestive disease (outside hospitals) or other cancers (outside hospitals).

Lung cancer deaths were common among males at retirement age. There is no evidence of province effect. Most misclassifications of lung cancer deaths were classified as other cancer (outside hospital).

Stroke deaths were common among males at retirement age similar to lung cancer. They were more common in central and southern regions of the country. Many of stroke deaths were registered as deaths from mental and nervous system disorders or other cardiovascular diseases (outside hospital) or other cause (in hospital).

In conclusion death registry is underreported liver and lung cancer deaths and substantially under reported stroke deaths. High proportions of deaths from the model reflect high proportions of misclassifications. After adjusted for misclassification, it is dramatically increasing trends for liver and lung cancer deaths during study period with the exception for 1997 and 2004. For stroke deaths, it is a gradually increasing trend.

Acknowledgements

I would like to express my gratitude and appreciation to my supervisor, Asst. Prof. Dr. Phattawan Tongkumchum and, my co-supervisor, Prof. Dr. Don McNeil for their invaluable assistance and helpful guidance on this thesis. Thanks also to Asst. Prof. Dr. Chamnein Choonpradub and Asst. Prof. Dr. Apiradee Lim, for their encouragement and suggestions.

I am grateful to the Thailand Bureau of Policy and Strategy and the National Health Security Office, Ministry of Public Health, for providing the data.

I would like to acknowledge Graduate School, Prince of Songkla University for funding this study and awarding me a scholarship.

Finally, I would like to thank sister and brother for their encouragement throughout my study.

Nattakit Pipatjaturon

Contents

Contents (cont.)

List of Tables

List of Figures

List of Figures (cont.)

List of Figures (cont.)

List of Acronyms

AUC	Area under the curve
CVD	Cerebrovascular Disease
DCI	Democratic confidence intervals
DR	Death registration
GU	Genitourinary
HIV	Human immunodeficiency virus
ICD-10	International Classification of Diseases and Related Health Problem 10th Revision
InH	In-Hospital
OutH	Outside-Hospital
NCDs	Non-communicable diseases
ROC	Receiver operating characteristic
SMR	Standardized mortality ratio
SPICE	Setting Priorities using Information on Cost-Effectiveness
TES	Thailand Epidemiology Stroke
VA	Verbal autopsy
WHO	World Health Organization

Chapter 1

Introduction

This thesis offers methods for correcting misreported multinomial outcome data with application to liver cancer, lung cancer, and stroke mortality in Thailand. The methods involve analysis of verbal autopsy (VA) sample for estimating proportion of deaths for each cause using logistic regression model with demographic factors, province of residence, location of death, and registered cause of death. The proportion of deaths in province outside the VA study was estimated using triangulation method. Then, utilize these findings to correct death registration (DR) data from 1996 to 2009 for liver cancer, lung cancer, and stroke deaths.

1.1 Background

Misreported cause of death in death registration database (DR) is common in developing countries (Mather *et al*., 2005). About 40% of death certificates in Thailand give the cause of death as “ill-defined”, thus many specific causes go largely under-reported. This limits their public utility.

A verbal autopsy (VA) survey was carried out in 2005. It aims to build capacity among Thai health professionals to improve the quality of cause of death recorded at registration (Rao *et al*., 2010).

Deaths from preventable diseases and premature deaths are a major problem in Thailand (Bureau of Policy and Strategy, 2010). Liver cancer, lung cancer and stroke are main non-communication diseases (NCDs) that cause of burden and death.

Liver cancer mortality was high (Jemal *et al*., 2010; Vatanasapt *et al*., 2002, Sripa *et al*., 2007; Viratroumanee *et al*., 2009). Age-standardized liver cancer mortality was 31.0 per 100,000 in Thailand in 2004 whereas it was 13.0 for Japan (WHO, 2008).

The rising death rates from lung cancer have been observed for both sexes (Kamnerdsupaphon *et al*., 2008). The lung cancer incidence rates among Thai women exceed those of women from many European countries, such as Germany and Finland (Jemal *et al*., 2010).

A stroke was ranked first among the top 15 causes of death in Thailand in 2005 based on the VA adjustment (Porapakkham *et al*., 2010). It is also a leading cause of disability and death among people aged 45 and above (Jalayondeja *et al*., 2011).

1.2 Rationale for study

A cause of death is important for providing mortality statistics at country level as well as at regional level. However, causes of death from DR data are of low quality (Mathers *et al*., 2005). The reasons for misclassification of the causes of death include a lack of properly trained skills for a physician in identifying the cause death for chain of illness, and a lack of medical knowledge for a 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 valid numbers of deaths for improved DR database and thus vital statistics system in Thailand.

VA is a research method, initiated by World Health Organization (WHO), helping to determine probable causes of death in cases that there was no medical record or formal medical attention given. When DR cause of death is misclassified, 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 VA questionnaires and survey conducted by a physician with a training certificate for 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 Thailand to find a solution for the low quality causes of death.

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). 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, for example, in case of HIV (Chutinantakul *et al*., 2014).

1.3 Objectives

This study offers methods for correcting misreported of liver cancer, lung cancer, and stroke mortality in Thailand. The specific objectives are as follows:

1. To construct appropriate models for correcting misreported liver cancer, lung cancer and stroke deaths based on the Thai VA study in 2005.

2. To estimate numbers of liver cancer, lung cancer and stroke deaths of the entire province in Thailand between 1996 and 2009.

1.4 Literature reviews

Literature review relates to mortality reporting system in Thailand, misclassification of causes of death and methods to correct them.

Overview of mortality reporting system in Thailand

In 1991, the Bureau of Registration Administration, Ministry of Interior was dominated as the central agency responsible for civil registration. Causes of deaths in hospitals are recorded using a Thai version of the standard International Form of Medical Certificate of Causes of Death, with an additional column in which the certifying physician records one cause in Thai to be used for registration purposes. This cause is entered in DR database by district officers. For unnatural deaths, causes are certified by a physician following forensic investigation (Rao *et al*., 2010).

For outside hospital deaths (65% of all deaths), local officers or village heads inquire the cause of death from family members and seek document evidence. Then record the reported cause of death in Thai. A complete DR database with a single cause in Thai for each death is transferred to the Ministry of Public Health, where the causes of death are coded according to ICD-10 (Rao *et al*., 2010). In 1996-2009, a gradual increase in the number of deaths from 288,941 in 1997 to 389,468 in 2008 in the DR database was reported.

Misclassification of deaths registry

Studies have reported misclassification in death registry for many developing countries. For example, Iranian death registry, about 20% of death statistics were recorded in misclassified categories such as septicemia, senility without mention of psychosis symptoms and other ill-defined conditions (Pourhoseingholi *et al*., 2011). A study in a district of Sri Lanka reported that cause of death was misclassified in 15% (Rampatige *et al*., 2013). In Brazil, about 6% of defined cause of death detected after investigation registered ill-defined condition as being due to injuries. In Thailand about 40% of death

statistics were recorded ill-defined (Rao *et al.*, 2010). It is known that the low quality of mortality data in Thailand's existing death registry. Because most causes of death especially death outside hospitals were reported by lay people. Valid causes of death data have been known to public health policy planners (Chuprapawon et al., 2003).

Methods to correct misclassification

A cross-referencing method has been used to correct misclassification cause of death in death registration database (Porapakkham *et al*., 2010; Pattaraarchachai *et al*., 2010; Polprasert *et al*., 2010; Rao *et al*., 2010; França *et al*., 2012). A correct cause of death was obtained from a small validation sample. Firstly, the data were adjusted for undercount of death registration, which expected to yield an estimate of true total number of deaths by age and sex in the region. Secondly, deaths with missing values (sex, age, and province) were handling by proportionately distributed across categories or imputation or omitting. Thirdly, the reported cause of death structure was corrected for the discrepancy arising from improper death certification and coding that had resulted in deaths being allocated to ill-defined cause (categories of limited public health utility). The adjusted proportionate cause structure was fitted to the estimated total deaths to derive mortality by age, sex, and cause. This method has been widely used and it used in previous analysis of Thai VA data. However, Byass (2010) concluded that uncertainties remain. Moreover, the simple cross-referencing method ignored the effect of sex-age groups and locality of the deceased, which could give incorrect estimate due to confounding.

Statistical approaches can also be used to estimate misclassification parameters. Many researchers (Lyles, 2002; Stamey *et al.*, 2008; Pourhoseingholi *et al*., 2009) have

introduced classical procedures that rely on asymptotic results and supplemental data in order to estimate unknown misclassification parameters. Two approaches to correct for misclassification are recommended (Stamey *et al.*, 2008). Firstly, a small validation sample is combined with main-study sample yielding more accurate parameter estimates (Lyles, 2002). Secondly, Bayesian analysis in which subjective prior information on at least some subset of the parameters is used to re-estimate death statistics (Whittemore and Gong, 1991; Sposto *et al.*, 1992). The Bayesian approach was used when no gold standard data set available. The Bayesian model was used to estimate the burden of gastrointestinal cancer, liver cancer (Pourhoseingholi *et al.*, 2010a), colorectal cancer (Pourhoseingholi *et al.*, 2009a), and gastric cancer (Pourhoseingholi *et al.*, 2009b). The Bayesian method estimates the parameters of a Poisson regression, where counts can be misclassified across the groups. It was claimed that the method is flexible. The likelihood approaches were also used in this case. It is forced to assume that misclassification is known when insufficient validation data are available. However, many mortality counts are not Poisson distribution due to over dispersion. This leads to limitation of using these methods.

1.5 Data

The VA and DR data sets were obtained from the Bureau of Policy and Strategy, the Ministry of Public Health of Thailand.

The DR data

The civil registration system of Ministry of Interior has provided electronic death data to the Ministry of Public Health since 1996. The data available comprise information from death certificates. Annual registered deaths aged five year and older including

percent of liver cancer, lung cancer, stroke, and ill-defined causes from 1996 to 2009 are shown in Table 2.1. Percent ill-defined ranged from 34.90% in 1997 to 42.25% in 1999.

Table 1.1 All-cause, liver cancer, lung cancer, stroke and ill-defined deaths in Thailand, 1996-2009

year	all-cause	% liver cancer	% lung cancer	% stroke	% ill-defined
1996	331,222	1.61	0.93	1.86	35.97
1997	288,941	1.73	0.92	1.74	34.90
1998	290,782	2.17	1.07	1.22	37.78
1999	347,657	2.36	1.39	1.94	42.25
2000	349,324	2.71	1.78	2.40	41.42
2001	358,601	3.13	1.96	3.37	38.64
2002	357,867	3.28	2.16	3.89	38.24
2003	359,383	3.68	2.36	5.27	33.72
2004	358,415	3.66	2.24	5.39	38.99
2005	384,689	3.78	2.55	4.70	38.75
2006	381,307	4.00	2.55	3.88	38.81
2007	385,136	4.05	2.66	3.92	38.70
2008	389,468	4.25	2.74	3.96	38.36
2009	386,401	4.12	2.85	4.02	38.33

The VA data

VA is increasing being used especially in setting without complete DR and is widely used in several countries (Mathers *et al*., 2005; Prasartkul *et al*., 2007). It is a method for providing accurate cause of death. The VA data was considered as a gold standard reference for cause of death where causes of death from DR are of low quality.

The recent VA study in Thailand was conducted in 2005 in nine provinces by the Setting Priorities using Information on Cost-Effectiveness (SPICE) analysis project. A multistage stratified cluster sampling was used to get a national representative of deaths

from two provinces of four regions (North, Northeast, Central and South), together with Bangkok. Study design, sampling plan and procedure have been described elsewhere (Rao *et al*., 2010).

The VA study assessed cause of death from a sample of 9,644 records. Since under-five deaths were separately analysed in another study, the study sample was reduced to 9,495 deaths aged five years and older (3,212 in-hospital and 6,283 outside-hospital deaths). Figure 1.1 shows nine provinces (ChiangRai, Phayao, UbonRatchathani, Loei, Bangkok, SupanBuri, Nakhonnayok, Chumphon and Songkhla) in the VA study with sample sizes.

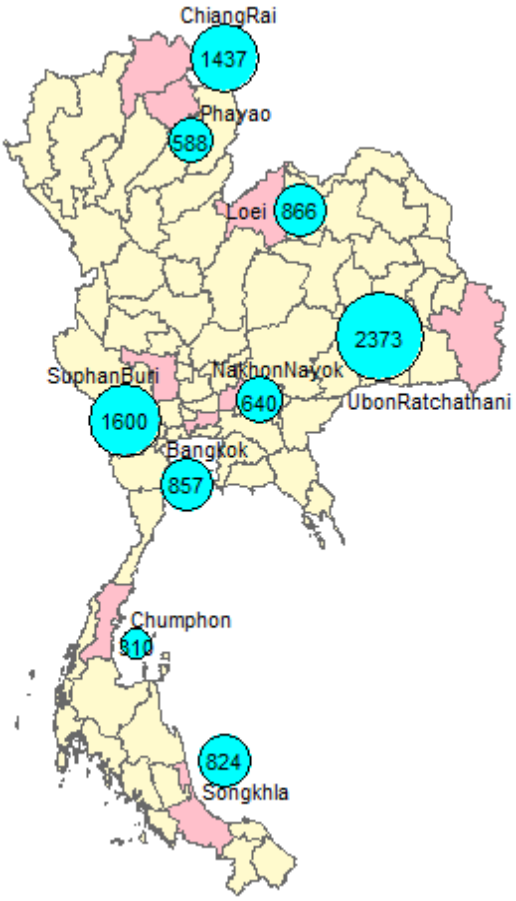


Figure 1.1 Map showing the nine provinces with sample size

Data collections on deceased persons were province, gender, age, location of death (in or outside hospitals), ICD-10) code reported on the death certificate, and VA-assessed ICD-10 code. The causes of death were grouped into 21 major causes based on the chapter-block classification of ICD-10 mortality tabulation (WHO, 2004). Groups with small counts (mainly less than 200) were combined into larger groups using medical considerations (apart from septicaemia, which received special attention due to over-reporting). Table 1.2 shows proportions of 21 major causes of deaths.

Table 1.2 Definition of cause groups

Cause of death group	Number of deaths	Percent
1:TB (A15-A19)	195	2.1
2:Septicemia (A40-A41)	77	0.8
3:HIV (B20-B24)	512	5.4
4:Other Infectious (A, B)$^{-}$	219	2.3
5:Liver Cancer (C22)	500	5.3
6:Lung Cancer^{+} (C30-C39)	320	3.4
7:Other Digestive Cancer (C15-C26)$^{-}$	290	3.1
8:Other Cancer (C^{-}, D00-D48)	697	7.3
9:Endocrine (E00-E99)	647	6.8
10:Mental, Nervous (F00-F99, G00-G99)	223	2.3
11:Ischemic (I20-I25)	617	6.5
12:Stroke (I60-I69)	1,076	11.3
13:Other CVD (I^{-})	540	5.7
14:Respiratory (J00-J99)	801	8.4
15:Digestive (K00-K93)	489	5.2
16:GenitoUrinary (N00-N99)	412	4.3
17:Ill-defined (R00-R99)	501	5.3
18:Transport Accident (V00- V99)	536	5.6
19:Other injury (W00-W99, X00-X59)	327	3.4
20:Suicide (X60-X84)	158	1.7
21:All other	358	3.8
Total	9,495	100.0

$^{+}$ *Respiratory/thoracic*, $^{-}$*exclude above*

1.6 Path diagram

The cross tabulation between VA-assessed and DR-reported ICD-10 codes were used to assess the misclassification. To correct misclassification, the DR-reported cause was

then considered as a main determinant for the VA-assessed specific cause. Other determinants are location, gender, age, and province.

For analysis of liver cancer deaths data, the binary outcome is VA-assessed ICD-10 code as deaths from liver cancer or other.

The location of death and DR-reported ICD-10 codes were combined and categorized into 16 groups: 8 DR-reported ICD-10 code groups each for the two locations (in and outside the hospitals).

Gender and age were combined and classified into 7 groups by sex: ages 5-29, 30-39, 40-49, 50-59, 60-69, 70-79 and 80+years for each sex.

Nine provinces (Bangkok, Nakhon Nayok, Suphan Buri, Ubon Ratchathani, Loei, Phayao, Chiang Rai, Chumphon, and Songkhla) were included in the VA study. Figure 1.2 shows path diagram.

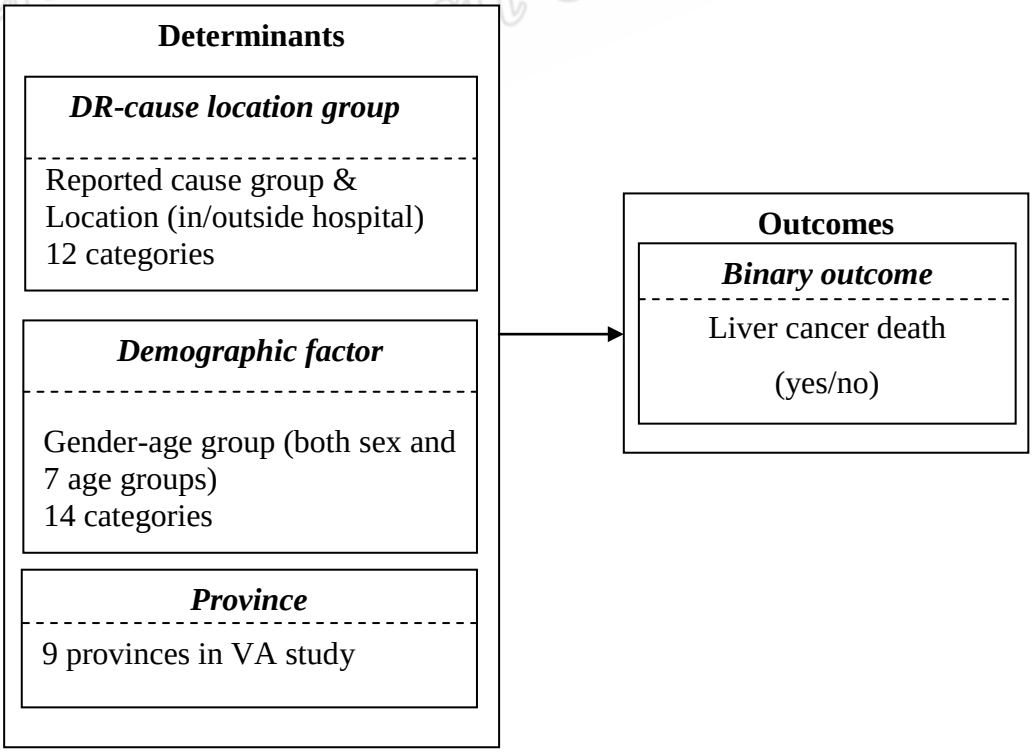


Figure 1.2 Path diagrams

Path diagrams for analysis of deaths from lung cancer and stroke were similar to liver cancer. Difference was only DR-cause location group that have different categories for liver cancer, lung cancer and stroke.

1.7 Plan of Thesis

This thesis contains four chapters. Chapter 1 as introductory chapter presents an overview of the rationale for study. Chapter 2 provides a description of the methodology. Chapter 3 states the preliminary data analysis of the three studies comprising estimating liver cancer, lung cancer, and stroke deaths based on verbal autopsy survey and estimating number of liver cancer, lung cancer, and stroke deaths in Thailand. The last chapter states the summaries and conclusions of the three studies. Limitations, suggestions, and recommendations for further study were also added.

Chapter 2

Methodology

This chapter describes methodology. Deaths at ages five year and older from the VA survey are used as a study sample. Deaths from DR database are target population. The sample data were analyzed using logistic regression model for each outcome intern including liver cancer, lung cancer, and stroke death. Deaths classified by gender, age group, province, cause and location of death, were the basis for this analysis.

2.1 Steps of analysis

The data analysis involves an issue of how to make use of the existing sample data to arrive at accurately estimate number deaths in target population. The study sample comprises 9,495 deaths aged 5 years and older in 2005 VA study. The target population comprises all reported Thai deaths aged 5 years and older in 1996-2009. A step of analysis is summarized in Figure 2.1.

The 2005 VA data for age five year and older were analysed. Logistic regression was used when the outcome is binary. Logistic regression with only one determinant is referred to as the simple model. The simple model with DR cause location as a determinant was fitted to each cause group. Then, the full model with three determinants was fitted to each cause group.

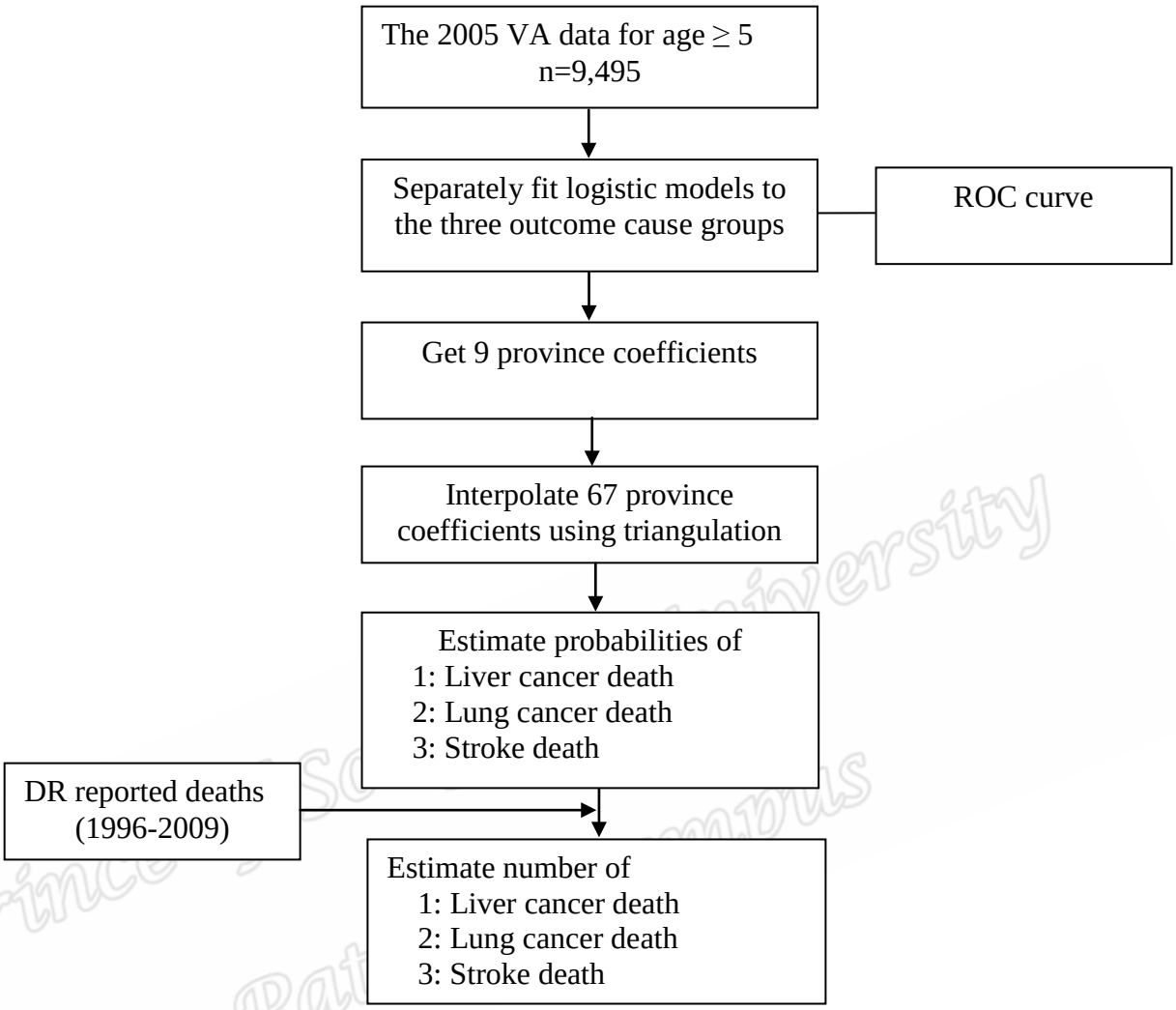


Figure 2.1 Diagram of analysis process

2.2 Logistic regression model

Logistic regression model is appropriate when the outcome takes one of only two possible values representing presence or absence of an attributes of interest. The model formulated the logit of the probability that a person died from the specific cause of death as an additive linear function of determinants (Hosmer and Lemshow 2002, Venable and Ripley 2002). The simple model when the determinant is a categorical variable is expressed as

$$\ln (p_i/(1-p_i)) = \mu + \alpha_i \quad (2.1)$$

where p_i is the probabilities of death due to the specific cause of death, μ is a constant, and α_i is the parameter of DR cause location group *i*. The simple model was compared with the full model (2.2), which includes an additive linear function of the determinant factors. The full model with three categorical determinants could be formulated as

$$\text{logit } (p_{ijk}) = \log (p_{ijk}/(1-p_{ijk})) = \mu + \alpha_i + \beta_j + \gamma_k \quad (2.2)$$

where p_{ijk} is the probabilities of death due to the specific cause of death of the *i*, *j* and *k* groups of predictor factors, μ is a constant, α_i, β_j, and γ_k refer to province, gender-age group, and cause-location group, respectively. The estimated probability of the selected cause of death can be obtained as follows:

$$p_{ijk} = 1/(1+\exp(-(\mu + \alpha_i + \beta_j + \gamma_k))) \quad (2.3)$$

The 95% confidence intervals of percentages of deaths were obtained from the model with sum contrasts (Venables and Ripley 2002; Tongkumchum and McNeil 2009; Kongchouy and Sampantarak 2010).

A receiver operating characteristic (ROC) curve assesses how well a model predicts a binary outcome. Denoting the predicted outcome as 1 (deaths due to selected cause) if $p_{ijk} \geq c$ or 0 if $p_{ijk} < c$, the ROC curve plots sensitivity against the false positive rate, as *c* varies. The ROC curve passes through the upper left corner, providing area under the curve (AUC) close to 1.

2.3 Confidence intervals

Democratic confidence intervals

Sum contrasts rather than conventional treatment contrasts was used when fitting the model. Confidence intervals based on sum contrasts has an advantage in 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. They are more appropriate compared to the corresponding confidence intervals based on the treatment contrasts. The confidence intervals compare percentage of liver cancer deaths in each category of determinant with the overall percentage. They applied equitably to each category, whereas the commonly used confidence intervals based on treatment contrasts measured the difference from a reference group that is taken to be fixed and thus does not have a confidence interval.

Democratic confidence intervals for logistic regression

Method for constructing democratic confidence intervals for logistic regression involves the process of estimating standard errors. The simple situation is comparison of two proportions in a 2 by 2 table. The notations used are described as follows.

For j = 1 or 2, let $p_j = s_j/n_j$ denotes the proportion of adverse outcomes and $r_j = n_j/n$ denotes the ratio of cases in category j. The number of successes is s_j, the sample size is n_j, $n = \sum_{j=1}^{n} n_j$ and the observed overall proportion $p = \sum_{j=1}^{2} s_j /n$. The logit of a proportion p_j takes the form $f(p_j) = \ln(p_j/(1-p_j))$. The data range is thus transformed from (0, 1) to $(-\infty, \infty)$.

Suppose that x is a binary determinant in a logistic regression model being fitted to two grouped data. The equation expressing one of the two contrasts in terms of the individual logit of proportions take the form $\boldsymbol{\alpha}^* = \boldsymbol{D_1}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the column vector containing the two logit of proportions. Solving this equation gives $\boldsymbol{\alpha} = \boldsymbol{C}_1\boldsymbol{\alpha}^*$ where $\boldsymbol{C}_1$ is the inverse of the matrix $\boldsymbol{D}_1$. The first column of $\boldsymbol{C_1}$ is omitted to obtain the desired contrast matrix $\boldsymbol{C}$, which is then specified when fitting the logistic regression.

Let $f(p)$ denote the logit of proportion p. The equations we use are as follows.

$$\alpha_1^* = r_1 f(p_1) + r_2 f(p_2) \neq f(p) \qquad (2.4)$$

$$\alpha_2^* = f(p_2) - f(p) \qquad (2.5)$$

The matrix $\boldsymbol{D_1}$ comprises equations (2.4) and (2.5) and the matrix $\boldsymbol{C}$ then takes the form $\begin{bmatrix} 1 \\ -r_1/r_2 \end{bmatrix}$ for group 1 and $\begin{bmatrix} 1 \\ -r_2/r_1 \end{bmatrix}$ for group 2. The standard errors that result when a logistic regression is fitted using $\boldsymbol{C}$ as the contrast matrix are used to obtain confidence intervals for the logit proportions after adjusted for intercept bias. The confidence interval for the omitted group was obtained by repeating the procedure with this group included and another omitted.

This enables us to construct a graph showing confidence intervals for each of the two proportions being compared, by transforming the confidence intervals for the logits back to confidence intervals for the proportions, using the inverse of the logit function, it follows that $p_j = 1/(1+\exp(-f(p_j))$ for group j.

The simple logistic regression model with the binary factor takes the additive form $\ln(p_j/(1-p_j) = a+b_j$, where a and b_j are coefficients and the proportion itself is thus

expressed as $p_j = 1/(1+\exp(-a-b_j)$, for j=1, 2. The confidence intervals for comparing proportions of group *j* are obtained from $1/(1+\exp(-\{(a^*+b_j \pm 1.96\times SE(b_j)\})$, where a^* is $f(p)$ and $SE(b_j)$ is the standard error of b_j. The constant *a* is replace by a^* to adjusted for bias due to logit of the overall proportion is not the same as the mean of the logit p_1 and logit p_2. The simple model can be extended to situation with many factors. A problem still arises because of logit of a proportion has a skewed distribution. To fix it, two constants are needed, not just one as stated in a previous study (Kongchouy and Sampantarak, 2010). A suggested model that allows for skewness takes the form $k+a_1\times b_j$ where a_1=1 if the estimated percentage is closer to the overall mean or to be determined otherwise.

Spatial triangulation method

For liver cancer, the logistic regression models gave coefficient, standard error and p-value for 9 province, 12 gender-age groups, and 16 DR-cause location groups. For the remaining 67 provinces, we used a simple "spatial triangulation method" to interpolate the coefficients. To do this, triangles were drawn linking the nine VA provinces. These triangles were set at planes, like roofs on poles which heights corresponding to their model coefficient values at the vertices of the triangles. The coefficients are estimated, based on the latitude and longitude of their central points.

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

$$\mathbf{a} + (\text{longitude}(\text{prov}_1)\times\mathbf{b}) + (\text{latitude}(\text{prov}_1)\times\mathbf{c}) = \text{coef}(\text{prov}_1) \quad (2.6)$$

$$\mathbf{a} + (\text{longitude}(\text{prov}_2)\times\mathbf{b}) + (\text{latitude}(\text{prov}_2)\times\mathbf{c}) = \text{coef}(\text{prov}_2) \quad (2.7)$$

$$\mathbf{a} + (\text{longitude}(\text{prov}_3)\times\mathbf{b}) + (\text{latitude}(\text{prov}_3)\times\mathbf{c}) = \text{coef}(\text{prov}_3) \quad (2.8)$$

The coefficient for any province *j* within a triangle was now given by equation (2.9) as follows:

$$\textbf{coef(prov}_j\textbf{)} = a + (\text{longitude}(\text{prov}_j)\times b) + (\text{latitude}(\text{prov}_j)\times c) \quad (2.9)$$

Coefficients for provinces outside triangles are obtained similarly by extrapolation.

Correcting causes of death

The estimated probabilities of cause specific deaths (liver cancer, lung cancer and stroke) from the models were applied to total number of reported deaths in the nine provinces by gender-age group and DR cause-location from the DR data in 2005. The numbers of deaths for cause specific death in the nine provinces in 2005 were obtained after adjusting for misreporting based on the simple and full models.

The coefficients for province, gender-age group and DR cause-location were used to estimate proportions and numbers of deaths. Thus, the VA-estimated number of deaths in 2005 was obtained.

Assuming the models were correct for years before and after 2005, they were extended to annual deaths in the DR database for years 1996-2009. Thus, the VA-estimated deaths from 1996 to 2009 for liver cancer, lung cancer and stork were obtained.

Similarly, the resultant cause specific proportions of liver cancer and stroke deaths were then intern applied to the annual numbers of registered deaths by provinces by gender-age group and DR cause-location during that period.

2.4 Methods for presenting results

Crude percent of deaths for each cause were presented using bar plot superimposed with their corresponding 95% confidence intervals from the model. To show geographical variation of mortality, thematic and range maps were used.

Confidence interval plot

The model results were presented using confidence interval plot. A 95% confidence interval plot was used to show the pattern of proportion for each factor level adjusted for other factors from the model. The confidence intervals were constructed based on standard errors of differences between the proportion and its overall mean, graphed as a vertical line containing a dot denoting the adjusted proportion deaths.

Thematic map

A thematic map was used to show geographical variation in mortality. It shows one province in dark shade to indicate that the province has high percent of deaths, while showing another province in light shade to indicate that the province has low percent.

Range map

A range map displays data according to specified data ranges. The ranges are shaded with difference colors. Ranges maps are used to show the geographical distribution of province coefficient and percent of deaths base on model. Ranges map of province coefficients and percent of death were divided into three groups based on the quartile.

The graphical displays and statistical analyses were performed using R program version 3.0.1 (R Core Team, 2013)

Chapter 3

Results

In this chapter, we applied methods described in Chapter 2 to the data focusing on deaths from liver cancer, lung cancer, and stroke. The chapter begins with analysis of VA data and follows by extension to DR data. First, we construct cross-tabulation showing misclassification causes of deaths between DR and VA. Second, we show number of cases of the VA sample when liver cancer is chosen as a binary outcome classify by three determinants. Third, we show confidence intervals of proportion of deaths and ROC curve from liver cancer, lung cancer, and stroke. Geographical variations of proportion of deaths by three causes were also presented using range map. Finally, number of estimated deaths for liver cancer, lung cancer, and stroke was presented using area plots.

3.1 Analysis of VA data

The assessed causes of deaths (VA group) were crossed check with the reported causes (DR group) and bubble plot was used to display results.

Figure 3.1 shows cross tabulation between DR and VA groups. Bubble size represents number of deaths. Assuming that assessed VA cause group correct, misclassification was observed among 21 causes. Most of the causes including liver cancer, lung cancer, and stroke are under-reported whereas septicemia, ill-defined, and all other were over-reported.

Assessed deaths from liver cancer in the VA group are 500 cases whereas reported deaths from liver cancer in the DR group are 281 cases. There are 236 cases in agreement (47.20%). Liver cancer deaths were most likely to have reported causes as liver cancer (236), ill-defined (97), digestive disease (49), other cancer (48), other digestive cancer (39), lung cancer (7), GU (7), and other (17, all of the rest comprising 14 causes).

Deaths from lung cancer in the VA group are 320 cases whereas deaths from lung cancer in the DR group are 218 cases. There are 164 cases in agreement (51.25%). Lung cancer deaths were most likely to have registered causes as lung cancer (164), ill-defined, other cancer, respiratory, and other (all the rest comprising 17 causes).

Deaths from stroke in the VA group are 1,076 cases but deaths from stroke in the DR group are 364 cases. There are 267 cases in agreement (24.81%). Stroke deaths were most likely to have registered causes as stroke (267), ill-defined (535), septicemia, respiratory, mental and nerve, other CVD, endocrine, all other, and other (all of the rest comprising 13 causes).

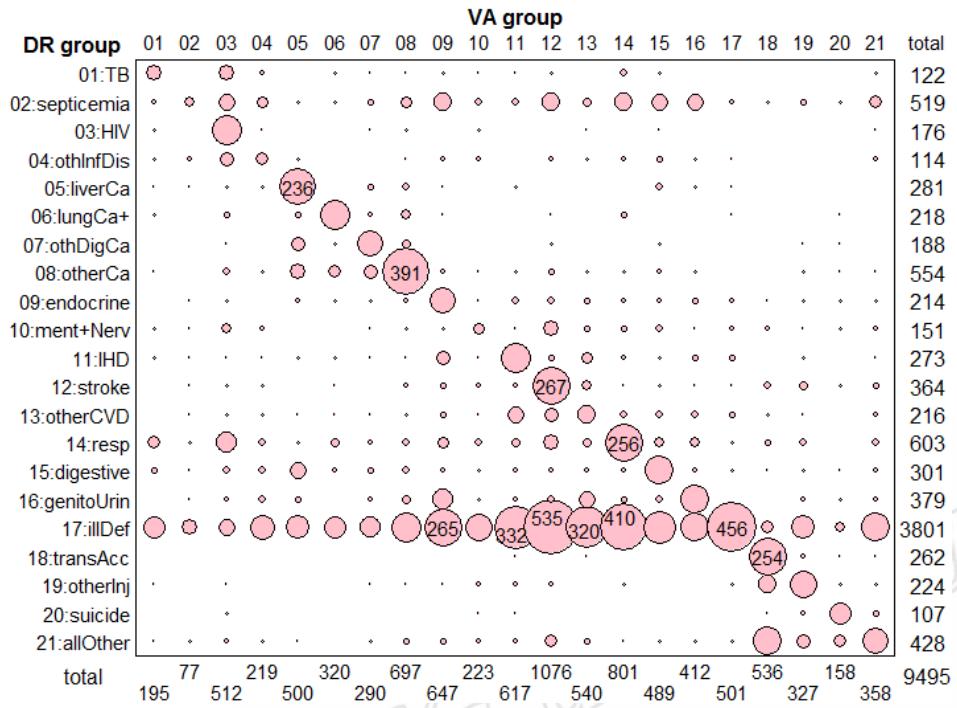


Figure 3.1 Cross tabulation of 21 major cause groups between DR and VA

The sample size is 9,495 cases. Number of cases of the VA sample classified by liver cancer deaths (yes/no), province, gender and age group, and DR cause location group are shown in Table 3.1. On average deaths from liver cancer was 5.27%. About 9.69% and 7.99% had places of residence as Ubonratchatani and Phayao provinces, respectively. It was 11.9% occurred in males aged 50-59. Among DR reported liver cancer 85.29% in hospital and 83.57% outside hospital were correctly due to liver cancer. Among DR ill-defined cause 1.38% in hospital and 2.7% outside hospital were assessed as liver cancer. Among DR reported digestive disease 1.85% in hospital and 33.09% outside hospital were assessed as liver cancer. Among DR reported other cancer 0.63% in hospital and 11.9% outside hospital were assessed as liver cancer. Among DR reported other digestive cancer 15.48% in hospital and 25% outside

hospital assessed as liver cancer. Among DR reported lung cancer 0.93% in hospital and 5.45% outside hospital assessed as liver cancer.

Table 3.1 VA sample classified by liver death/other, province, gender-age group, DR cause group and location (in-hospital and out-hospital)

Determinants	Deaths in VA survey		Determinants	Deaths in VA survey	
	liver (n=500)	other (n=8,995)		liver (n=500)	other (n=8,995)
Province			DR group and location		
Bangkok	31	826	inH-liver cancer	58	10
NakhonNayok	17	623	outH-liver cancer	178	35
UbonRatchatha	230	2143	inH-ill-defined	6	429
Loei	49	817	outH-ill-defined	91	3275
Phayao	47	541	inH-digestive disease	3	159
ChiangRai	66	1371	outH- digestive disease	46	93
SuphanBuri	30	1570	inH-other cancer	1	158
Chumphon	8	302	outH-other cancer	47	348
SongKhla	22	802	inH-other digestive	13	71
Gender age group			outH- other digestive	26	78
M:5-39	16	1035	inH-lung cancer	1	107
F:5-39	4	419	outH-lung cancer	6	104
M:40-49	50	565	inH-GU	1	159
F:40-49	21	270	outH-GU	6	213
M:50-59	85	625	inH-other group	2	2034
F:50-59	33	387	outH-other group	15	1722
M:60-69	95	862			
F:60-69	50	634			
M:70-79	64	1075			
F:70-79	33	1001			
M:80+	26	882			
F:80+	23	1240			

Simple logistic regression model was first fitted to the data with DR cause location group as the only determinant. Then, full model with three determinants was fitted to the data. The same analysis was performed for deaths from lung cancer and stroke.

Among nine provinces, Ubon Ratchathani had the largest number of total deaths (2,373), while Chumporn had the lowest (310). The VA- assessed deaths gave 500 liver cancer deaths whereas only 236 liver deaths (47.20%) were correctly DR- reported.

Figure 3.2 shows percentage of deaths from liver cancer, lung cancer, and stroke deaths in the top, middle, and bottom panels, respectively. A bar chart shows crude percentage whereas 95% confidence intervals show adjusted percentages obtained from the full models using weighted sum contrasts. A red horizontal line shows an average. To distinguish the bar chart and 95% confidence interval, a non-linear vertical axis scale was used. The values derived from the VA assessment and from the model are similar indicating no confounding and no effect of transformation back from logit to probability. The confidence intervals above the average line reflect the groups that were more likely to die from liver cancer.

DR-cause location group was significant in the simple model and all three determinants were significant in the full model. The 95% confidence intervals for both Ubonratchatani and Phayao were higher than average, whereas for SuphanBuri it is lower. Therefore, effect of province on misclassification is observed. Ubonratchatani and Phayao provinces were most likely to have high level of liver cancer death. Males aged 40–69 were more likely to have high level of liver cancer deaths. Therefore, these age groups were likely to have high levels of under-reporting. Some of liver cancer deaths were reported as digestive disease (outside hospital), other cancer (outside

hospital), and other digestive cancer. These are the group in which liver cancer deaths were often misclassified.

The VA- assessed deaths gave 320 lung cancer deaths whereas only 164 lung cancer deaths (51.25%) were correctly DR- reported.

DR-cause location group was significant in the simple model and all three determinants were significant in the full model. Therefore, effect of province on misclassification was not observed. Although the province was not significant we still keep it in the model as basis for estimating lung cancer deaths for every province in the country. Males aged 60–79 were more likely to have high level of lung cancer deaths. Therefore, these age groups were likely to have high levels of under-reporting. Some of lung cancer deaths were reported as other cancer (outside hospital). These are the group in which liver cancer deaths were often-misclassified.

The VA- assessed deaths gave 1,076 stroke deaths whereas only 267 stroke deaths (25.02%) were correctly DR- reported.

DR-cause location group was significant in the simple model. The DR-cause location and gender-age-group were highly statistically significant, but there was no significant evidence of province effect in the full model. The 95% confidence intervals for Bangkok, SuphanBuri, and Songkhla were higher than average, whereas for Ubonratchatani and Chiangrai they are lower. Therefore, effect of province on misclassification was observed. Males aged 60–79 and females aged above 70 were more likely to have high level of stroke deaths. Therefore, these age groups were likely to have high levels of under-reporting. Substantial numbers of stroke deaths were

reported as mental and nerve, other cardiovascular diseases (outside hospital), and other cause (in hospital). These are the group in which stroke deaths were often misclassified.

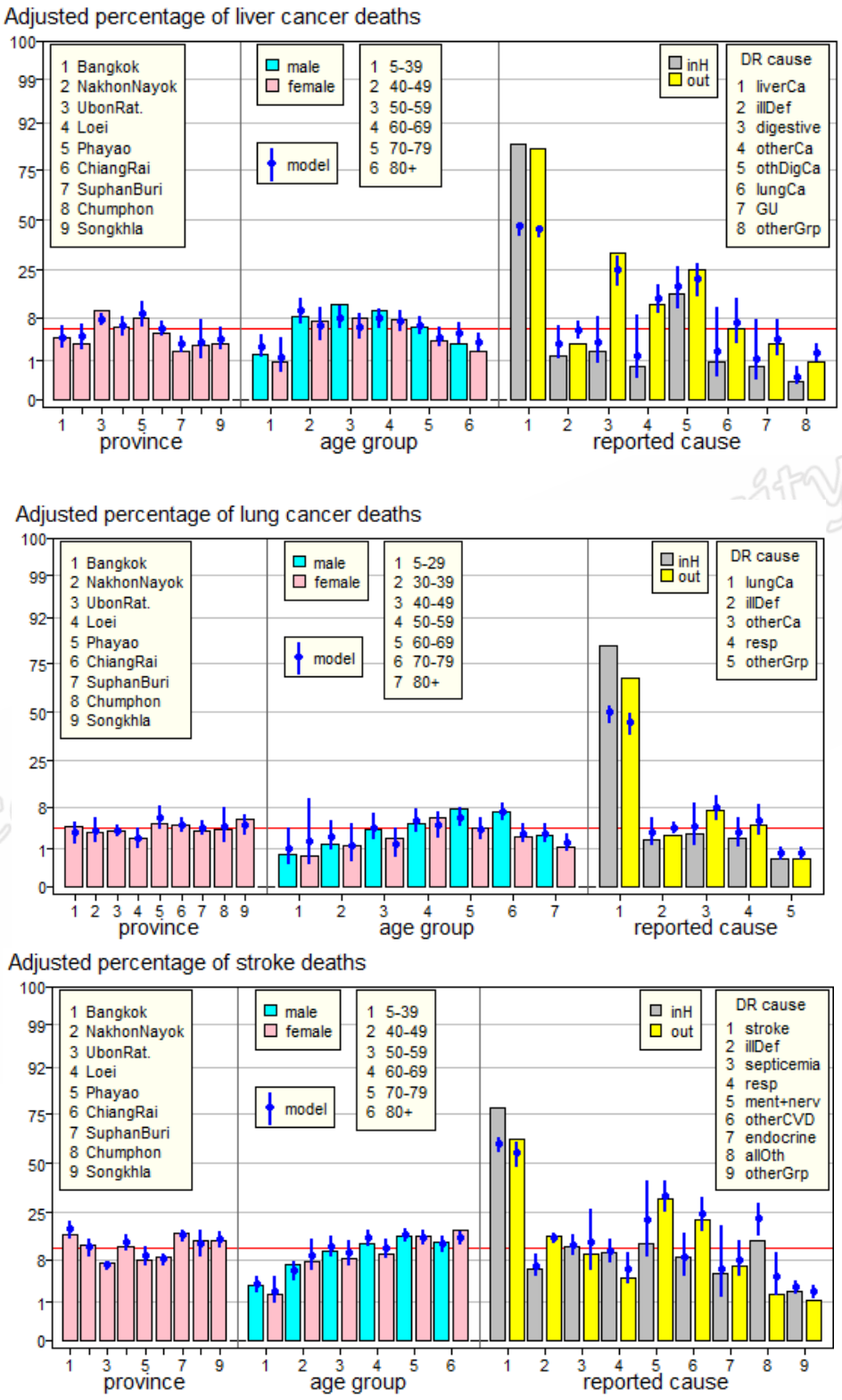


Figure 3.2 Crude and adjusted percentage of liver cancer, lung cancer and stroke deaths

The full model with three determinants (province, gender-age group, DR-cause location) were assessed using the ROC curve and compared with simple cross-referencing model with only one determinant (DR-cause location). The ROC cureve shows how well a model predicts a binary outcome. The cut-off point gives a total predicted number agreement of the number of VA-assessed liver cancer deaths. Figure 3.3 shows ROC cure for liver cancer deaths. The full model for liver cancer gives 62.6% sensitivity, 97.9% specificity, and an AUC of 0.84, whereas the simple model gives 56.4% sensitivity, 98.5% specificity, and an AUC of 0.8. The full model reduced the error from 20% to 16%. The full model has the ability to predict the correct cause of liver cancer death slightly better than the simple cross-referencing model.

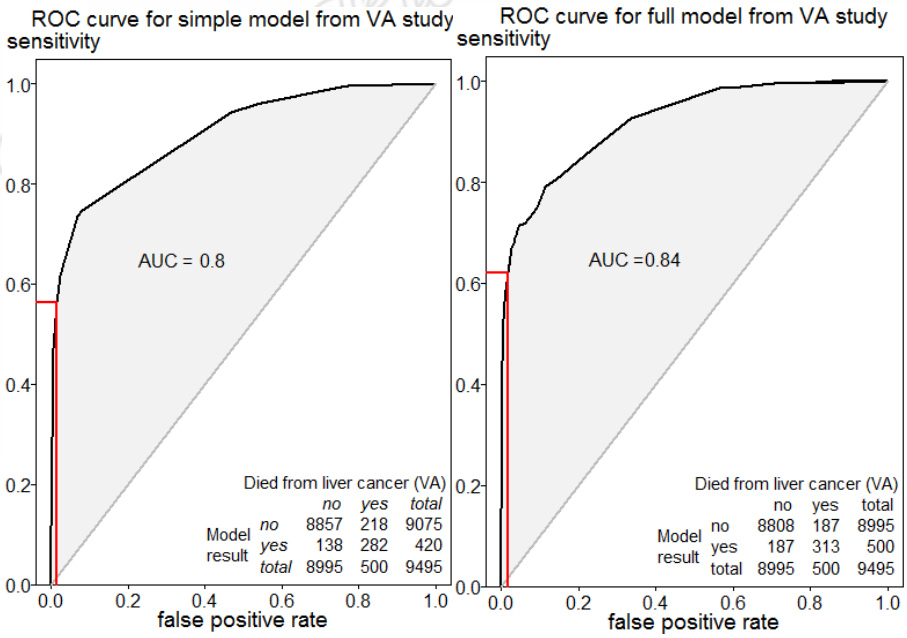


Figure 3.3 ROC curve for simple and full model of liver cancer deaths

Figure 3.4 shows ROC cure for lung cancer death. The full model for lung cancer gives 55.3% sensitivity, 98.5% specificity, and an AUC of 0.79, whereas the simple model gives 51.2% sensitivity, 99.4% specificity, and an AUC of 0.72. The full model reduced

the error from 28% to 21%. The full model has the ability to predict the correct cause of liver cancer death slightly better than the simple cross-referencing model.

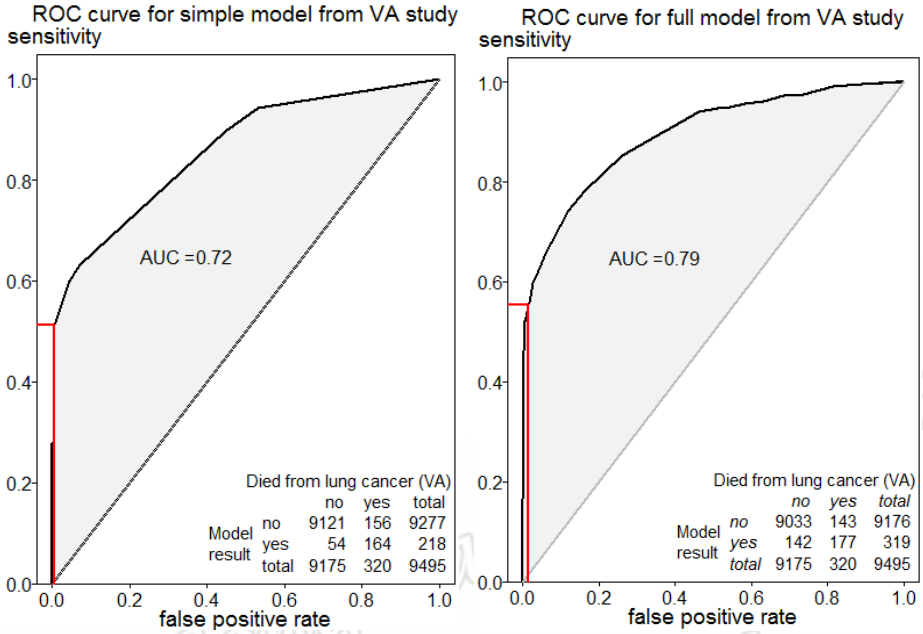


Figure 3.4 ROC curve for simple and full model of lung cancer deaths

Figure 3.5 shows ROC curve for stroke death. The full model for stroke gives 41.6% sensitivity, 92.6% specificity, and an AUC of 0.67, whereas the simple model gives 30.9% sensitivity, 96.7% specificity, and an AUC of 0.57. The full model reduced the error from 43% to 33%. The full model has the ability to predict the correct cause of liver cancer death better than the simple cross-referencing model.

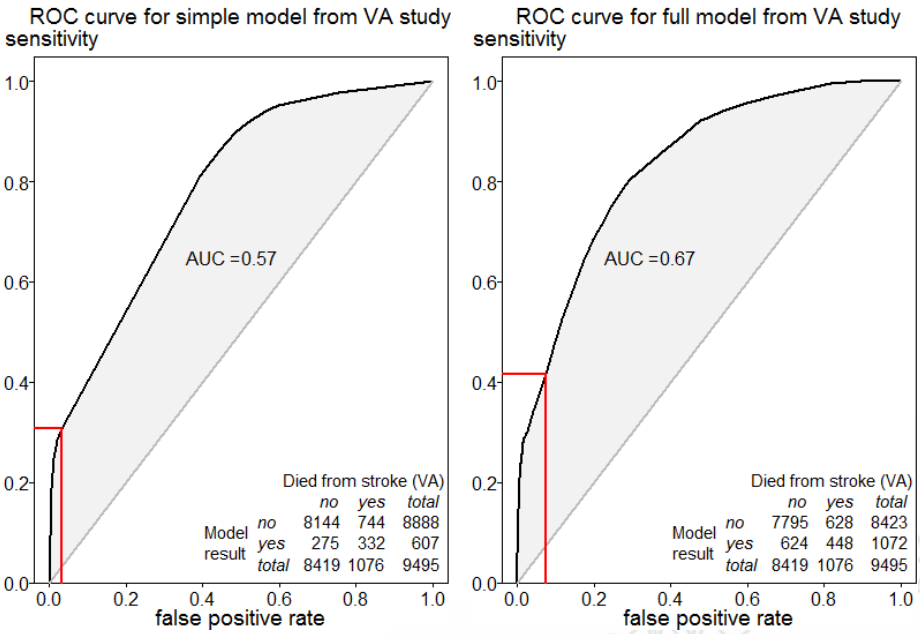


Figure 3.5 ROC curve for simple and full model of stroke deaths

3.2 Interpolation of province coefficients

The 9 province coefficients were obtained from the full model whereas remaining 67 province coefficients were obtained from interpolation using triangulation method as described in Chapter 2. Figures 3.6-3.8 show province coefficients for liver cancer, lung cancer, and strokes deaths.

Substitute the coefficients of DR cause location to Equation 2.2, estimated probabilities of death from simple model for each category of DR cause location were obtained. Similarly, substitute the coefficients of province, gender age group, and DR cause location to Equation 2.4, estimated probabilities of death from full model for each combination of determinants were obtained. The probabilities were then applied to total deaths in DR data.

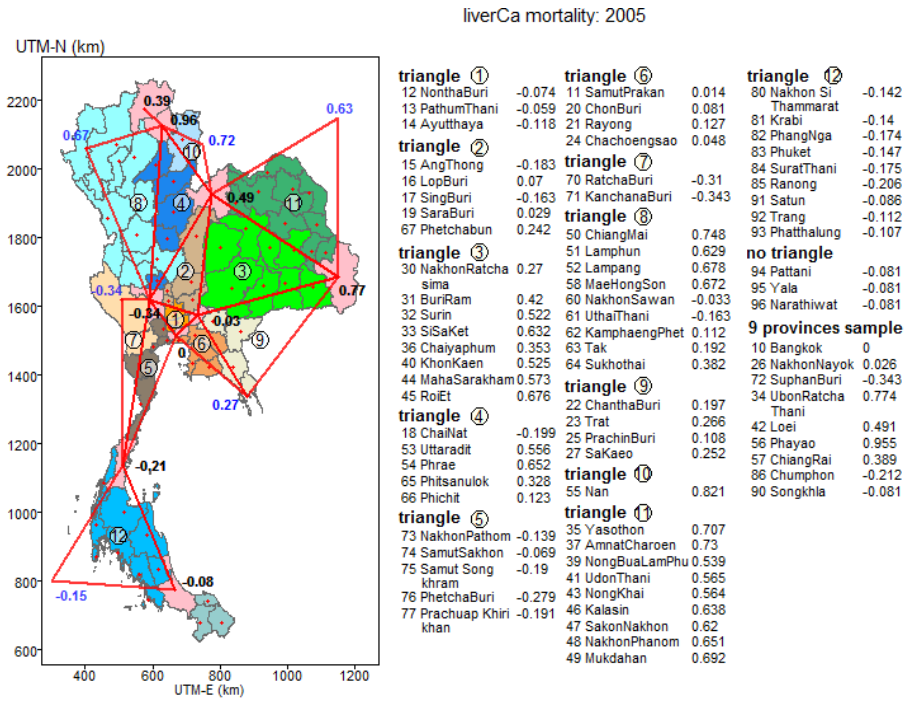


Figure 3.6 Province coefficients of liver cancer model

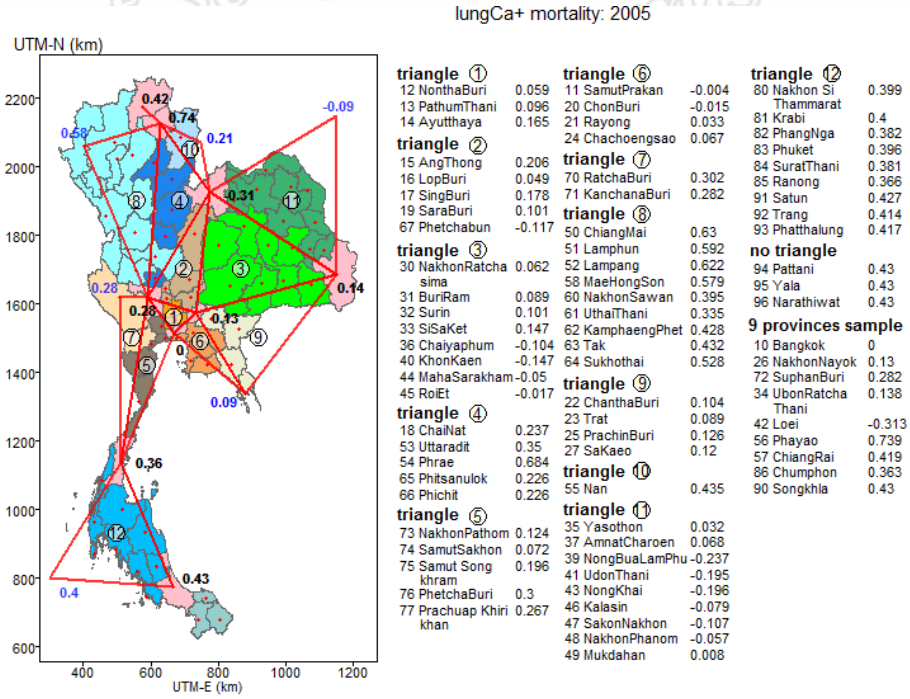


Figure 3.7 Province coefficients of lung cancer model

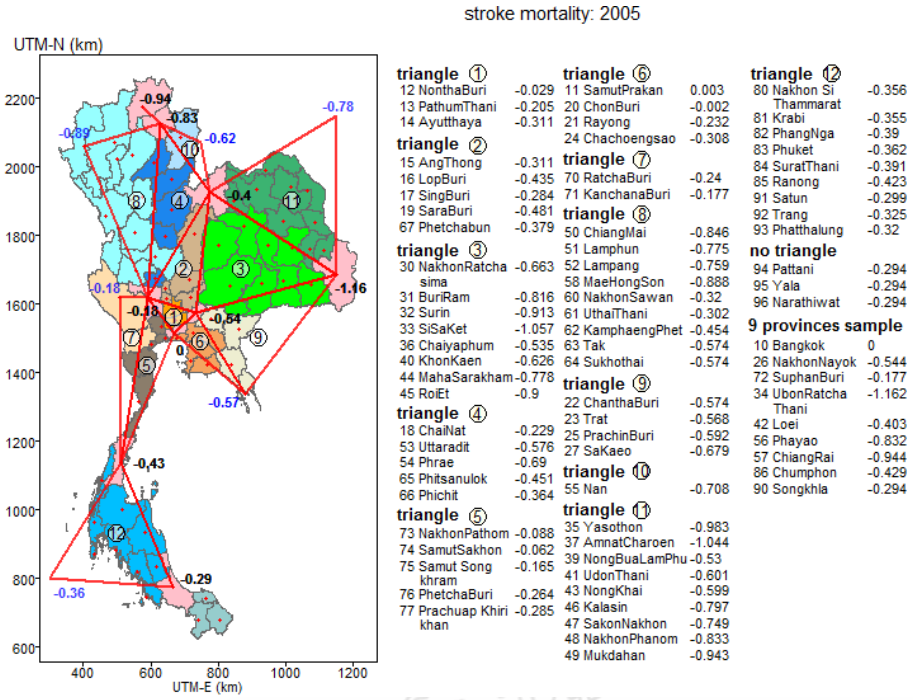


Figure 3.8 Province coefficients of stroke model

3.3 Extension to DR data

The adjusted percentages obtained from Equation 2.2 based on the coefficients of gender-age group, DR-cause location, and provinces were applied to total death for each combination of determinants in the DR data in 2005 to get the estimated numbers of deaths. Figure 3.9 shows area plot of estimated numbers of deaths when applied adjusted percentage to DR data in 2005. The numbers of reported deaths and numbers estimated deaths from simple model were also shown. The estimated total numbers of liver cancer deaths in 2005 were 13,975 (males) and 6,388 (females). The inflation factor are 1.58 and 1.75, that higher than the reported totals of 8,869 and 3,641, respectively.

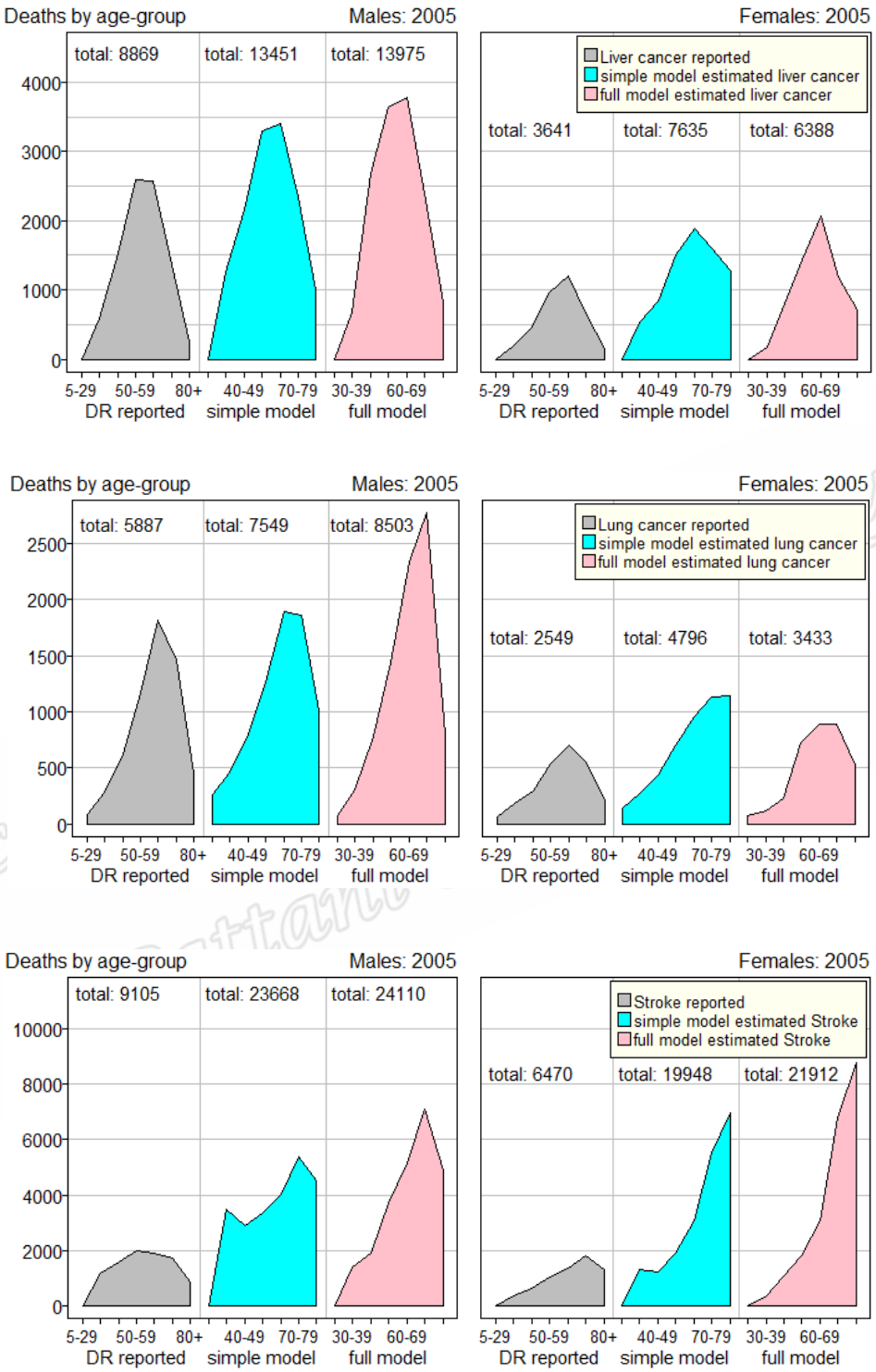


Figure 3.9 Reported and estimated deaths from simple and full models in 2005 for liver cancer (top panel), lung cancer deaths (middle panel) and stroke (bottom panel)

The adjusted percentages from the simple and full models were then applied to DR data in 1996-2009. Figure 3.10 shows area plots of DR reported, simple model estimated and full model estimated number of deaths from liver cancer, lung cancer, and stroke by gender-age groups in 1996-2009. The area of each colour strip denotes the number of deaths in each age group.

The total numbers of liver cancer deaths reported for 14 years are 147,158. The estimated total numbers of liver cancer deaths from the simple and full models are 260,508 and 249,922, respectively. The total number of DR reported liver cancer deaths were lower than those estimated by simple model and full model by factors of 1.77 and 1.69, respectively. While simple model gave large proportions of liver cancer deaths at ages below 40 years, these were reduced when full model allowing for province, gender and age was used. For the older age groups, cause of liver cancer deaths was already improved in accuracy by the simple model.

When separated by gender, the total number of DR reported liver cancer deaths were lower than those estimated by simple model and full model by factors of 1.58 and 1.64 for males. For females the total number of DR reported liver cancer deaths were lower than those estimated by simple and full models by factors of 2.19 and 1.83, respectively.

Similar to liver cancer, the total numbers of lung cancer deaths reported for 14 years are 93,310. The estimated total numbers of lung cancer deaths from the simple and full models are 148,029 and 140,651, respectively. The total number of DR reported lung cancer deaths were lower than those estimated by simple model and full model by factors of 1.77 and 1.69, respectively.

For males the simple model gave large proportions of lung cancer deaths at ages below 40 years, these were reduced when full model was used. For the older males, cause of lung cancer deaths was already improved in accuracy by the simple model.

For females the simple model gave large proportions of lung cancer deaths at ages below 50 years, these were reduced when full model was used. For the older females, cause of lung cancer deaths was already improved in accuracy by the simple model. The total number of DR reported liver cancer deaths were lower than those estimated by simple model and full model by factors of 1.38 and 1.54 for males. For females the total number of DR reported liver cancer deaths were lower than those estimated by simple model and full model by factors of 2.04 and 1.44.

For stroke, the total numbers of deaths reported for 14 years are 157,537. The estimated total numbers of stroke deaths from the simple and full models are 549,653 and 570,245, respectively. The total number of DR reported stroke deaths were lower than those estimated by simple model and full model by factors of 3.49 and 3.61, respectively.

The simple model gave large proportions of stroke deaths at ages below 40 years, these were reduced when full model was used. For the older age groups, cause of stroke deaths was already improved in accuracy by the simple model. The total number of DR reported stroke deaths were lower than those estimated by simple model and full model by factors of 3.28 and 3.27 for males. For females the total number of DR reported stroke deaths were lower than those estimated by simple model and full model by factors of 3.78 and 4.10.

The area plots for liver cancer, lung cancer and stroke deaths clearly reveal that numbers of deaths were under reported especially for the earlier years. Similar patterns are seen with number of deaths increase in recent years.

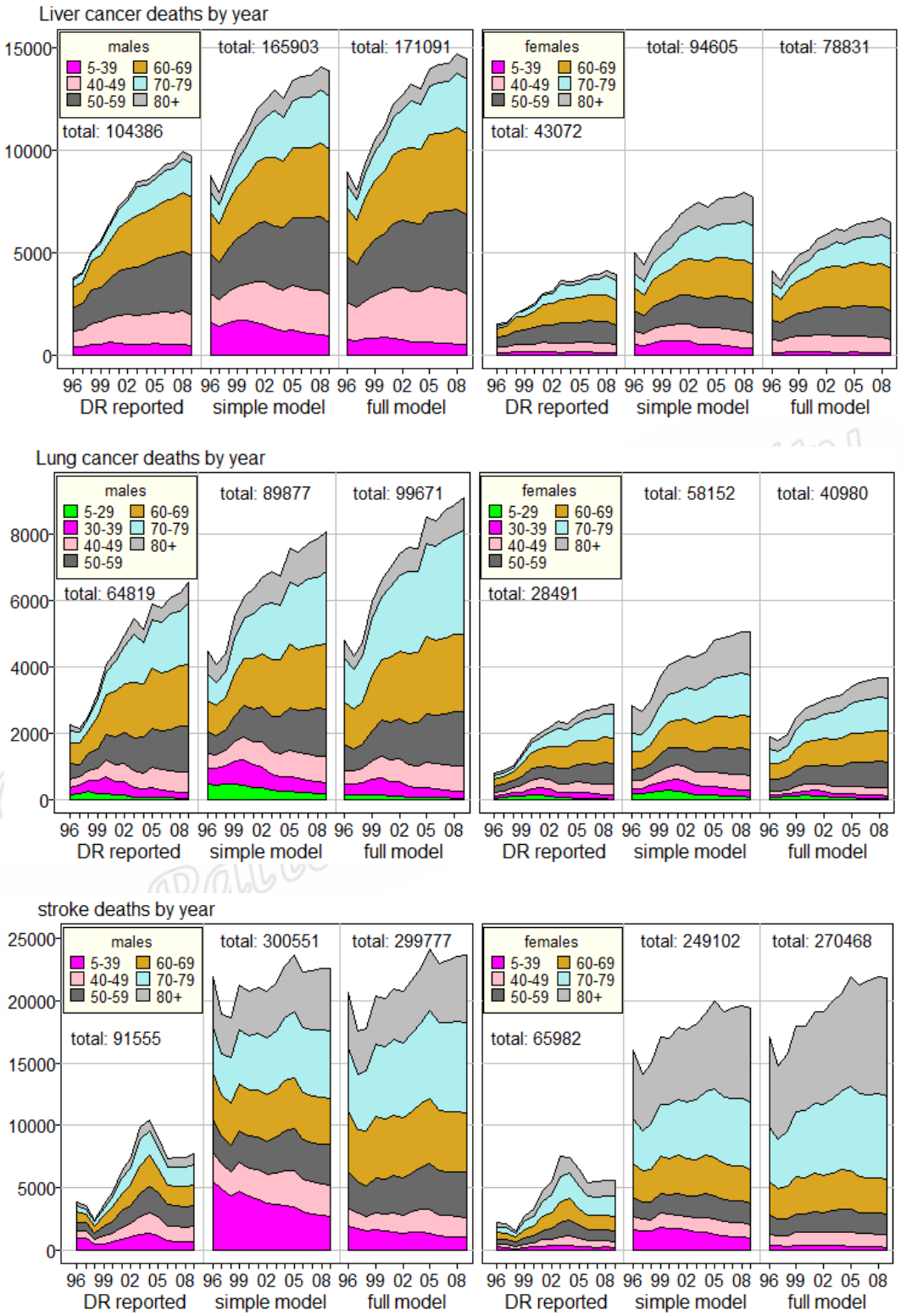


Figure 3.10 DR reported, simple model estimated and full model estimated of liver cancer, lung cancer, and stroke deaths by gender-age groups in 1996-2009

The estimated numbers of death for each cause from 1996 to 2009 were used to calculate percentage of death for each province. Figure 3.11 shows estimated percentages of deaths from liver cancer, lung cancer, and stroke for 1996, 2000, 2005, and 2009. The percentages were ranked as low, medium and high. The dark color represents high percentage of mortality for each cause. The graph shows that the regional patterns do not change much, even though the numbers of deaths have substantial trends,

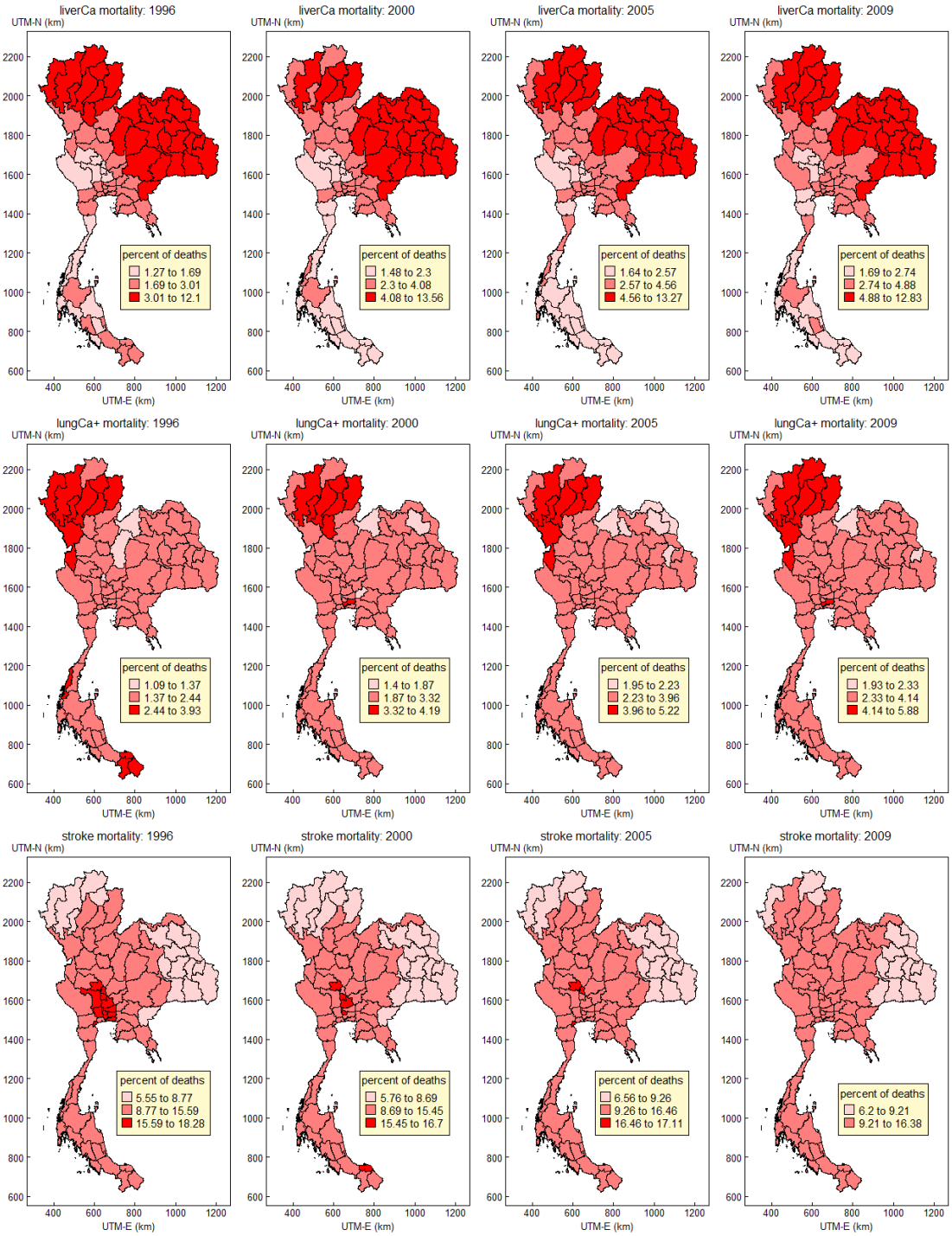


Figure 3.11 Range maps of percentages of estimated deaths from liver cancer, lung cancer and stroke by region in 1996, 2000, 2005 and 2009

Chapter 4

Discussions and Conclusions

Death due to 21 causes was considered as a multinomial outcome. Since the outcome is a nominal variable with 21 levels, the appropriate model for systematic analysis of death by ICD10 code is multinomial regression. However, it is simpler and more informative to separately fit logistic regression models to each outcome cause groups, and then rescale the results to ensure that the total numbers of deaths estimated for each group match those reported in the corresponding populations.

This study described the methods based on logistic regression for correcting misreported cause of death and illustrated the methods using death from liver cancer, lung cancer, and stroke. The data were from 2005 VA sample.

Separate logistic regression model of VA-assessed causes of deaths with demographic factors is found to be appropriate to use. It allowing for gender-age, province, and DR cause location predicted causes specific deaths with higher sensitivity and specificity compared to those derived from simple model (simple cross-referencing method). The models do not ensure that adjusted death counts in each year match reported totals because they aggregate results from separate logistic regression models.

Although multinomial model ensures that adjusted and reported totals match, simply scaling totals from separate logistic models gives similar results. Advantage of using logistic regression model to analyse data is that it can handle general determinants. The logistic function has many desirable properties. Its range is between 0 and 1 when the

independent variable varies from -∞ to ∞, so the logistic regression model can be used to model the probability of an individual death. In addition, logistic regression can control confounding and assess interaction very effectively when there are several confounders or the confounder is a continuous variable (Hosmer and Lemshow, 2000). Moreover, it can be used to calculate an odds ratio and its confidence interval directly, so that the results can be interpreted easily. The probability of a given subject death form a specific disease can also be calculated.

Another advantage in using the logistic regression model is that it gives confidence intervals for adjusted percentages of specific causes of deaths for levels of each risk factor adjusted for other risk factors. These confidence intervals, when compared with the bar charts of sample percentages, provide evidence of confounding bias.

Model based on sum contrasts (Tongkumchum and McNeil, 2009; Kongchouy and Sampantarak, 2010) provided democratic confidence intervals (DCI) of adjusted percentages of specific deaths. The DCI 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. They applied equitably to each category, whereas the commonly used confidence intervals based on treatment contrasts measured the difference from a reference group that is taken to be fixed and thus does not have a confidence interval.

The model can be simply extended to the larger target population comprising all deaths in Thailand for longer periods of time and other populations in similar context.

4.1 Discussions

Methodology

To reduce costs from conducting VA study for the whole country, we proposed an analysis of the VA data using appropriate statistical methods to a large-scale VA study, for example, in case of HIV (Chutinantakul *et al*., 2014), transport accident (Klinjan *et al*., 2014) and liver cancer (Waeto *et al.*, 2014).

The importance of evaluating the reliability and validity of causes of death in mortality statistics has long been recognized in public health (Moriyama 1989). Periodic validation of the quality of diagnostic information ensures that countries have a more confident basis on which to develop their policies and guide health planning (Khosravi *et al.*, 2008). VA survey is generally the most reliable method to determine causes of death (Lahti and Penttilä 2001; Rao *et al.*, 2007; Rao *et al.*, 2010). However, conducting a survey is expensive and time consuming. It is important for public authorizes to pay attention on quality of death registry rather than conducting verbal autopsy.

This method enables public health researchers to estimate percentages of specific causes of deaths in countries where there is low quality for recorded cause of deaths but reliable sample data such as a VA study should be available.

Liver cancer mortality

DR under reported liver cancer by a factor of 1.64 for males and 1.83 for females.

Liver cancer deaths were common among males aged 40-69, in agreements with other studies (Jemal *et al*., 2010). High proportionate among females aged 60-69 was also observed.

Liver cancer deaths were most common in upper north and northeast of the country. It is in agreement with others (Jemal *et al*., 2010; Sripa *et al*., 2007; Viratroumanee *et al*., 2009) that observed high incidence rates in the northeast. In particular, the overall ratio of mortality to incidence is almost one, meaning that the higher incidence rate indicating the higher mortality. The geographical inequality of liver cancer in Thailand (Faramnuayphol *et al*., 2008) was also supported our finding.

The mis-reported of liver cancer deaths in our study and in the previous report of the 2005 VA data using cross-referencing method are in agreement with our study is slightly higher (Pattaraarchachai *et al*., 2010). Most misclassifications of liver cancer deaths were classified as other digestive cancer or digestive disease (outside hospitals) or other cancers (outside hospitals).

Lung cancer mortality

DR is under reported lung cancer deaths by a factor of 1.54 for males and 1.44 for females. It suggested that lung cancer deaths were being diagnosed reasonably accurately in death registry. The simple cross-referencing method distorted age distribution and could lead to understanding that lung cancer was common among the elderly especially for females. This finding agrees other study, where lung cancer deaths were observed not to contribute significantly to ill-defined cancer coding (Porapakkham *et al*., 2010).

Lung cancer deaths were common among males aged 60-79, in agreements with other studies. Lung cancer was found to be common among Thai patients aged 50 years or more (Deesomchok *et al*, 2005).

No evidence of regional effect is found in this study, but the study on cancer control in Thailand using cancer registration data has found high incidence rates of lung cancer in the northern region (Vatanasapt *et al*., 2002). A geographical variation on lung cancer deaths in 2000 also has 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.

Most misclassifications of lung cancer deaths were classified as other cancer (outside hospital).

Stroke mortality

Death registry is substantially under reported stroke deaths by a factor of three for males and a factor of four for females. The simple cross-referencing method distorted age distribution and could lead to understanding that stroke was common among young adults especially for males.

Logistic regression model of stroke showed that VA-assessed stroke deaths were more likely in elderly compared to death registry, but many of those deaths were registered as deaths from mental and nervous system disorders or other cardiovascular diseases (outside hospital) or other cause (in hospital). Under reporting were more common in central and southern regions of the country.

Stroke deaths were common among deaths in the age group 50 or more in males, and 60 or more in females. Our findings of high proportionate of stroke deaths in elderly are not surprising. According to the Thailand Epidemiology Stroke (TES), high crude prevalence among adults aged 75 years and above was observed (Hanchaiphiboolkul *et al.*, 2011). An average age of stroke onset was 65 years (Suwanwela, 2014).

Proportion of stoke deaths varies with province. Stroke deaths were most common in central and southern region. Geographical variation on stroke deaths in 2000 also has been reported of high standardized mortality ratio (SMR) in Bangkok and lowest in upper north-eastern region (Faramnuayphol *et al.*, 2008). In addition, Thailand Epidemiology Stroke (TES) found stroke prevalence by regions was highest in Bangkok and lowest in north-eastern region (Hanchaiphiboolkul *et al.*, 2011).

This study found high percentages of stroke deaths especially deaths outside hospitals. Some under reported stroke deaths as the cause because of coding error such as stroke was ascribed to other CVD (Brown *et al*., 2007). The percentage of estimated stroke deaths was highest at aged 60 and over whereas the percentage of DR reported stroke was highest at aged 50-59, especially the inflation factor in elder is higher than other age group.

4.2 Conclusions

The unreliable cause of death from death registration database in countries like Thailand necessitates extensive adjustment to the data in order to derive plausible liver cancer mortality by gender, age and regions or provinces or districts. The data with more reliable cause of deaths from well-designed research such as the VA study (Rao *et al.*, 2010; Pattaraarchachai *et al*., 2010; Polprasert *et al*, 2010; Porapakkham *et al*.,

2010) together with appropriate statistical methods are very useful for making adjustment to imperfect registration data. This study reported the utility of statistical methods in analysing existing data to derive estimates of liver cancer deaths in Thailand from 1996 to 2009.

The statistical methods used in this study can be applied to available mortality data in middle income countries where their national vital registration data are of low quality and supplementary reliable data are available.

This method enables public health researchers to estimate percentages of specific causes of deaths in countries where there is low quality for recorded cause of deaths but reliable sample data such as a VA study should be available.

4.3 Limitations

This study has some limitations to be addressed.

Firstly, 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 years before 2005 when reporting practices were distorted by the HIV epidemic.

Secondly, we fixed the province effect rather than make it random across region. This may lead to slightly lower in standard error because VA study was based on cluster sampling. Cluster sampling gave standard error larger than simple random sampling (Lumley, 2010)

Thirdly, a bias may arise from 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, which located on the deep southern part of Thailand.

Lastly, this study assumed annual total death is accurate. Thus, we did not adjust for undercount of death registration, which yield the true total number of death. This assumption may lead to under estimate of mortality.

4.4 Recommendations for further study

Our findings suggest a substantial misclassification of stroke deaths but not so high for liver cancer and lung cancer mortality in the Thai population. Therefore policy makers who determine research and treatment priorities should pay more attention to quality of death registry.

The use of standard VA methods adapted to Thailand enabled a plausible assessment of cause-specific mortality patterns and a substantial reduction of ill-defined diagnoses. Validation studies enhance the utility of findings from the application of verbal autopsy. Regular implementation of well design VA in Thailand could accelerate development of the quality and utility of death registration data especially for deaths outside hospitals.

References

Bureau of Policy and Strategy. 2010. Public Health Statistic 2010. Nonthaburi: Ministry of Public Health. Thailand.

Byass, P. 2010. Integrated multisource estimates of mortality for Thailand in 2005. Population Health Metrics, 8-10.

Chuprapawon, C. 2003. A study on quality of cause of death data in hospitals: A research report (in Thai). Health Information System Development Office. Nonthaburi: Ministry of Public Health. Thailand.

Chutinantakul, A., Tongkumchum, P., Bundhamcharoe, K. and 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(25), 1-8.

Deesomchok, A., Dechayonbancha, N. and 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 Medical Association Thai. 88, 1236-1241.

Faramnuayphol, P., Chongsuvivatwong, V. and Panarunothai, S. 2008. Geographical variation of mortality in Thailand. Journal of Medical Association Thai. 91(9), 1455-60.

França, E., Rao, C., Abreu, D.M.X.D., Souza, M.D.F.M.D. and Lopez, A.D. 2012. Comparison of crude and adjusted mortality rates from leading causes of death in northeastern Brazil. Revista Panamericana de Salud Pública. 31(4), 275-282.

Hanchaiphiboolkul, S., Poungvarin, N., Nidhinandana,S., Suwanwela, N.C., Puthkhao, P., Towanabut, S., Tantirittisak, T., Suwantamee, J. and Samsen, M. 2011. Prevalence of Stroke and Stroke Risk Factors in Thailand: Thai Epidemiologic Stroke (TES) Study. Journal of the Medical Association of Thailand. 94(4), 427–436.

Hosmer, D.W. and Lemeshow, S. 2000. Applied Logistic Regression (2nd ed). New York: John Wiley and Sons.

Jalayondeja, C., kaewkungwal, J., Sullivan, P. E, Nidhinandana, S., Pichaiyongwongdee, S. and Jareinpituk, S. 2011. Factors related to community participation by stroke victims six month post-stroke. Southeast Asian Journal of Tropical Medicine and Public Health. 42(4), 1005–1013.

Jemal, A., Center, M.M., DeSantis, C. and Ward, E.M. 2010. Global Patterns of Cancer Incidence and Mortality Rates and Trends, Cancer Epidemiology Biomarkers Prevention Journal. 19(8), 1-15.

Kamnerdsupahon, P., Srisukho, S., Sumitsawan, Y., Lorvidhaya, V. and Sukthomya, V. 2008. Cancer in Northern Thailand. Biomedical Imaging and Intervention Journal, 4(3).

Khosravi, A., Rao, C., Naghavi, M., Taylor, R., Jafari, N. and Lopez, A.D. 2008. Impact of misclassification on measure of cardiovascular disease mortality in the Islamic Republic of Iran: a cross-sectional study. Bull World Health Organ. 86, 688-696.

Kijsanayotin, B., Ingun, P. and 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. 39.

Klinjun, N., Lim, A. and Bundhamcharoen, K. 2014. A logistic regression model for estimating transport accident deaths using verbal autopsy data. Asia-Pacific Journal of Public Health.

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

Lahti, R.A., and Penttilä, A. 2001. The validity of death certificates: routine validation of death certification and its effects on mortality statistics. Forensic Science International, 115(1), 15-32.

Lumley, T. 2010. Complex Surveys: A Guide to Analysis Using R, Wiley.

Lyles, R.H. 2002. A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. Biometrics. 58(4), 1034–1036.

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

Moriyama, I.M. 1989. Problems in measurement of accuracy of cause-of-death statistics. American Journal of Public Health. 79, 1349-50.

Pattaraachachai, J., Rao, C., Polprasert, W., Porapakkham, Y., Poa-in, W., Singwerathum, N. and Lopez, A.D. 2010. Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. Population Health Metrics. 8-12.

Polprasert, W., Rao, C., Adair, T., Pattaraachachai, J., Porapakkham, Y. and Lopez, A.D. 2010. Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. Population Health Metrics. 8:13.

Porapakkham, Y., Rao, C., Pattaraachachai, J., Polprasert, W., Vos, T., Adair, T. and Lopez, A.D. 2010. Estimated causes of death in Thailand, 2005: implications for health policy. Population Health Metrics. 8-14.

Pourhoseingholi, M.A., Faghihzadeh, S., Hajizadeh, E., Abadi, A. and Zali, M.R. (2009a). Bayesian estimation of colorectal cancer mortality in the presence of misclassification in Iran. Asian Pacific Journal of Cancer Prevention. 10(4), 691-4.

Pourhoseingholi, M.A., Faghihzadeh, S., Hajizadeh, E. and Abadi, A. (2009b). Bayesian analysis of gastric cancer mortality in Iranian population. Gastroenterology and Hepatology from bed to bench. 3(1).

Pourhoseingholi, A., Safaee, A., Pourhoseingholi, M.A., Moghimi-Dehkordi, B., Habibi, M., Vahedi, M. and Zali, M. R. 2010. Prevalence and demographic risk factors of gastrointestinal symptoms in Tehran province. Journal for Public Health. 8(7), 42-6.

Pourhoseingholi, M.A., Pourhoseingholi, A, Vahedi, M., Moghimi, D.B., Safaee, A., Ashtari, S. and Zali, M.R. 2011. Alternatives for Cox regression model: using parametric models to analyze the survival of cancer patients. Iranian Journal of Cancer Prevention. 1, 1-9.

Prasartkul, P., Porapakkham, Y., Vapattanawong, P. and Rittirong, J. 2007. Development of a Verbal Autopsy Tool for Investigating Cause of death: the Kanchanaburi Project. Journal of Population and Social Studies, 15(2), 1-22.

R, Core Team R. 2013. A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, Retrieved October 26, 2010, from http://www.R-project.org/

Rao, C., Porapakkham, Y., Pattaraarchachai, J., Polprasert, W., Swampunyalert, N. and Lopez, A.D. 2010. Verifying causes of death in Thailand: rationale and methods for empirical investigation. Population Health Metrics. 8:11.

Rampatige, R., Gamage, S., Peiris, S. and Lopez, A.D. 2013. Assessing the reliability of causes of death reported by the Vital Registration System in Sri Lanka: medical records review in Colombo.Health Information Management Journal. 42(3), 20.

Sposto, R., Preston, D.L., Shimizu, Y. and Mabuchi, K. 1992. The effect of diagnostic misclassification on non-cancer and cancer mortality dose–response in A-bomb survivors. Biometrics. 48, 605-17.

Sripa, B., Kaewkes, S., Sithithaworn, P., Mairiang, E., Laha, T., Smout, M., Pairojkul, C., Bhudhisawasdi, V., Tesana, S., Thinkamrop, B., Bethony, J.M., Loukas, A. and Brindley, P.J. 2007. Liver Fluke Induces Cholangiocarcinoma. PLOS MED. 4(7), 1148-1155.

Stamey, J.D., Young, D.M. and Seaman, J.W. 2008. A Bayesian approach to adjust for diagnostic misclassification between two mortality causes in Poisson regression. Statistics in medicine. 27(13), 2440-2452.

Suwanwela, N.C. 2014. Stroke Epidemiology in Thailand. Journal of Stroke. 16(1), 1–7.

Tangcharoensathien, V., Faramnuayphol, P., Teokul, W., Bundhamcharoen, K. and Wibulpholprasert, S. 2006. A critical assessment of mortality statistics in Thailand. Bulletin of the World Health Organization. 84, 233-239.

Tongkumchum, P. and McNeil, D. 2009. Confidence interval using contrasts for regression model. Songklanakarin Journal Science Technology. 31(2), 151-6.

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

Venables, W.N. and Ripley, B.D. 2002. Modern Applied Statistics with S (4rd ed). New York: Springer-Verlag.

Viratroumanee, C., Pramyothin, P., Limwongse, C., Suwannasri, P. and Assawamakin, A. 2009. Glutathione S-Transferase P1 Variant Plays a Major Contribution to Decreased Susceptibility to Liver Cancer in Thais. Asian Pacific Journal of Cancer Prevention. 10,783-788.

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

Whittemore, A.S. and Gong, G. 1991. Poisson regression with misclassified counts: application to cervical cancer mortality rates. Applied Statistics. 40, 81-93.

World Health Organization. 2004. ICD-10 International Statistical Classification of Disease and Related Health Problems, Geneva, Switzerland.

World Health Organization. The global burden of disease, 2008. [Cited 2013 Nov 17]. Available from: URL: http://www.who.int/healthinfo/global_burden_disease/en/

Appendix I

Article I: "Estimating Liver Cancer Deaths in Thailand based on Verbal Autopsy Study"

JRHS 2014; 14(1): 18-22

JRHS
Journal of Research in Health Sciences
journal homepage: www.umsha.ac.ir/jrhs

Original Article

Estimating Liver Cancer Deaths in Thailand based on Verbal Autopsy Study

Salwa Waeto (MSc)[a], Nattakit Pipatjaturon (MSc)[a,b], Phattrawan Tongkumchum (PhD)[a,c*], Chamnein Choonpradub (PhD)[a], Rattikan Saelim (PhD)[a,c], Nifatamah Makaje (PhD)[a,c]

[a] Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand
[b] The Office of Diseases Prevention and Control 9th Phitsanulok, Phitsanulok, Thailand
[c] Centre of Excellence in Mathematics, CHE, Si Ayutthaya Road, Bangkok, Thailand

ARTICLE INFORMATION

Article history:
Received: 14 September 2013
Revised: 01 November 2013
Accepted: 05 December 2013
Available online: 10 December 2013

Keywords:
Mortality
Logistic regression
Confidence intervals
Sum contrasts

** Correspondence*
Phattrawan Tongkumchum (PhD)
Tel: +66 86 2876915
Fax: +66 73 312 179
E-mail1: tphattra@bunga.pn.psu.ac.th
E-mail2: phattrawan@gmail.com

ABSTRACT

Background: Liver cancer mortality is high in Thailand but utility of related vital statistics is limited due to national vital registration (VR) data being under reported for specific causes of deaths. Accurate methodologies and reliable supplementary data are needed to provide worthy national vital statistics. This study aimed to model liver cancer deaths based on verbal autopsy (VA) study in 2005 to provide more accurate estimates of liver cancer deaths than those reported. The results were used to estimate number of liver cancer deaths during 2000-2009.

Methods: A verbal autopsy (VA) was carried out in 2005 based on a sample of 9,644 deaths from nine provinces and it provided reliable information on causes of deaths by gender, age group, location of deaths in or outside hospital, and causes of deaths of the VR database. Logistic regression was used to model liver cancer deaths and other variables. The estimated probabilities from the model were applied to liver cancer deaths in the VR database, 2000-2009. Thus, the more accurately VA-estimated numbers of liver cancer deaths were obtained.

Results: The model fits the data quite well with sensitivity 0.64. The confidence intervals from statistical model provide the estimates and their precisions. The VA-estimated numbers of liver cancer deaths were higher than the corresponding VR database with inflation factors 1.56 for males and 1.64 for females.

Conclusion: The statistical methods used in this study can be applied to available mortality data in developing countries where their national vital registration data are of low quality and supplementary reliable data are available.

Citation: Waeto S, Pipatjaturon N, Tongkumchum P, Choonpradub C, Saelim R, Makaje N. Estimating Liver Cancer Deaths in Thailand based on Verbal Autopsy Study. J Res Health Sci. 2014;14(1):18-22.

Introduction

Quality of mortality data is a major problem in providing reliable national vital statistics of developing countries. In Thailand, mortality data are also questionable because the coverage is incomplete and causes of deaths often mis-specified [1]. Causes of deaths have been coded according to the World Health Organization's International Classification of Diseases (ICD). Nearly 40% of death certificates give the ICD-code cause of R00-99 "ill-defined"[1-3], and thus many specific causes, including liver cancer, go largely under-reported, whereas less than 4% of Japan's deaths are ill-defined [4]. Japan is considered as one of the most developed countries in Asia and it has a reliable vital registration database [1].

In 2005, the Ministry of Public Health of Thailand proposed a verbal autopsy (VA) study to build capacity among Thai health professionals (physicians, paramedical staff, biostatisticians and epidemiologists) to critically assess vital registration (VR) data and improve the quality of causes of death recorded at registration [2,5-7]. The assessment process was based on medical record review for inside hospital deaths and standard verbal autopsy questionnaires for outside hospital deaths. It provided a reasonable basis for ascertaining the true underlying cause of death. The results have yielded corrected estimates of the true underlying cause of death pattern. The validity of the VA in the Thai context is accurate at some levels. In fact, for some site-specific cancers, the sensitivity scores were higher than 75% [6]. However, Byass [8] concluded that uncertainties remain and suggested further research in the area of probabilistic modeling. Therefore, appropriate statistical methods are needed for beneficial use of the VA data to provide reliable national vital statistics of a particular cause of deaths.

This study focuses on liver cancer mortality which is high in Thailand [9-12]. Age-standardized liver cancer mortality was 31.0 per 100,000 in Thailand in 2004 whereas it was 13.0 for Japan [4]. However, such comparison is complicated by the fact that these countries have quite different age distributions (only 4.9% of the Thai population in 2005 was aged 70 [13] or more compared with 15.0% of the Japanese population in 2006 [14]).

There are two kinds of liver cancer. Hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA)[11]. The ICD-10

code for HCC is C22.0 and for CCA is C22.1. HCC and CCA have different etiology [9-10,12,15] but Thai death certificates code both as C22.9 (unspecified liver cancer).

The objectives of our study were to estimate percentages of liver cancer deaths in Thailand based on data from the 2005 VA study and to apply the adjusted percentages to numbers of liver cancer deaths reported from the VR database from 2000 to 2009. The goal was to increase reliability and precision of the national liver cancer mortality data in Thailand.

Methods

Data sources

This study used secondary data from the VA survey. The VA was designed to verify causes of death for nationality representative sample of deaths that occurred in Thailand using multistage stratified cluster sampling technique. The sample was drawn from the VR database and the sampling unit was a registered death of Thai citizen, who was permanent resident in Thailand. Full details of the sampling procedures were explained elsewhere[2].

The VA study was carried out in 2005 based on a sample of 3,316 in-hospital and 6,328 outside-hospital deaths from 28 selected districts in nine provinces[2,5-7], giving 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 data were separated by field (d), grouped fields (c) and (e) into the 21 leading causes of death for each location plus all other cause group, and thus found inflation factors for determining percentages of deaths in specific cause groups. The 22 groups were classified according to the ICD-10 Mortality Tabulation categories[16] and each group had to be large enough for statistical analysis. The cause group based on the VA count ranged from 77 for septicemia (A40-41) to 1,076 for stroke (I60-69). There were 500 deaths for liver cancer (C22).

Statistical Methods

The outcome was liver cancer death (yes/no) and the determinants were province, gender-age group and VR cause location. The logistic regression model[17-18] was used for describing the relation between the outcome and determinants. This model formulated 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:

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

In this model μ is constant, α_i, β_j and γ_k refer to province, gender-age group and VR cause-location, respectively. The province factor has nine levels corresponding to the nine provinces in the VA sample. The gender-age group factor has 13 levels, by classifying age into seven groups (0-29, 30-39,…,70-79, 80+) for males and six groups for females (no females aged below 30 died from liver cancer). The VR cause-location factor has 12 levels, corresponding to the six most likely VR cause groups (liver cancer, other digestive cancer, other cancer, digestive, ill-defined and septicemia, and other causes) for liver cancer in the VA study and the two locations (in or outside hospital).

The model as described in equation (1) was fitted based on treatment contrasts with Bangkok as a reference group to get the nine province coefficients compared to Bangkok. To assess the goodness of fit of the model the Receiver Operating Characteristic (ROC) curve was used. It showed how well a model predicts a binary outcome. The interpretation of how well a model predicts a binary outcome is made by the area under the ROC curve. In particular, the more the area under the curve, the more accurate the model is. Denoting the predicted outcome as 1 (liver cancer) if $p \geq c$, or 0 (other death) if $p<c$, it plotted sensitivity (proportion of positive outcomes correctly predicted by the model) against the false positive rate (proportion of all outcomes incorrectly predicted), as c varies. In our case, we chose c to give predicted liver cancer deaths in agreement with the liver cancer deaths in the VA study, which were 500 cases.

The province coefficients from the model were then used to extrapolate the province coefficients for the rest of the country using triangulation method. To get confidence intervals of adjusted percentage of liver cancer deaths the model based on sum contrasts was used. The adjusted percentages of liver cancer deaths were presented using graphs of confidence intervals. Thus, the estimated probabilities of liver cancer deaths were obtained.

Sum contrasts

Sum contrasts[19-20] was used to obtain confidence intervals for comparing means/proportions with the overall mean/proportion. 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. The confidence intervals based on sum contrasts are used because they are more appropriate compared to the corresponding confidence intervals based on the treatment contrasts. The confidence intervals compare percentage of liver cancer deaths in each category factor with the overall percentage. They applied equitably to each category, whereas the commonly used confidence intervals based on treatment contrasts measured the difference from a reference group that is taken to be fixed and thus does not have a confidence interval.

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. Coefficients for provinces inside triangles were obtained by solving three linear equations via linear algebra.

Coefficients for provinces outside triangles were obtained similarly by extrapolation. The interpolated values for all 76 provinces reflect regional variation of liver cancer mortality compared to the reference province (Bangkok).

Applied the estimated probabilities of liver cancer deaths to the VR data

Finally, we applied the estimated probabilities of liver cancer deaths from the model to the target population (all reported Thai deaths 2000-2009). To do this, we used the interpolated values for the province effects, and we assumed that the model was valid for years before and after 2005. By doing this, the numbers of deaths were estimated for each gender-age group and year. The area plot was used to show estimated liver cancer deaths for each gender-age group for each year during 2000-2009. All statistical analysis, graphs and maps were carried out using the R program version 3.0.1.

Results

Preliminary Results

According to the 9,644 cases in the VA study, it was assessed that 500 deaths were due to liver cancer. Of the 500 VA liver cancer deaths, the most likely VR reported causes were liver cancer (236), other digestive cancer (39), other cancer (48), digestive (49), ill-defined or septicemia (99), and all other (29).

Figure 1 shows the percentage of assessed liver cancer deaths in nine provinces, 13 gender-age-groups and 12 VR reported cause-location groups. More than 80% of reported liver cancer deaths were really due to liver cancer. But among deaths outside hospital, 33% of those reported as digestive disease and 25% of those reported as other digestive cancer were really due to liver cancer.

Logistic Regression Model

The *P*-value for a factor in the regression model is the probability of χ^2 being greater than D, the tail area of a chi-squared distribution with k-1 degrees of freedom, 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. The three factors in the logistic regression model are highly statistically significant (P<0.001).

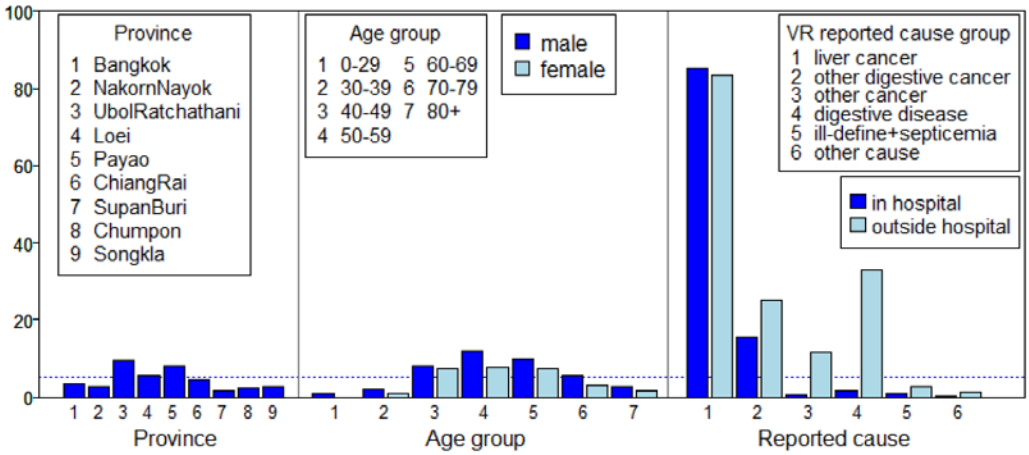


Figure 1: Percentage of liver cancer deaths by province, gender-age group and VR cause-location

Figure 2 shows the ROC curve of logistic regression model. 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. Note that just using the reported cause to predict the true cause has sensitivity 0.47. Only 236 cases out of 500 liver cancer deaths were correctly reported.

Figure 3 shows confidence intervals of percentage deaths due to liver cancer from logistic model based on sum contrasts. The model suggests that the percentages of liver cancer in Payao Province in the north and Ubonratchatanee Province in the northeast were higher than the average percentage, whereas Supanburee Province in central Thailand was lower than the average.

For gender age groups, males had higher percentages than those of females. The percentages of liver cancer deaths were higher than average in ages 40-49, 50-59 and 60-69 for males, and in age 60-69 for females.

For the VR cause-location, deaths in hospital due to liver cancer were more likely to be reported as liver cancer (85.2%) and other digestive cancer (15.4%). For deaths outside hospital, they were more likely to be reported as liver cancer (83.5%) and other digestive cancer (25.0%).

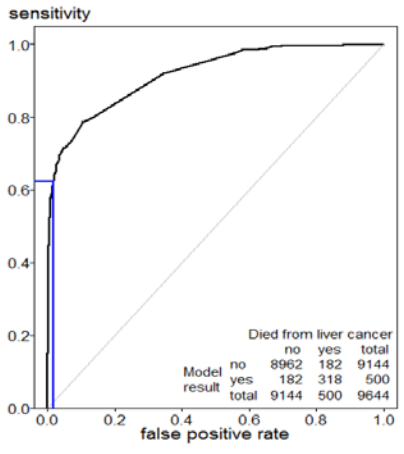


Figure 2: Receiver Operating Characteristic (ROC) curve and cross-classifying observed and estimated outcome

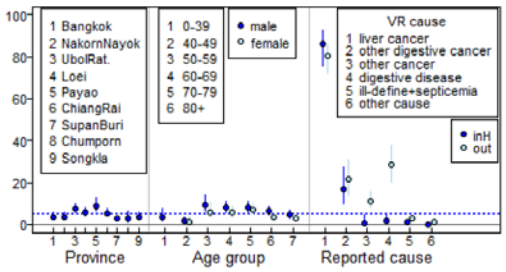


Figure 3: Confidence intervals for comparing liver cancer percentage with overall percentage (dotted line)

The estimated probabilities of liver cancer deaths from the model were applied to the VR data for males and females by age groups from 2000 to 2009. Over the decade 2000-2009, the estimated numbers of liver cancer deaths were 134,244 (males) and 58,548 (females). These are 56% and 64% higher than the reported totals of 85,873 and 35,643, respectively. Figure 4 compares numbers of liver cancer deaths between VA estimated and VR reported deaths using area plot.

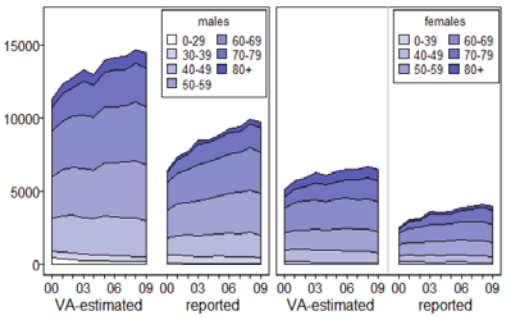


Figure 4: Area plot for number of liver cancer deaths in 2000-2009

Discussion

This study adjusted number of reported liver cancer deaths from the VR database using the logistic regression model based on the 2005 VA data of liver cancer deaths and three determinants including province, gender-age group and VR cause location group.

The model showed that province, gender-age group and VR cause location group were highly statistically significant related to liver cancer deaths. The liver cancer deaths were more likely to occur in Payao Province in the north and Ubonratchatanee Province in the northeast of Thailand. This finding was in agreement with previous studies[9,11-12] that reported high liver cancer incidence rates in the northeast of Thailand. In particular, the overall ratio of mortality to incidence is almost one, meaning that the higher rate of the incidence indicating the higher rate of the mortality. Moreover, the geographical inequality of liver cancer in Thailand[21] supported our finding.

It is well known that liver cancer mortality varies with gender and age[9]. It more pronounces among males and elderly. The results in this study were of high percentages in males. For age the percentages of liver cancer deaths were higher than average in ages 40-49, 50-59 and 60-69 for males and marginally higher in age 60-69 for females. Therefore, estimating number of liver cancer deaths is necessary to take these demographic factors into account.

Moreover, our adjusted results showed that the deaths due to liver cancer were more likely to be correctly reported with more than 80% for both deaths in hospitals and outside hospitals. The misreported of liver cancer deaths were also not very high in the previous report of the 2005 VA data using different methodology[7]. For misreported cases, their cause of deaths were more likely to be recorded as other digestive cancer for both deaths in and outside hospitals, and digestive disease and other cancers for only deaths outside hospitals. So cause of deaths recoding has to be more concerned.

The estimates number of liver cancer deaths over the decade 2000-2009 were 134,244 (males) and 58,548 (females). These were 56% and 64% higher than the reported totals of 85,873 and 35,643, respectively. The estimates numbers of liver cancer deaths tended to be a little increased with year. It may be related to changing in age distribution of Thai population[13].

The strength of this study is the methodologies used. Logistic regression is commonly used in public health research. According to our knowledge, it has not been applied to the verbal autopsy study. Other methods such as capture-recapture[22] were used for estimation incomplete data not misclassification data. The capture-recapture technique is applicable to estimating the size of populations of mobile objects like HIV-mobility/incidence. In the case of mortality, the method is applied to estimate the undercount. In our case, the liver cancer mortality was misclassification not undercounted.

There is a limitation in our study. The verbal autopsy study was based on cluster sampling. We fixed the province effect because cluster sampling gave standard error larger than simple random sampling[23].

The unreliable cause of death from vital registration database in countries like Thailand necessitates extensive adjustment to the data in order to derive plausible liver cancer mortality by gender, age and regions or provinces or districts. The data with more reliable cause of deaths from well-designed research such as the VA study[2,5-7] together with appropriate statistical methods are very useful for making adjustment to imperfect registration data. This study reported the utility of statistical methods in analyzing existing data to derive estimates of liver cancer deaths in Thailand from 2000 to 2009.

Conclusions

The statistical methods used in this study can be applied to available mortality data in developing countries where their national vital registration data are of low quality and supplementary reliable data are available.

Acknowledgments

This research was supported by Centre of Excellence in Mathematics, the Commission on Higher Education, Thailand. We would like to thank Professor Don McNeil for his helpful guidance and Dr. Kanitta Bundhamcharoen from Thai Ministry of Public Health for providing us the data.

Graduate School, Prince of Songkla University supported scholarship for Salwa Waeto and Nattakit Pipatjaturon.

Conflict of interest statement

The authors have no conflict of interests to declare.

Funding

This study was funded by Centre of Excellence in Mathematics, the Commission on Higher Education, Thailand.

References

1. Mathers CD, Fat DM, Inoue M, Rao C, Lopez AD. Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bull World Health Organ*. 2005;83(3):171-177.
2. Rao C, Porapakkham Y, Pattaraarchachai J, Polprasert W, Sawanpanyalert N, Lopez AD. Verifying causes of death in Thailand: rationale and methods for empirical investigation. *Popul Health Metr*. 2010;8:11.
3. Tangcharoensathien V, Faramnuayphol P, Teokul W, Bundhamcharoen K, Wibulpholprasert S. A critical assessment of mortality statistics in Thailand. *Bull World Health Organ*. 2006;84:233-239.
4. World Health Organization. The global burden of disease: 2004 update. Geneva: WHO; 2008. [updated 2013; cited 26 November 2013]; Available from: http://www.who.int/healthinfo/global_burden_disease/en/.
5. Pattaraarchachai J, Rao C, Polprasrt W, Porapakkham Y, Pao-in W, Singwerathum N, Lopez AD. Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. *Popul Health Metr*. 2010;8:12.
6. Polprasert W, Rao C, Adair T, Pattaraarchachai J, Porapakkham Y, Lopez AD. Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: application of verbal autopsy methods. *Popul Health Metr*. 2010;8:13.
7. Porapakkham Y, Rao C, Pattaraarchachai J, Polprasert W, Vos T, Adair T, Lopez AD. Estimated causes of death in Thailand, 2005: implications for health policy. *Popul Health Metr*. 2010;8:14.
8. Byass P. Integrated multisource estimates of mortality for Thailand in 2005. *Popul Health Metr*. 2010;8:10.
9. Jemal A, Center MM, DeSantis C, Ward EM. Global Patterns of Cancer Incidence and Mortality Rates and Trends, *Cancer Epidem Biomar*; 2010;19(8):OF1-OF15.
10. Vatanasapt V, Sriamporn S, Vatanasapt P. Cancer Control in Thailand. *Jpn J Clin Oncol*. 2002;32:S82-S91.
11. Sripa B, Kaewkes S, Sithithaworn P, Mairiang E, Laha T, Smout M, Pairojkul C, Bhudhisawasdi V, Tesana S, Thinkamrop B, Bethony JM, Loukas A, Brindley PJ. Liver Fluke Induces Cholangiocarcinoma. *PLOS MED*. 2007;4(7):1148-1155.
12. Viratroumanee C, Pramyothin P, Limwongse C, Suwannasri P, Assawamakin A. Glutathione S-Transferase P1 Variant Plays a Major Contribution to Decreased Susceptibility to Liver Cancer in Thais. *Asian Pac J Cancer P*. 2009;10:783-788.
13. United Nations, Population Division. *World Population Prospects: The 2006 Revision, Volume III: Analytical Report*. New York: United Nation Publication; 2007.
14. National Institute of Population and Social Security Research. *Population Statistics of Japan 2008*. Tokyo: NIPSSR; 2008.
15. Ahmed F, Perz JF, Kwong S, Jamison PM, Friedman C, Bell BP. National Trends and Disparities in the Incidence of Hepatocellular Carcinoma, 1998–2003. *CDC*. 2008;5:3.
16. World Health Organization: *ICD-10 International Statistical Classification of Diseases and Related Health Problems*. Geneva: WHO; 2004.
17. Hosmer, DW, Lemeshow S. *Applied Logistic Regression* 2nd ed. New York: John Wiley and Sons; 2000.
18. Kleinbaum DG, Klein M. *Logistic Regression: A Self-Learning Text*. 2nd ed. New York: Springer-Verlag; 2002.
19. Venables W, Ripley B. *Modern Applied Statistics with S*. 4th ed. New York: Springer-Verlag; 2002.
20. Tongkumchum P, McNeil D. Confidence intervals using contrasts for regression model. *Songklanakarin J Sci Technol*. 2009;31(2):151-156.
21. Faramnuayphol P, Chongsuvivatwong V, Panarunothai S. Geographical Variation of Mortality in Thailand. *J Med Assoc Thai*. 2008;91(9):1455-1460.
22. Khazaei S, Poorolajal J, Mahjub H, Esmailnasab N, Mirzaei M. Estimation of the frequency of intravenous drug users in Hamadan City, Iran, using the capture-recapture method. *Epidemiol Health*. 2012;34:e2012006.
23. Lumley T. *Complex Surveys: a guide to analysis using R*. Manhattan: Wiley; 2010.

Appendix II

Article II: "Estimating Liver Cancer Deaths in Thailand: Methodologies to Optimize the Use of Verbal Autopsy Data"

TITLE PAGE

Type of manuscript	Original article
Title	Estimating Liver Cancer Deaths in Thailand: Methodologies to Optimize the Use of Verbal Autopsy Data
Authors	Nattakit Pipatjaturon[1, 2] (M.Sc. Biostatistics) Arinda Ma-a-lee[1] (M.Sc. Research Methodology) Phattrawan Tongkumchum[1] (Ph.D. Statistics) Attachai Ueranantasan[1] (Ph.D. (Research Methodology)
Affiliation	[1] Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, 94000, Thailand [2] The office of Diseases Prevention and Control 9th Phitsanulok, Phitsanulok, 65000, Thailand
Correspondence	Phattrawan Tongkumchum Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, 94000, Thailand, Tel: +66 86 2876915 Fax: +66 73 312 179 e-mail: phattrawan@psu.ac.th, phattrawan@gmail.com
Running title	Estimating Deaths due to Liver Cancer
Word count	Abstract: 232 Text excluding abstract and references: 3067

Estimating Liver Cancer Deaths in Thailand: Methodologies to Optimize the Use of Verbal Autopsy Data

Abstract

Mortality statistics are essential for national policies on intervention and resource allocation. Mortality statistics derived from death registration (DR) in Thailand need to be estimated because the DR is currently considered both incomplete and inaccurate. Verbal autopsy (VA) survey was carried out to verify cause of death in the DR and thus the VA-assessed cause are more reliable than the registered cause. In this paper, we described statistical methods used to estimate liver cancer deaths. The methods involve analysis of the VA data with liver cancer deaths as an adverse outcome using logistic regression. A categorical variable of registered causes for liver cancer deaths is a predictor. Demographic factors (age and gender) and locality of the deceased (province) are covariates. The models with and without demographic factors were compared. The Goodness of Fit of the models was assessed using a receiver operating characteristic (ROC) curve. The estimates were applied to number of death in the DR data and thus the estimated numbers of liver cancer deaths were obtained. The misreported cases were mainly for deaths outside hospital and more likely to be reported as digestive disease, other digestive cancer and other cancers. Regional variations for liver cancer were observed and suggested that liver cancer deaths were more frequent in the north and northeast. The methods enable health professionals to estimate any specific causes of death when DR causes of death are inaccurate.

Key words: Logistic regression model, ROC, Triangulation method

Introduction

Mortality statistics in the nation are essential for monitoring health and planning appropriate health services. In Thailand, 65% of deaths occur outside hospitals. Although death registration (DR) system attempts to document deaths and their causes covering the whole country, information on causes of death (COD) were considered as low quality (Mathers et al. [1]) due to 35-40% of deaths being ill-defined (Pattaraarchachai et al. [2] , Rao et al. [3]) and extensive misclassification of specific causes of death (Tangcharoensathien et al. [4]).

Verbal autopsy (VA) has been widely used for the assessment of COD in countries where death registration systems are weak and most people die at home without medical certification of the COD. The VA method determines the cause of death from data collected about the symptoms and signs of illness and the events preceding death. It has procedures to ensure that the data collected is of high quality.

The recent VA survey in Thailand was carried out in 2005. It was claimed as the first complete national application of the WHO methodology to Thailand. It built capacity among Thai health professionals (physicians, paramedical staff, biostatisticians and epidemiologists) to critically assess death registration data and improve the quality of COD recorded at registration. In previous studies, the 2005 VA data were used to estimate various COD including liver cancer (Pattaraarchachai et al. [2], Rao et al. [3], Porapakkham et al. [5], Polprasert et al. [6]). However, the simple cross-referencing method used in these studies, ignored the effect of gender-age group and locality of the deceased, which could give incorrect estimates due to confounding (Carmichael [7]).

There is still more rooms, to improve COD at national level based on VA data using appropriate statistical model (Byass [8]) that allows for confounding bias to be detected.

This study aims to propose methodologies to optimize the use of the VA data for estimating numbers of deaths for inaccurate COD in DR database. The methods were illustrated using liver cancer deaths. Firstly, we fitted logistic regression model to the VA data with liver cancer as an adverse outcome. The main predictor is DR reported COD for liver cancer. Secondly, we included demographic variables (gender, age and province) into the model to predict liver cancer death. Thirdly, we interpolated coefficients for province outside the VA study. Finally, we applied the estimated probabilities to number of death in DR data in 1996 to 2009. Thus, estimated numbers of liver cancer deaths were obtained.

Material and methods

The data analysis involves an issue of how to make use of the existing sample data to arrive at accurately estimate number deaths in target population. The study sample comprises 9,495 deaths aged 5 years and older in 2005 VA study. The target population comprises all reported Thai deaths aged 5 years and older in 1996-2009. A step of analysis is summarized in Figure 1.

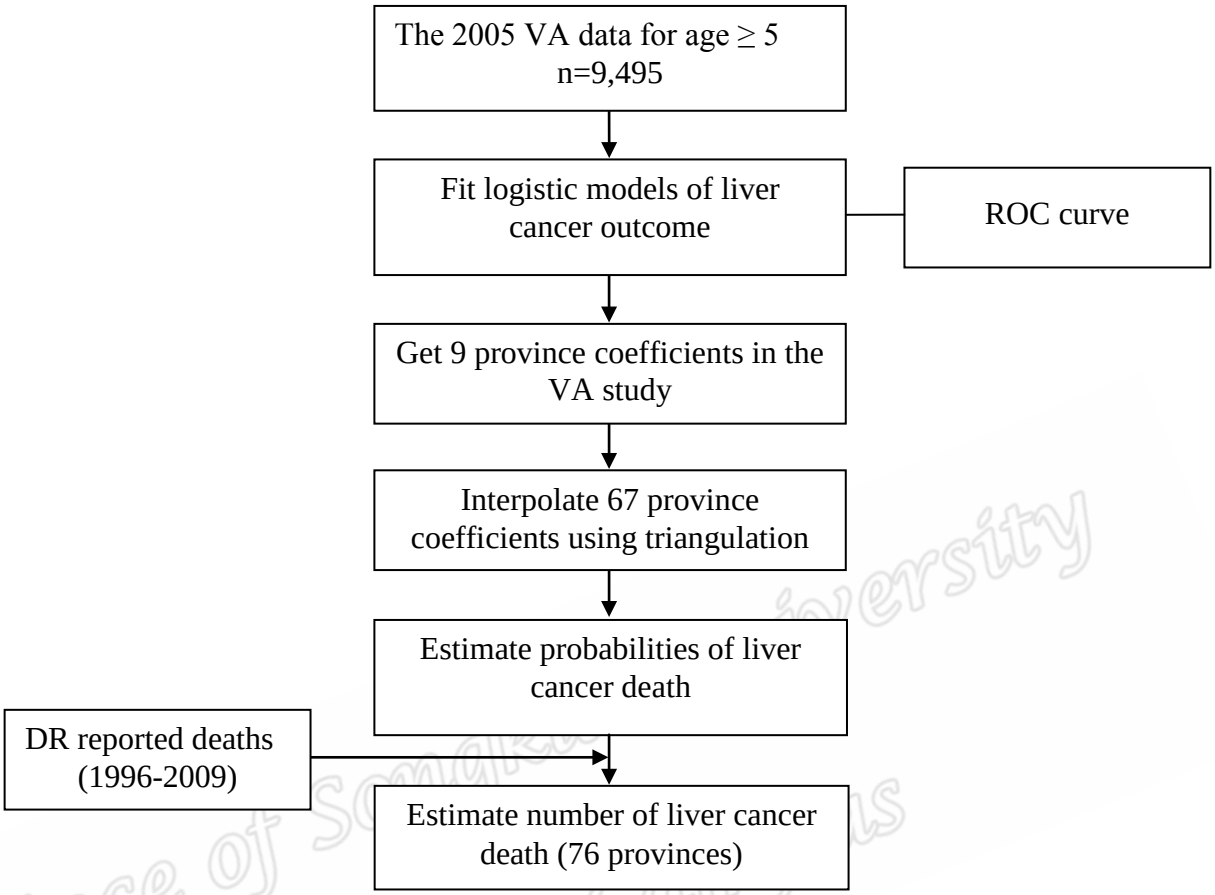


Figure 1 Diagram of analysis process

The VA data

The VA study was carried out in 2005 by the Setting Priorities using Information on Cost-Effectiveness (SPICE) analysis project team to assess causes of death based on a sample of 9,644 deaths (3,316 in-hospital and 6,328 outside-hospital deaths). A clustered sample was taken from 28 selected districts in nine provinces (Rao et al. [3]). Districts were selected by two-stage stratification where Bangkok and pairs of provinces from the four regions were first randomly selected. Stratification was on the number of deaths in the regional province or district. Then, death certificates to be assessed were randomly selected from the 28 districts using the probability-proportional-to-size method. The nine provinces were ChiangRai, Phayao,

UbonRatchathani, Loei, Bangkok, SupanBuri, Nakhonnayok, Chumphon, and Songkhla superimposed with sample size in blue bubbles as shown in Figure 2.

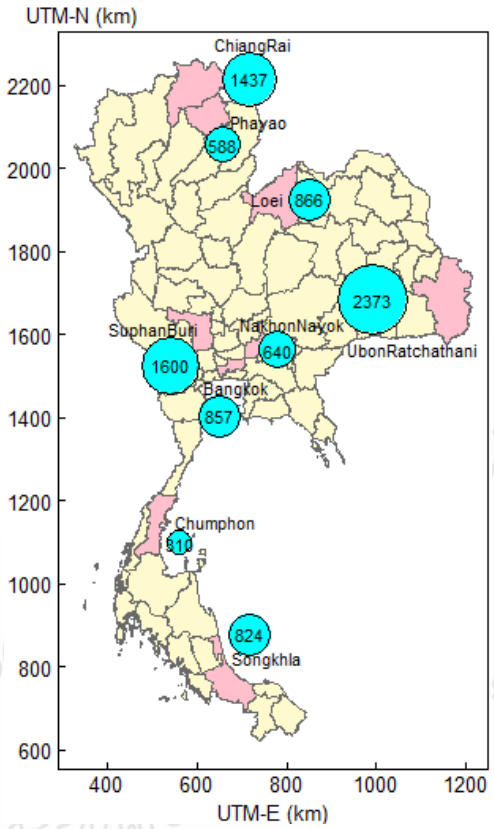


Figure 2 Map of Thailand showing the nine provinces with sample size of 2005 VA survey

Since no cases died of liver cancer at age less than five years, the study sample was reduced to 9,495 deaths aged five years and older (3,212 in-hospital and 6,283 outside-hospital deaths). Data collected on deceased persons were province, gender, age, location of death (in or outside hospitals), the DR-reported International Statistical Classification of Diseases (ICD-10) code reported on the death certificate, and the VA-assessed ICD-10 codes. The VA-assesses ICD-10 codes are more reliable than the DR-reported ICD-10 codes.

The DR data

The DR data were information from death certificates including ICD-10 codes, age, gender, province and location. However, the DR data are not complete and Carmichael [7] described problems in Thai data. Age was not recorded for 0.73% of death certificates from 1996-2009. It had decreased to 0.02% in 2009. Province was not recorded for 1.08% of death certificates from 1996-1999 and 3.5% from 2000-2004. It had decreased to 0.2% for 2005-2009, and it was only 0.01% in 2009. Ill-defined cause of death is ICD-10 code R00-R99, which includes senility (R54). Percentages of deaths recorded in these codes averaged 37.7% (37.9% in 2009), ranging from 10.8% for males aged 10-19 to 75.5% for females aged 80+. Cases with missing province were omitted.

Data analysis

The VA-assessed ICD-10 codes were grouped into 21 major causes according to the chapter-block classification of ICD-10 codes based on mortality tabulation (World Health Organization [9]). The groups were also necessary to be large enough for statistical analysis. The 21 major cause groups are defined in Table 1. Numbers of death for each group were ranged from 77 for septicemia to 1,076 for stroke. The liver cancer deaths were 500.

Table 1. Definition of 21 major cause groups.

Cause of death groups	Cause of death groups
1:TB (A15-A19)	11:Ischemic (I20-I25)
2:Septicemia (A40-A41)	12:Stroke (I60-I69)
3:HIV (B20-B24)	13:Other CVD (I^{-})
4:Other Infectious (A, B)$^{-}$	14:Respiratory (J00-J99)
5:Liver Cancer (C22)	15:Digestive (K00-K99)
6:Lung Cancer^{+} (C30-C39)	16:GenitoUrinary (N00-N99)
7:Other Digestive Cancer (C15-C26)$^{-}$	17:Ill-defined (R00-R99)
8:Other Cancer (C^{-}, D00-D48)	18:Transport Accident (V00- V99)
9:Endocrine (E00-E99)	19:Other injury (W00-W99, X00-X59)
10:Mental, Nervous (F00-F99, G00-G99)	20:Suicide (X60-X84)
	21:All other

$^{+}$*Respiratory/thoracic,* $^{-}$*exclude above*

To identify mis-classification, cross tabulation between the VA-assessed ICD-10 and the DR-reported ICD-10 codes was created as shown in Figure 3. A total of 500 deaths were assessed as liver cancer deaths. Their DR reported causes were liver cancer (236), ill defined (97), digestive (49), other cancer (48), other digestive cancer (39), lung cancer (7), genitourinary (7), and all other (17). The ill-defined, digestive, other cancer, other digestive cancer, lung cancer, genitourinary and all others are among the causes that liver cancer deaths are often misreported.

	VA group																				
DR group	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	total
01:TB	48	0	46	4	0	2	1	1	2	1	1	2	0	10	2	0	0	0	0	0	2	122
02:septicemia	6	20	53	27	2	3	8	24	62	13	10	60	14	62	52	52	6	3	8	2	32	519
03:HIV	2	0	164	1	0	0	1	2	0	3	0	0	1	0	1	0	0	0	0	0	1	176
04:othInfDis	2	4	36	28	3	0	0	1	5	5	0	5	2	5	9	3	1	0	0	0	5	114
05:liverCa	1	1	3	3	236	0	9	11	1	0	2	0	0	0	11	2	1	0	0	0	0	281
06:lungCa+	2	0	9	0	7	164	6	18	1	0	0	1	0	8	0	0	1	0	0	1	0	218
07:othDigCa	0	0	1	0	39	3	124	15	0	0	0	2	0	0	2	0	0	0	1	1	0	188
08:otherCa	1	0	12	3	48	32	35	391	4	1	3	8	3	3	4	2	0	0	2	1	1	554
09:endocrine	0	1	3	0	6	2	3	5	129	1	11	13	9	5	6	7	4	1	2	3	3	214
10:ment+Nerv	2	1	21	4	0	0	1	3	3	27	1	42	8	7	12	1	4	5	1	2	6	151
11:IHD	2	1	1	1	2	2	1	1	35	1	166	7	23	6	3	9	9	0	2	0	1	273
12:stroke	0	1	3	2	0	1	0	6	9	6	5	267	21	1	2	1	1	13	14	2	9	364
13:otherCVD	0	1	2	3	1	1	1	3	9	1	50	35	60	12	10	10	9	2	1	0	5	216
14:resp	31	2	83	11	2	17	4	12	24	11	17	45	17	256	20	20	3	8	10	0	10	603
15:digestive	7	1	12	12	49	4	7	11	6	3	4	5	5	4	153	6	2	1	3	1	5	301
16:genitoUrin	0	1	5	12	7	0	4	16	83	2	4	12	49	8	12	151	0	0	5	2	6	379
17:illDef	89	41	50	106	97	89	82	167	265	139	332	535	320	410	187	146	456	32	97	18	143	3801
18:transAcc	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	254	4	1	1	262
19:otherInj	1	0	1	0	0	0	1	1	1	4	4	6	0	3	0	0	2	63	132	2	3	224
20:suicide	0	0	2	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	5	90	7	107
21:allOther	1	2	5	2	1	0	2	9	8	4	5	31	7	1	3	2	2	153	40	32	118	428
total	195	77	512	219	500	320	290	697	647	223	617	1076	540	801	489	412	501	536	327	158	358	9495

Figure 3 Cross tabulation between VA and DR cause groups

The binary outcome was the VA-assessed ICD-10 codes (liver cancer deaths (C22) or others). The determinants were province, gender, age, location of death, and the DR-reported ICD-10 codes that often misreported for liver cancer deaths.

Gender and age group were combined into 12 levels of a gender-age group factor with seven age groups (5-39, 40-49, 50-59, 60-69, 70-79 and 80+) for each sex. Since the effects of gender and age as determinants of liver cancer death might not be additive, there is an advantage in combining them to form a single factor corresponding to all gender-age group combinations.

According to the VA data, liver cancer deaths were reported with eight different causes. The location (in or outside hospitals) of death and the DR-reported ICD-10 codes were combined into 16 levels of a DR-cause location factor.

Logistic regression model

The simple model with DR cause location as a determinant was first fitted to explore the misclassification. Then, the full model with three determinants was fitted to compare effect of demographic factors. The model formulated the logit of the probability p_{ijk} of death due to liver cancer as an additive linear function of the three determinants as

$$\log (p_{ijk}/(1-p_{ijk})) = \mu + \alpha_i + \beta_j + \gamma_k \quad (1)$$

where p_{ijk} is the probability of liver cancer death in each of the i, j and k groups of determinants, μ is a constant, α_i, β_j, and γ_k refer to province, gender-age group, and cause-location group, respectively. This equation may be inverted to give an expression for the probability p_{ijk} as

$$p_{ijk} = 1/(1+\exp(-(\mu + \alpha_i + \beta_j + \gamma_k))). \quad (2)$$

The model provided confidence intervals (CIs) for percentages of liver cancer deaths for levels of each determinant adjusted for other determinants, using sum contrasts method (Tongkumchum and McNeil [10] and Kongchouy and Sampantarak [11]). The confidence intervals were compared with bar charts of sample percentages to assess evidence of confounding bias.

Sum contrasts

The CIs based on sum contrasts have an advantage in that they provide a simple criterion for classifying levels of the determinants into three groups according to whether each corresponding CI exceeds, crosses, or below the overall mean. They are more appropriate compared to the corresponding CIs based on the treatment contrasts.

The CIs compare percentage of liver cancer deaths in each category with the overall percentage. They applied equitably to each category, whereas the commonly used CIs based on treatment contrasts measured the difference from a reference group that is taken to be fixed and thus does not have a confidence interval.

The ROC curve

A Receiver Operating Characteristic (ROC) curve gives error rate. It plots sensitivity against the false positive rate and shows how well a model predicts a binary outcome.

The ROC curve passes through the upper left corner, providing area under the curve (AUC) close to 1 indicating perfect fit.

Triangulation method

The logistic regression model provided nine province coefficients, 12 coefficients for gender-age group and 16 coefficients for DR-cause location. The nine province coefficients were used to interpolate coefficients for the remaining 67 provinces outside the VA study using a triangulation method. Triangles were drawn linking the nine VA provinces with coefficient values at the vertices of the triangles. The coefficients of these provinces were estimated based on the latitude and longitude of their central points.

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

$$\mathbf{a} + \text{long}_1 \times \mathbf{b} + \text{lat}_1 \times \mathbf{c} = \beta_1 \quad (3)$$

$$\mathbf{a} + \text{long}_2 \times \mathbf{b} + \text{lat}_2 \times \mathbf{c} = \beta_2 \quad (4)$$

$$\mathbf{a} + \text{long}_3 \times \mathbf{b} + \text{lat}_3 \times \mathbf{c} = \beta_3 \quad (5)$$

where $long_1$, $long_2$, and $long_3$ are longitudes of the provinces, lat_1, lat_2, and lat_3 are latitudes of the provinces, β_1, β_2, and β_3 are coefficients of the provinces in the triangle.

The coefficient for any province *j* within a triangle was now given by equation (6) as follows:

$$\textbf{coef(prov}_j\textbf{)} = a + long_j \times b + lat_j \times c. \quad (6)$$

For provinces located outside triangles, their coefficients were obtained similarly by extrapolation.

Extension to the target population

Target population is all deaths aged five year and older in Thailand in 2005. A larger population is all deaths aged five years and older in Thailand from 1996-2009, for which DR data is available. Extending model results to population in 2005 is reasonable if the sample is representative. The VA study sample was representative of the population of reported deaths in Thailand in 2005 because random sampling was used at each stage of the clustering to select 28 districts and at the final step when deaths certificates were selected from these districts. However, when the target population is extended to include years before and after 2005, it must be assumed that the patterns of misreporting of deaths in these years are the same as in 2005. Thus, the VA-estimated for liver cancer deaths from 1996 to 2009 were obtained.

The graphical displays and statistical analyses were performed using R program version 3.0.1 [12].

Results

Figure 4 shows bar chart of percentages of liver cancer deaths by province, age group and DR cause-location superimposed with adjusted percentages and their 95% confidence intervals from the full model. The three factors in the full model are statistically significant (p-value <0.001). The horizontal red line is average percentage (7%). The percentages of liver cancer deaths in UbonRatchatani and Phayao were higher than average. Liver cancer death for males aged 40–69 were higher than average. Among deaths outside hospitals, 33% reported as digestive disease, 25% reported as other digestive cancer, and 10% of other cancers were really due to liver cancer.

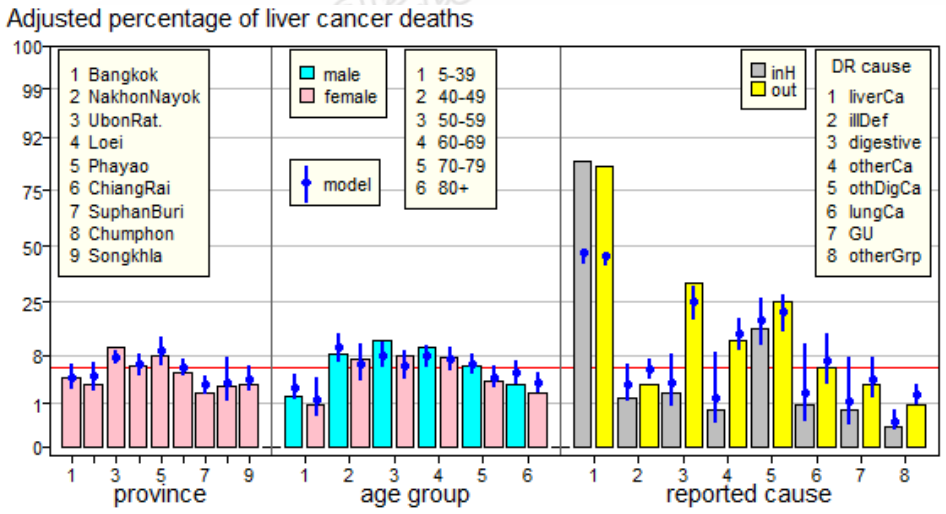


Figure 4 Liver cancer death by province, gender-age and DR cause-location

Figure 5 shows ROC curves of the logistic regression model. The cut-off point in the ROC gives 500 predicted liver cancer deaths, in agreement with the observed number from the VA study, for which the sensitivity is 0.64 and the false positive rate is 0.02. Using the reported cause to predict the true cause has sensitivity 0.47. Only 236 cases out of 500 liver cancer deaths were correctly reported. 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 62.6% sensitivity, 97.9% specificity, and AUC 0.84, whereas the simple model gives 56.4% sensitivity, 98.5% specificity, and AUC of 0.8. The full model reduced the error from 20% to 16%. The full model has the ability to predict the correct cause of liver cancer death slightly better than the simple cross-referencing model.

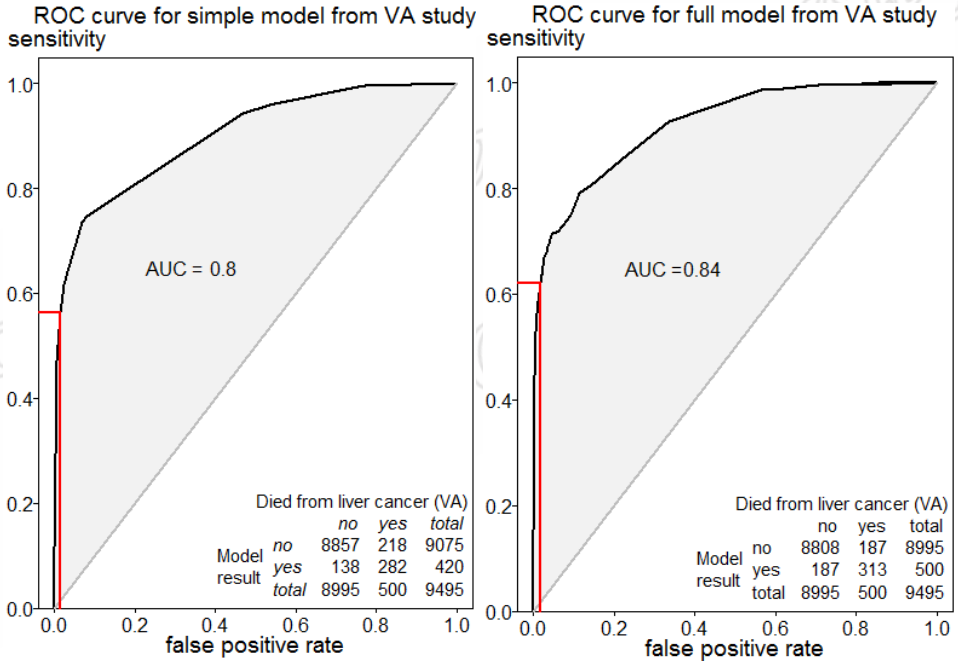


Figure 5 ROC curve for simple and full model from VA study

Figure 6 shows geographical variation of liver cancer death based on the adjusted percentages from the model. The adjusted percentages were classified into three categories. It indicates that liver cancer deaths were more frequent in the north and northeast.

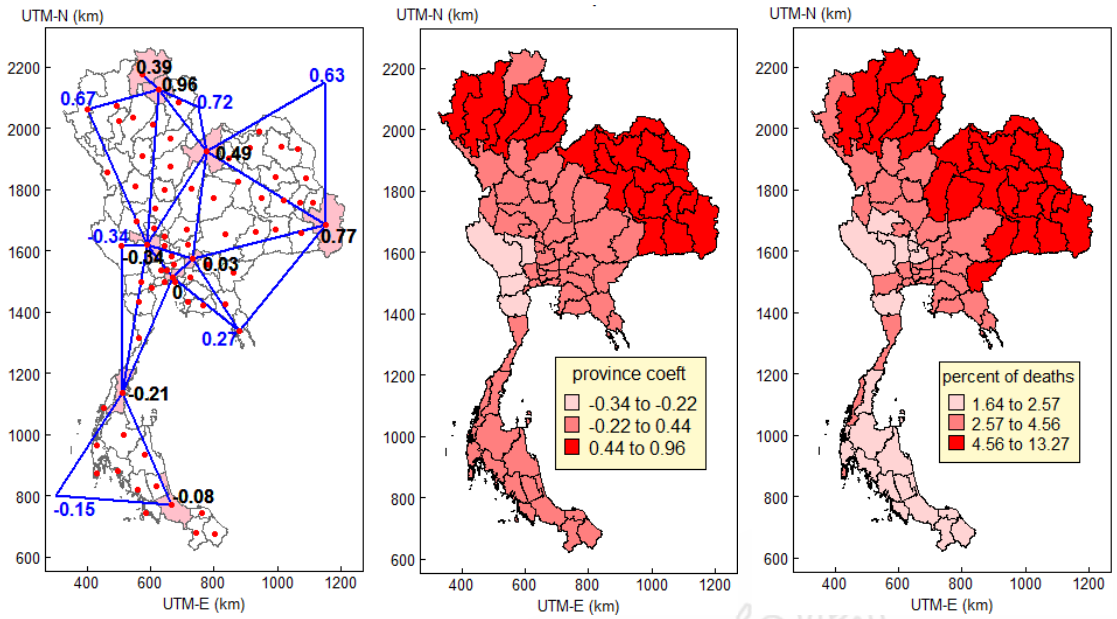


Figure 6 Coefficients for nine provinces from model and eight provinces from interpolated method (left panel), coefficients for every province (middle panel) and adjusted percentages of liver cancer death in 2005 (right panel)

The estimated percentages of liver cancer deaths from the model were applied to the DR data by gender-age groups and DR-cause location, and provinces from 1996 to 2009. The area plots in Figure 7 clearly reveal that numbers of liver cancer deaths were under reported especially for the earlier years.

The total numbers of liver cancer deaths reported for 14 years are 147,458. The estimated total numbers of liver cancer deaths from the simple and full models are 260,508 and 249,922, respectively. The total number of DR reported liver cancer deaths were lower than those estimated by simple model and full model by factors of 1.77 and 1.69, respectively. While simple model gave large proportions of liver cancer deaths at ages below 40 years, these were reduced when full model allowing for province, gender and age was used. For the older age groups, cause of liver cancer deaths was already improved in accuracy by the simple model.

When separated by gender, the total number of DR reported liver cancer deaths were lower than those estimated by factors of 1.58 (simple model) and 1.64 (full model) for males. For females they were lower than those estimated by factors of 2.19 (simple model) and 1.83 (full model), respectively.

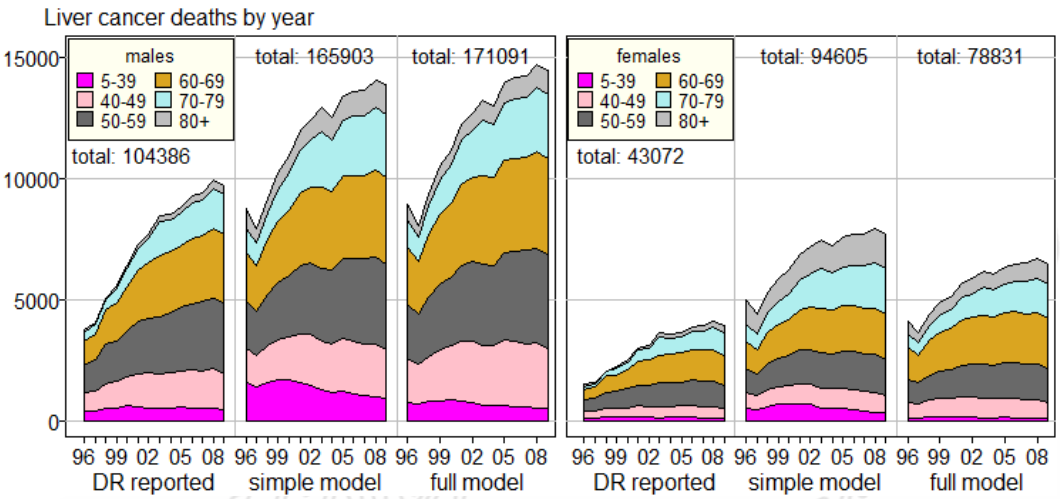


Figure 7 DR reported, simple model estimated and full model estimated of liver cancer deaths by gender-age groups in 1996-2009

Discussion and conclusion

This study illustrates the methods of using the VA data to allocate causes of death into their correct groups in the DR data in 1996 to 2009. The methods comprise fitting logistic regression to liver cancer cause group of the VA data. Logistic regression model of VA-assessed causes of deaths with demographic factors is found to be appropriate to use. It allows for gender-age, province, and DR-cause location predicted causes specific deaths with higher sensitivity and specificity compared to those derived from simple model (simple cross-referencing method).

Logistic regression can assess confounding and interaction effectively when there are several confounders (Hosmer and Lemshow, 2004 [13]). Moreover, it can be used to calculate percentage and its confidence interval, so that the results can be interpreted easily.

The model with sum contrasts are more appropriate compared to the corresponding model based on the treatment contrasts. It provides confidence interval for every category. The confidence interval based on sum contrasts applied equitably to each category, whereas the commonly used confidence intervals based on treatment contrasts measured the difference from a reference group that is taken to be fixed and thus does not have a confidence interval. These confidence intervals are compared with bar charts of sample percentages to assess evidence of confounding bias. ROC cure gives error rates and area plots show results by gender and year.

The logistic model showed that liver cancer deaths often occurred in males aged 40-69 in Payao Province in the north and UbonRatchatanee Province in the northeast. Misreported causes of death for liver cancer mainly occurred for deaths outside hospitals and they were more likely to be digestive disease (ICD-10 codes are K00-K99), other digestive cancer and other cancer (ICD-10 codes are all C with exception for C30-C39 and D00-D48).

When extended, the model results to provinces outside the VA study in 2005, regional patterns of liver cancer mortality have been observed. The estimated liver cancer deaths varied by provinces and they ranged from 1.64 to 13.27% of all cause deaths. They were more likely to occur in provinces of the north and the northeast. This result is supported by the findings in the previous study (Faramnuayphol et al. [14]). They found that people from the upper northeast faces higher deaths from liver cancer with 17 times higher than people from the lower south.

Methodology used in this study is added value to the data in national level and it will help increasing number of mortality studies in developing countries. This study will be

useful for research in liver cancer mortality meta analyses. The model can be extended to the larger target population comprising all deaths in Thailand for longer periods of time and it can be used to forecast liver cancer deaths and other specific causes and compare to other methods (Ugarte et al. [15]).

In conclusion, the methods can be applied to available mortality data similar to the VA data in developing countries where their national vital registration data are of low quality.

Acknowledgements

We would like to thank Professor Don McNeil for his helpful guidance and Dr. Kanitta Bundhamcharoen from Thai Ministry of Public Health for providing us the data. Graduate School, Prince of Songkla University supported scholarship for Nattakit Pipatjaturon and the Commission on Higher Education supported scholarship for Arinda Ma-a-lee.

References

[1] C. D. Mathers, D. M. Fat, M. Inoue, C. Rao and A. D. Lopez, 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) (2005),171-177.

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

[3] C. Rao, Y. Porapakkham, J. Pattaraarchachai, W. Polprasert, N. Swampunyalert and A. D. Lopez, Verifying causes of death in Thailand: rationale and methods for empirical investigation, Population Health Metrics 8 (11) (2010), doi:10.1186/1478-7954-8-11.

[4] V. Tangcharoensathien, P. Faramnuayphol, W. Teokul, K. Bundhamcharoen and S. Wibulpholprasert, A critical assessment of mortality statistics in Thailand: potential for improvements, Bulletin of the World Health Organization 84 (3) (2006), 233-238.

[5] Y. Porapakkham, C. Rao, J. Pattaraarchachai, W. Polprasert, T. Vos, T. Adair and A. D. Lopez, Estimated causes of death in Thailand, 2005: implications for health policy, Population Health Metrics 8(14) (2010), doi:10.1186/1478-7954-8-14.

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

[7] G. Carmichael, Exploring Thailand's mortality transition with the aid of life tables, Asia Pacific Viewpoint, 52(1) (2011), 85-105.

[8] P. Byass, Integrated multisource estimates of mortality for Thailand in 2005, Population Health Metrics 8(10) (2010), doi:10.1186/1478-7954-8-10.

[9] World Health Organization, ICD-10: International Statistical Classification of Diseases and Related Health Problems: tenth revision–2nd ed. 3 v., Geneva, Switzerland, 2004.

[10] P. Tongkumchum and D. McNeil, Confidence intervals using contrasts for regression model, Songklanakarin J. Sci. Technol 31(2) (2009), 151-156.

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

[12] R Development Core Team R. 2012, A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, Retrieved October 26, (2010), from http://www.R-project.org/

[13] Hosmer Jr, D. W., and Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

[14] P. Faramnuayphol, V. Chongsuvivatwong and S. Panarunothai, Geographical Variation of Mortality in Thailand, Journal of the Medical Association of Thailand 91(9) (2008), 1455-60.

[15] M. D. Ugarte, T. Goicoa, J. Etxeberria and A. F. Militino, Projections of cancer mortality risks using spatio-temporal P-spline models, Stat Methods Med Res 21(5) (2012), 545-560.

Appendix III

Article III: "Estimating Lung Cancer Deaths in Thailand based on Verbal Autopsy Study in 2005"

Running title:

ESTIMATING LUNG CANCER DEATHS IN THAILAND

Full title: ESTIMATING LUNG CANCER DEATHS IN THAILAND BASED ON VERBAL AUTOPSY STUDY IN 2005

Authors and Corresponding author information:

First author: Nattakit Pipatjaturon[1,2], M.Sc. (Biostatistics)

Second author: Phattrawan Tongkumchum[2], Ph.D. (Statistics)

Third author: Attachai Ueranantasan[2], Ph.D. (Research Methodology)

Affiliation:

[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.

Corresponding author:

Phattrawan Tongkumchum

Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, 94000, Thailand.

Phone: 66 73 312179

Mobile: 66 86 2876915

e-mail: phattrawan.t@psu.ac.th or phattrawan@gmail.com

Number of tables 1

Number of figures 4

ESTIMATING LUNG CANCER DEATHS IN THAILAND BASED ON A VERBAL AUTOPSY STUDY FROM 2005

Abstract. The causes of death obtained from death certificates in Thailand are incomplete and inaccurate. Therefore, mortality statistics from death registrations (DR) are unreliable. Accurate mortality statistics are essential for national policies on intervention and care, and resource allocation. The Verbal Autopsy (VA) is a more reliable source for cause of deaths than the DR. We investigated the classification of lung cancer deaths in Thailand from 1996 to 2009 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. The misclassified cases were reported mainly as other cancers and respiratory disease. 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 cause of death in the DR database and reliable data such as the VA data are 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

Causes of death statistics are essential for monitoring the health of a nation and identifying priorities. The causes of death data 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*., 2010). Extensive misclassification of causes of death (Tangcharoensathien *et al.,* 2006) makes it necessary for mortality studies in Thailand to estimate numbers of deaths using other data source.

The VA in Thailand was conducted by the 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 effect of gender-age groups and location of the deceased. 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.

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

This study aims to estimate number of lung cancer deaths obtained from the DR during 1996-2009 using VA data from 2005 with a statistical model of lung cancer deaths taking into account demographic and medical factors. Thus, after correction for misclassified lung cancer deaths, a more accurate estimate of lung cancer death can 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. Twenty-eight districts were selected from the provinces similarly. 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 less than 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). The data obtained from each case were the province, gender, age, location of death (in or outside hospital), the DR-reported International Statistical Classification of Diseases (ICD-10) code reported on the death certificate, and the VA-assessed ICD-10 code.

Data analysis

We analyzed 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 for deaths. The groups had to be large enough for statistical analysis. The 21 groups are described elsewhere (Chutinantakul *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 the 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 (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 categorized 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 (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 (Hosmer and Lemshow 2002, Venables and Ripley 2002, McNeil 1996). 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 includes 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 and Ripley, 2002; Tongkumchum and McNeil, 2009; Kongchouy and Sampantarak, 2010; Sampantarak *et al*., 2011) instead of conventional treatment contrasts 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 2006) to show how well the simple and full models predict 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 *et al*., 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 years 1996-2009, the VA-estimated lung cancer deaths from 1996 to 2009 were thus obtained. Graphical displays and statistical analyses were performed using R program 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 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.

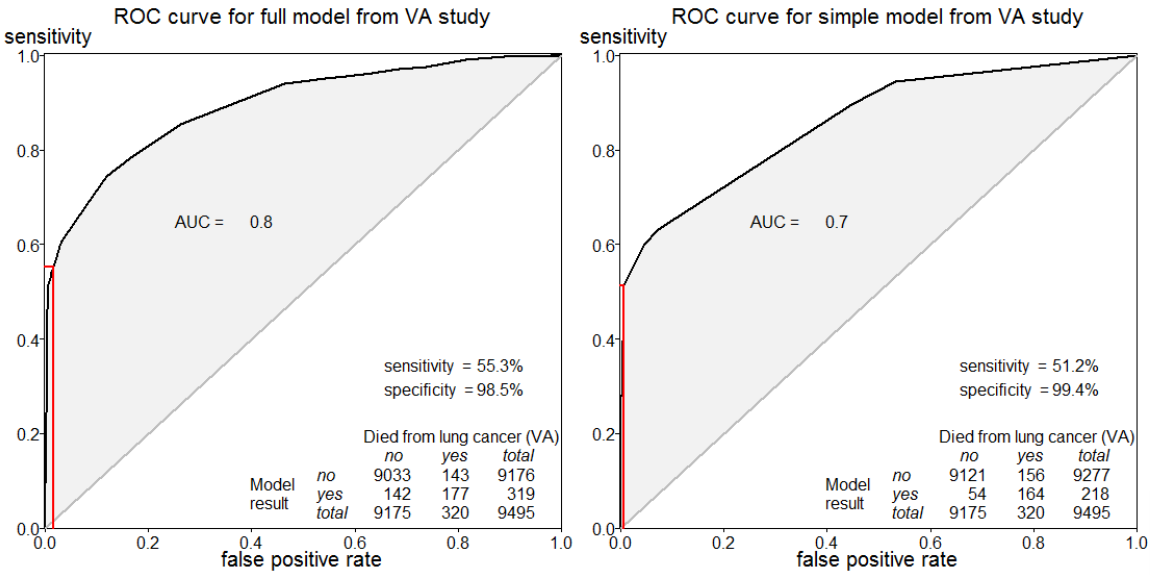


Fig 1. ROC curve for full fitted model and simple fitted model from VA study.

Adjusted percentages of lung cancer deaths

Figure 2 shows 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%). To distinguish the bar chart and 95% confidence interval, a non-linear vertical axis scale was used.

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

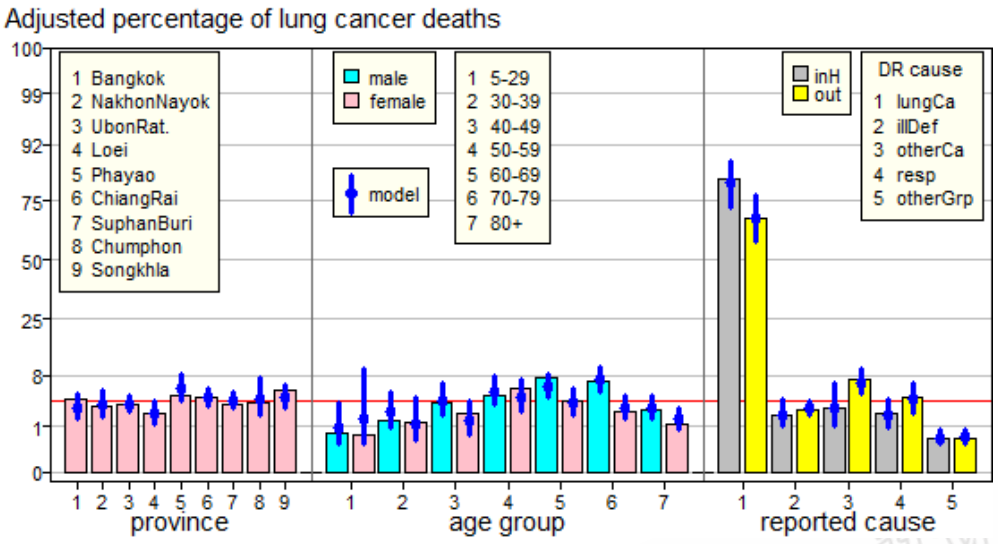


Fig 2. Adjusted percentage of lung cancer death by province, gender-age group and DR cause location.

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 numbers of lung cancer deaths 7,549 for males and 4,796 for females. The full model estimated numbers 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 corresponding DR reports.

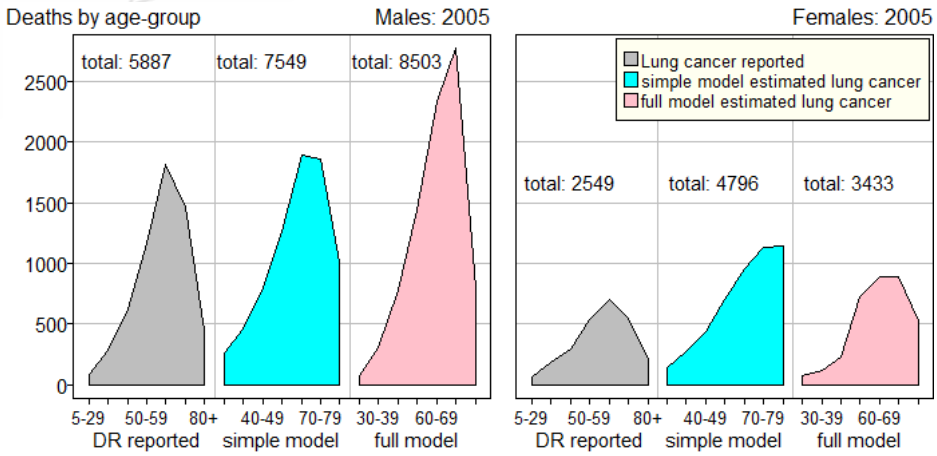


Fig 3. DR reports of lung cancer deaths and estimates from simple and full models in 2005 by age groups.

Figure 4 shows DR reported, simple model estimated and full model estimated numbers of lung cancer deaths by age group and year 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 and 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.

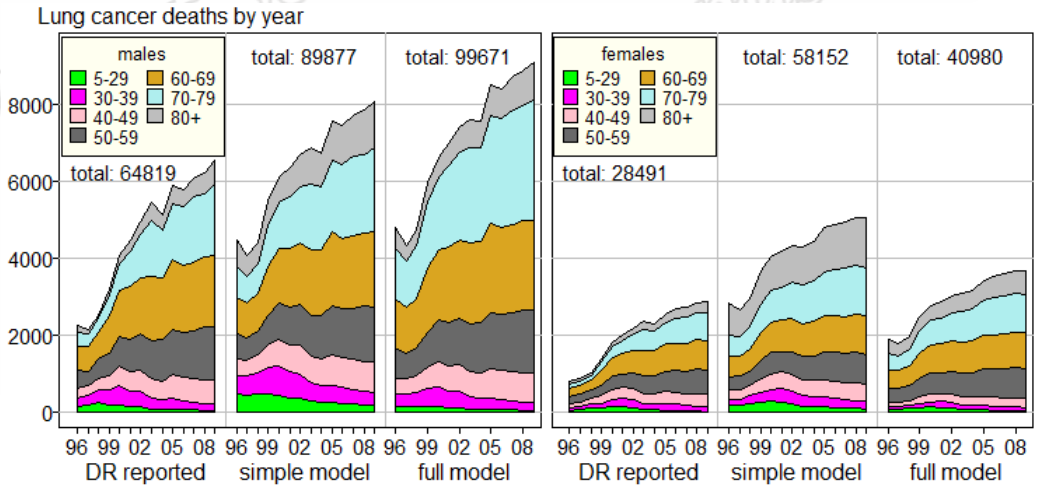


Fig 4. DR reports of lung cancer deaths and estimates from simple and full models of lung cancer deaths by age groups and years.

DISCUSSIONS

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, correct cause of death for individuals is uncertain, particularly for causes being reported as ill-defined or unknown cause. Goodness of fit of the model as assessed by the ROC curve indicates that the model adequately separates lung cancer deaths from others.

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 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 group (C-, D00-D48). The estimated numbers of lung cancer deaths from the models were higher than those reported.

Logistic regression is commonly used in health studies. According to our knowledge, it has not been applied to the VA study. 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 and it can 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 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 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).

No evidence of regional effect was found in this study but 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 also 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 correctly reported and some misclassifications due to other cancers. This agrees with a previous study, where lung cancer deaths were observed not to contribute significantly to ill-defined cancer coding (Porapakkham *et al*., 2010).

CONCLUSIONS

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

ACKNOWLEDGEMENTS

We are grateful to Prof. Don McNeil for guidance, support and assistance. We are also thankful to Dr. Kanitta Bundhamcharoen from Bureau of Policy and Strategy,

Ministry of Public Health Thailand for providing us the data. Finally, we thank to Graduate School, Prince of Songkla University for 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*:25. http://dx.doi.org/10.1186/s12963-014-0025-x.

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,* 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.

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. http://dx.doi.org/10.2349/biij.4.3.e46.

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.

Pattaraachachai, 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*:12. http://dx.doi.org/ 10.1186/1478-7954-8-12.

Polprasert, W., Rao, C., Adair, T., Pattaraachachai, 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*, 8:13. http://dx.doi.org/ 10.1186/1478-7954-8-13.

Porapakkham, Y., Rao, C., Pattaraachachai, 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*:14. http://dx.doi.org/ 10.1186/1478-7954-8-13.

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

Rao, C., Porapakkham, Y., Pattaraarchachai, J., Polprasert, W., Swanpunyalert, N., & Lopez, A. D. (2010). Verifying causes of death in Thailand: rationale and methods for empirical investigation. *Population Health Metrics*, 8:11. 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*, S82-91.

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.

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

Appendix IV Proceeding

Estimating Lung Cancer Deaths in Thailand based on the 2005 Verbal Autopsy Study

Nattakit Pipatjaturon [1], and Phattrawan Tongkumchum [2]

[1]*Office of Disease Prevention and Control, region 9 Phitsanulok, Hua-Ror subdistrict, Muang, Phitsanulok, 65000 Thailand*

Tel: 66879030804 E-mail: nattakit@hotmail.com

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

E-mail: tphattra@bunga.pn.psu.ac.th

Background: Death records in Thailand are currently considered both incomplete and inaccurate. They make lung cancer mortality statistics less reliable. Knowledge of reliable cancer mortality statistics can affect national policies on intervention and care, as well as related resource allocation.

Objectives: To improve quality of cancer deaths from national vital registration (VR) database using verbal autopsy (VA) study and to estimate lung cancer deaths in Thailand in 2005.

Methods: The VA data were 9,495 deaths aged 5 years and older. The cross tabulation of VA and VR causes groups of lung cancer was used to summarize numbers of deaths by cause. Logistic regression model of death due to lung cancer was fitted separately to the data classified by province, gender-age group and the VR cause-location group. Triangulation method was used to interpolate province outside the VA study. Finally, the estimated numbers of lung cancer deaths by province and gender-age groups were obtained.

Results: Lung cancer groups have different regional patterns. Highly statistical significant differences exist between the nine provinces in the VA study. VA estimates of lung cancer both male and female in 2005 are 8,503.4 and 3,433.4 cases. These are

44.4% and 34.7% higher than VR reported totals of 5,887 and 2,549, respectively. The death rates of lung cancer for both sexes are 42.0 and 17.0 per 100,000 populations in north region, respectively.

Conclusion: Reported lung cancer deaths in the target population in 2005 are substantially under-reported. These methods can apply to estimate other causes for further year. Reliable estimation on lung cancer burden can provide essential guidance for the Thai public health authorities of cancer prevention and control.

Keywords: Verbal Autopsy, Lung cancer, Logistic regression, Triangulation method


5th
ICPH-GMS
2013
5th
INTERNATIONAL CONFERENCE ON PUBLIC HEALTH
AMONG GMS COUNTRIES
Evidence-based Health Policy: Current Status and Future Prospects
Venue : National Theater
University of Public Health
University of Medicine (1)
28 - 29 September, 2013

5th International Conference on Public Health
among Greater Mekong Sub-Regional Countries
28-29 SEPTEMBER, 2013

UNIVERSITY OF PUBLIC HEALTH, YANGON, MYANMAR

LETTER OF INVITATION

Mr Nattakit Pipatjaturon
PhD Student,
Department of Mathematics and Computer Science,
Faculty of Science and Technology,
Prince of Songkla University,
Thailand

Date: 10 August, 2013

On behalf of the Organizing Committee, it is a great pleasure for me to inform you that your abstract entitled '**Estimating Lung Cancer Deaths in Thailand based on the 2005 Verbal Autopsy Study**' has been reviewed and accepted for ORAL presentation at 5th International Conference on Public Health among Greater Mekong Sub-regional Countries at University of Public Health, Yangon, Myanmar on 28 – 29 September, 2013. Therefore you are cordially invited to participate in the above mentioned conference. Exact date and time of oral presentation will be announced later.

Yours sincerely,

Professor Dr Nay Soe Maung
Rector of University of Public Health
Chairman of Organizing Committee, 5th ICPHGMS
Postal address: Corner of Bogyoke Aung San Road and Myoma Kyaung Road, Latha Township, Yangon, Myanmar 11131
website: http://www.uph-myanmar.org
e-mail naysoemg26@gmail.com
Off Ph +951 395207, Fax +951 395212

THI - 50

Certificate of Presentation

This certificate is awarded to

Mr Nattakit Pipatjaturon

in recognition of presentation in

"5TH INTERNATIONAL CONFERENCE ON PUBLIC HEALTH AMONG GMS COUNTRIES"

28th - 29th September, 2013
Yangon, Myanmar

Prof. Nay Soe Maung
Rector
University of Public Health
Yangon, Myanmar

Appendix V

R command for democratic confidence intervals for logistic regression model

```
# stroke.Rcm
setwd("d:/nattakit")
read.table("m.txt",h=T,as.is=T) -> m

glm.dci <- function(dat,yID,xIDs,reverse=FALSE,delta) {
 y <- dat[,yID]
 if (reverse==TRUE) y <- 1-dat[,yID]
 xs <- as.data.frame(dat[,xIDs])
 np <- length(xIDs)

 for (j in c(1:np)) {          # fit model with weighted sum contrasts
  nj <- length(unique(xs[,j]))
  nj1 <- tapply(xs[,j],xs[,j],length)
  rho <- 0*(1:nj)
  for (i in c(1:nj)) {
   rho[i] <- nj1[i]/sum(nj1)
  }

  D1 <- rbind(rho,cbind(diag(nj-1),0))
  C1 <- solve(D1)
  C <- C1[,-1]
```

```
 xs[,j] <- as.factor(xs[,j])
 contrasts(xs[,j]) <- C
}
glm(y~.,family=binomial,data=xs) -> mod
summary(mod) -> rez

for (j in c(1:np)) {                # repeat with last two levels reversed
 nj <- length(unique(xs[,j]))
 nj1 <- tapply(xs[,j],xs[,j],length)
 rho <- 0*(1:nj)
 for (i in c(1:nj)) {
  rho[i] <- nj1[i]/sum(nj1)
 }
 D1 <- rbind(rho,cbind(diag(nj-1),0))
 D1[nj,nj-1] <- 0
 D1[nj,nj] <- 1
 C1 <- solve(D1)
 C <- C1[,-1]
 xs[,j] <- as.factor(xs[,j])
 contrasts(xs[,j]) <- C
}
glm(family=binomial,y~.,data=xs) -> mod.a
summary(mod.a) -> rez.a

cf <- rez$coef                      # assemble coefficients and SEs
```

```
cf.a <- rez.a$coef
cfs <- NULL
ses <- NULL
npos <- 0
for (j in c(1:np)) {
 nj <- length(unique(xs[,j]))
 cfs <- c(cfs,cf[npos+c(2:nj),1],cf.a[npos+nj,1])
 ses <- c(ses,cf[npos+c(2:nj),2],cf.a[npos+nj,2])
 npos <- npos+nj-1
}

meanPc <- mean(100*y)              # create adjusted percents & their CIs
k <- -log(100/meanPc-1)
 pcs <- NULL
cilbs <- NULL
ciubs <- NULL
npos <- 0
for (j in c(1:np)) {
 nj <- length(unique(xs[,j]))
 cfj <- cfs[npos+c(1:nj)]
 sej <- ses[npos+c(1:nj)]
 nj1 <- tapply(xs[,j],xs[,j],length)
 rho <- 0*(1:nj)
 for (i in c(1:nj)) {
  rho[i] <- nj1[i]/sum(nj1)
```

```
}
dd <- delta                    # Marquardt damping constant
epsilon <- 0.00005             # convergence criterion
nit <- 20                      # maximum number of iterations
a0 <- 1                        # initial value of constant
a1 <- 1
it <- 0
aDiff <- 1
while ( (abs(aDiff)>epsilon) && ((it <- it+1) < nit) ) {
 adjPc <- ifelse(cfj<=0,100/(1+exp(-k-cfj)),100/(1+exp(-k-a1*cfj)))
 expCoef <- ifelse(cfj<=0,0,exp(-k-a1*cfj))
 F0 <- sum((adjPc/100)*rho) - meanPc/100
 DF0 <- sum(rho*(adjPc/100)^2*expCoef*cfj)
 a1 <- a0-dd*(F0/DF0)
 aDiff <- a1-a0
 a0 <- a1
}
meanj <- sum(adjPc*nj1)/sum(nj1)
DF <- meanPc/meanj
adjPc <- adjPc*DF
pcs <- c(pcs,adjPc)
sej <- ses[npos+c(1:nj)]
cilbs <- c(cilbs,DF*ifelse(cfj<=0,100/(1+exp(-k-cfj+1.96*sej)),
        100/(1+exp(-k-a1*cfj+1.96*sej))))
ciubs <- c(ciubs,DF*ifelse(cfj<=0,100/(1+exp(-k-cfj-1.96*sej)),
```

```
        100/(1+exp(-k-a1*cfj-1.96*sej))))
 npos <- npos+nj
}
if (reverse==TRUE) {
 cfs <- -cfs
 pcs <- 100-pcs
 zz1 <- cilbs
 zz2 <- ciubs
 cilbs <- 100-zz2
 ciubs <- 100-zz1
}
 cbind(cfs,ses,pcs,cilbs,ciubs)
}
str(m)


glm(data=m,family="binomial",y~factor(prov)+factor(SAG)+factor(VR.h)) -> mod1
summary(mod1)
drop1(mod1,test="Chisq") -> rez1
pval <- rez1$"Pr(>Chi)"[2:5]
pval1 <- ifelse(pval[1]<0.0001,"<0.0001",round(pval[1],4))
pval2 <- ifelse(pval[2]<0.0001,"<0.0001",round(pval[2],4))
pval3 <- ifelse(pval[3]<0.0001,"<0.0001",round(pval[3],4))


yID <- 13
xIDs <- c(12,9,15)
```

```
glm.dci(m,yID,xIDs,delta=0.01) -> rez

options(scipen=12)

windows(10,4)                                   # Figure 1
par(oma=c(0,0,0,0),mar=c(2.5,2,3,1),las=1,mgp=c(1.1,0.1,0),tcl=0.2)
n1 <- length(unique(m$prov))
n2 <- length(unique(m$SAG))
n3 <- length(unique(m$VR.h))

ylab <- "Stroke Mortality (%)"
titl <- "Thailand 2005"
xlab1 <- "Province"
xlab2 <- "Gender-Age Group"
xlab3 <- "Cause-Location Group"
xCoord <- c((1:n1),(n1+2):(n1+n2+1),(n1+n2+3):(n1+n2+n3+2))
pc1 <- rez[c(1:n1),3]
pc2 <- rez[(n1+1):(n1+n2),3]
pc3 <- rez[(n1+n2+1):(n1+n2+n3),3]

yCoord <- c(pc1,pc2,pc3)

cilb1 <- rez[1:n1,4]
cilb2 <- rez[(n1+1):(n1+n2),4]
cilb3 <- rez[(n1+n2+1):(n1+n2+n3),4]
```

```
ciub1 <- rez[1:n1,5]
ciub2 <- rez[(n1+1):(n1+n2),5]
ciub3 <- rez[(n1+n2+1):(n1+n2+n3),5]

xmin <- min(cilb1,cilb2,cilb3)
ymax <- max(ciub1,ciub2,ciub3)
ymin <- 0
ymax <- 100

plot(1,type="n",xlim=c(0.5,max(xCoord)+0.5),ylim=c(ymin,ymax),ylab="",xlab="",xa
xt="n")

meanPc <- 100*mean(m$y)
 abline(h=meanPc,col=2)
abline(v=c(n1+1,n1+n2+2),col="dimgrey")
dx <-  0.1
for (i in c(1:n1)) {
 points(xCoord[i]+c(0,0)+dx,c(cilb1[i],ciub1[i]),type="l",lwd=2)
}
for (i in c(1:n2)) {
 points(xCoord[n1+i]+c(0,0)+dx,c(cilb2[i],ciub2[i]),type="l",lwd=2)
}
for (i in c(1:n3)) {
 points(xCoord[n1+n2+i]+c(0,0)+dx,c(cilb3[i],ciub3[i]),type="l",lwd=2)
}
```

```
points(xCoord+dx,yCoord,pch=20)

pcCr1 <- 100*tapply(m$y,m$prov,mean)      # crude percentages
pcCr2 <- 100*tapply(m$y,m$SAG,mean)
pcCr3 <- 100*tapply(m$y,m$VR.h,mean)

points(xCoord[1:n1]-dx,pcCr1,pch=21,bg=3,cex=0.8)
points(xCoord[(n1+1):(n1+n2)]-dx,pcCr2,pch=21,bg=3,cex=0.8)
points(xCoord[(n1+n2+1):(n1+n2+n3)]-dx,pcCr3,pch=21,bg=3,cex=0.8)

axis(side=1,at=c(1:n1),lab=c(1:n1))
axis(side=1,at=c(2,4,6,8),lab=c(2,4,6,8))

axis(side=1,at=n1+1+c(1:n2),lab=c(1:n2))
axis(side=1,at=n1+n2+2+c(1:n3),lab=c(1:n3))
axis(side=1,at=n1+n2+2+c(2,4,6,8,10,12,14,16,18),lab=c(2,4,6,8,10,12,14,16,18))

at1 <- (1+n1)/2
axis(side=1,at=at1,lab=xlab1,tcl=0,padj=1.4)
at2 <- (n1+1+n1+n2+2)/2
axis(side=1,at=at2,lab=xlab2,tcl=0,padj=1.4)
at3 <- (n1+n2+3+n1+n2+n3+2)/2
axis(side=1,at=at3,lab=xlab3,tcl=0,padj=1.4)

text(at1,ymax,adj=c(0.5,1),paste("p-value ",pval1))
```

```
text(at2,ymax,adj=c(0.5,1),paste("p-value ",pval2))
text(at3,ymax,adj=c(0.5,1),paste("p-value ",pval3))

mtext(side=3,line=0.2,adj=-0.04,ylab)
mtext(side=3,line=0.2,adj=1,titl)

lg2 <- "Overall Mean"
legend("topleft",inset=c(0.4,0.15),leg=lg2,lwd=1,col=2,bg="ivory",
       y.intersp=0.8,x.intersp=0.4,cex=0.9)
```

Vitae

Name: Mr Nattakit Pipatjaturon

Student ID: 5420330010

Educational Attainment:

Degree	Name of institution	Year of Graduation
Dip in Nursing Science	College of Nursing Borommaratchachonnani Sawanpracharak	1989
B.P.H. (Public Health)	Sukhothai Thammathirat Open University	1992
M.Sc. (Bisostatistics)	Mahidol University	1999

Scholarship Award during Enrolment

- Scholarship from the Graduate school, Prince of Songkla University, Pattani campus Songkhla, Thailand
- Research Scholarship from Graduate school, Prince of Songkla University, Songkhla, Thailand
- Scholarship for visiting at University of Malaya in Malaysia from the Faculty of Science and Technology, Prince of Songkla University, Songkhla, Thailand

- Scholarship for proceeding from the Graduate school and Faculty of Science and Technology, Prince of Songkla University, Pattani campus Songkhla, Thailand

Work-Position and Address:

Position: Public Health Technical Officer

Address Epidemiology and intelligence section,
Office of disease prevention and control, 2nd Phitsanulok,
Department of Disease Control, Ministry of Public Health

List of Publication and Proceeding:

Publication:

Waeto, S., Pipatjaturon, N., Thongkumchum, P., Choonpradub, C., Saelim, R., and Makaje, N. 2014. Estimating Liver Cancer Deaths in Thailand based on Verbal Autopsy Study. *Journal of Research in Health Sciences*, 14(1): 18-22.

Accepted:

Pipatjaturon, N., Ma-a-lee, A., Thongkumchum, P., and Ueranantasan, A. 2016. Estimating Liver Cancer Deaths in Thailand: Methodologies to Optimize the Use of Verbal Autopsy Data. *Far East Journal of Mathematical Sciences.*

Submitted:

Pipatjaturon, N., Thongkumchum, P., Ueranantasan, A. 2015. Estimating Lung Cancer Deaths in Thailand based on Verbal Autopsy Study in 2005. *Journal of Science and Technology.*

International Conference:

Pipatjaturon Nattakit, N. and Tongkumchum, P. Estimating Lung Cancer Deaths in Thailand based on the 2005 Verbal Autopsy Study. 5th International Conference on Public Health Among GMS Countries 28 – 29 September 2013. Yangon, Myanmar.