Appendix I Article

"A logistic regression model for estimating transport accident deaths

using Verbal Autopsy data"

Asia-Pacific Journal of Public Health

http://aph.sagepub.com/

A Logistic Regression Model for Estimating Transport Accident Deaths Using Verbal Autopsy Data

Nuntaporn Klinjun, Apiradee Lim and Kanitta Bundhamcharoen

The online version of this article can be found at:

Published by:

$SAGE

http://www.sagepublications.com

On behalf of:

Asia-Pacific Academic Consortium for Public Health

Additional services and information for *Asia-Pacific Journal of Public Health* can be found at:

Email Alerts: http://aph.sagepub.com/cgi/alerts

Subscriptions: http://aph.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> OnlineFirst Version of Record - Apr 14, 2014

What is This?

A Logistic Regression Model for Estimating Transport Accident Deaths Using Verbal Autopsy Data

Nuntaporn Klinjun, MSc¹, Apiradee Lim, PhD¹,², and Kanitta Bundhamcharoen, PhD³

Abstract

This study aimed to create an appropriate model using verbal autopsy (VA) data to estimate transport accident deaths from vital registration data in Thailand. A sample of 9644 VA deaths was obtained from the Thai Ministry of Public Health. VA assessed transport accidents accounted for 546 deaths (5.7% of sample). Logistic regression was used to model transport accident deaths classified by 9 provinces, 16 gender–age groups, 14 combinations of vital reported cause groups, and place of death (in or outside hospital). The receiver operating characteristic curve was used to match the number of reported transport accident deaths to the number predicted by the model with sensitivity 73.8% and false positive rate 1.6%. The estimated transport accident deaths ranged from 1.68 to 2.65 times higher than the vital registration data reported according to gender–age groups.

Keywords

transport accident deaths, verbal autopsy, vital registration, estimating model, logistic regression

Introduction

Injuries and deaths from transport accidents, most of which are road traffic accidents, are a major problem around the world. Many of these are preventable.[1,2] The World Health Organization's Global Burden of Disease Project in 2004 estimated that more than 1.3 million people die each year and between 20 and 50 million are injured from road traffic accidents globally. These problems are expected to increase from the ninth leading cause of death in 2004 to be the fifth leading cause of death in 2030. Road traffic accident was one of the leading causes of death and disability in most countries in the Southeast Asia region in 2004.[3]

A survey of 11 countries in Southeast Asia in 2007 reported that Thailand had the highest estimated road traffic accident mortality with the rate of 25.4 per 100 000 population.[3] Data from the Royal Thai Police from 1983 to 2002 revealed that road traffic accident mortality rate had

¹Prince of Songkla University, Muang, Thailand
²Centre of Excellence in Mathematics, CHE, Bangkok, Thailand
³Ministry of Public Health, Nonthaburi, Thailand

Corresponding Author:
Apiradee Lim, Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, 94000, Muang, Thailand.
Email: api_45@hotmail.com

varied substantially from 6.1 per 100 000 population in 1983 and peaked at 28.2 per 100 000 population in 1995 before steadily decreasing to 20.9 per 100 000 population in 2002.[4] In addition, data from the Injury Surveillance System of 28 provincial hospitals from 1998 to 2007 showed that most of injury-caused deaths were from transport accidents.[5,6]

Good-quality data for transport accident deaths are essential to aid health planners with appropriate planning, problem solving, and decision making. However, data from vital registration (VR) in Thailand are of limited use as approximately 40% to 49% of deaths have the cause of death reported as "ill defined," which includes unknown, unspecified causes, and "senility."[7-10] As part of the efforts to reduce the misclassification of the causes of deaths, a verbal autopsy (VA) study was conducted in 2005 by the SPICE project (Setting Priorities using Information on Cost-Effectiveness analysis) to verify causes of deaths reported by VR in Thailand.[10-13] The study was carried out to investigate and determine the actual causes of deaths recorded in the VR system. The use of the 2005 VA data can be substantially improved by applying appropriate systematic biostatistical methods. Thus, this study aimed to generate a model based on the VA data for estimating transport accident deaths reported by VR in Thailand.

In this study, we focused on transport accident deaths with the codes V01 to V99 in the International Statistical Classification of Diseases and Related Health Problems (ICD-10) comprising land transport (V01-V89), water transport (V90-V94), air and space transport (V95-V97) and unspecified transport (V98-V99).[14]

Methods

Data Sources and Management

The main sources of data used for this study were 2 databases, namely the VA study and the VR, which is based on death certificates without autopsy, both obtained from the Bureau of Health Policy and Strategy, Thai Ministry of Public Health.

In 2005, a VA study comprised the deceased person's age, gender, province, cause of death (ICD-10 coded in death certificate and VA-assessed code) and place of death (in or outside hospital). The VA study was conducted for 9 provinces: Chiang Rai, Phayao, Loei, Ubon Ratchathani, Nakhon Nayok, Bangkok, Suphanburi, Chumphon, and Songkhla. The details of selecting provinces and districts are explained by Rao et al.[11] In this study, a total of 9644 deaths, out of which 546 were from transport accident, were analyzed. The causes of deaths were derived from the ICD-10 Mortality Tabulation List 1, which consists of 103 cause categories. Twenty leading registered causes of death in Thailand in 2005 were considered in the VA study according to the mortality distribution.[11] We grouped contiguous ages into 10-year age groups: 0 to 9, 10 to 19, 20 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, and ≥70 years. This was combined with gender to form gender-age groups for reducing gender and age group interaction.

Data management for systematic analysis of the VA data began with the use of the chapter-block classification of ICD-10 (the chapters consist of 21 blocks divided mainly by human body systems),[14] which we used to group causes of death after consultation with the Epidemiology Unit, Faculty of Medicine, Prince of Songkla University. We then created 22 major cause groups with VA counts. To ensure statistical accuracy, we aimed to have at least 200 deaths in each group, except for septicemia (overreported), perinatal and congenital child death (confined to ages 0-4 years), and suicide as groups deserving special attention.

Of these major cause groups, we found that VR causes of transport accident deaths could be attributed to 11 VA causes: "all other" (156 deaths), "other injuries" (64 deaths), "ill-defined" (33 deaths), "stroke" (13 deaths), "respiratory" (8 deaths), mental and nervous" (5 deaths), "septicemia" (3 deaths), "other CVD" (2 deaths), "endocrine" (1 death), "digestive" (1 death) and "suicide" (1 death). We then classified causes of deaths from these 11 groups into 7 groups. The 6 most

Table 1. Likelihood Ratio Test for Logistic Regression Model of Transport Accident Deaths by Province, Gender–Age Group and Vital Registration (VR) Cause–Location Group.

Factors	Deviance	df	P Value
Province	28.92	8	.0003
Gender–age group	152.50	15	<.001
VR cause–location group	1524.65	13	<.001
Error	1468.10	9607	

likely risk groups and other groups combined: "transport accident," "other injuries," "stroke," "respiratory," "ill defined," "all other," and "other groups" (the remaining cause groups). A new variable was created by combining VR prediction groups with the 2 locations of death (in hospital and outside hospital), which we named VR cause–location group.

Statistical Methods

The determinants in this study were province, gender–age group, and medical (VR cause–location group) components. The outcome was transport accident death coded as 0 (other death) and 1 (transport accident death).

Logistic regression is a model for analyzing binary outcomes with one or more predictors. This model formulates the logit of the probability P that a person died from a transport accident as an additive linear function of the 3 determinant factors as follows:

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

In this model P_{ijk} is the probability of death due to transport accident. μ is a constant and the terms α_i, β_j and γ_k are individual parameters that specify 9 provinces ($i; i = 1, 2, 3, \ldots, 9$), 16 gender–age groups ($j; j = 100, 110, \ldots, 170, 200, \ldots, 270$), and 14 VR cause–location groups ($k; k = 110, 120, \ldots, 170, 210, \ldots, 270$), respectively.

A receiver operating characteristic (ROC) curve was used to diagnose how well a model predicts a binary outcome.[15,16] After that, the model was applied to estimate the number of transport accident deaths in 2005 from VR database with a total of 395 374 deaths in Thailand. Of these, there were 10 914 deaths from transport accidents.

Data management and statistical analysis were performed using R statistical software.[17]

Results

Of the 9644 deaths in the VA study sample, 546 deaths were from transport accident (5.7%).

Estimation of Model

A logistic regression model with the 3 determinants: province, gender–age group, and VR cause–location group were created. Each determinant had a P value less than .001 indicating that the associations of the 3 factors with transport accident deaths were significant. The model yielded a χ^2 of 1468.1 with 9607 degrees of freedom, and a P value of approximately 1, as shown in Table 1.

Figure 1 shows bar charts of the crude percentages of deaths and model-based 95% confidence intervals of the adjusted percentage of VA-assessed estimates from transport accidents for
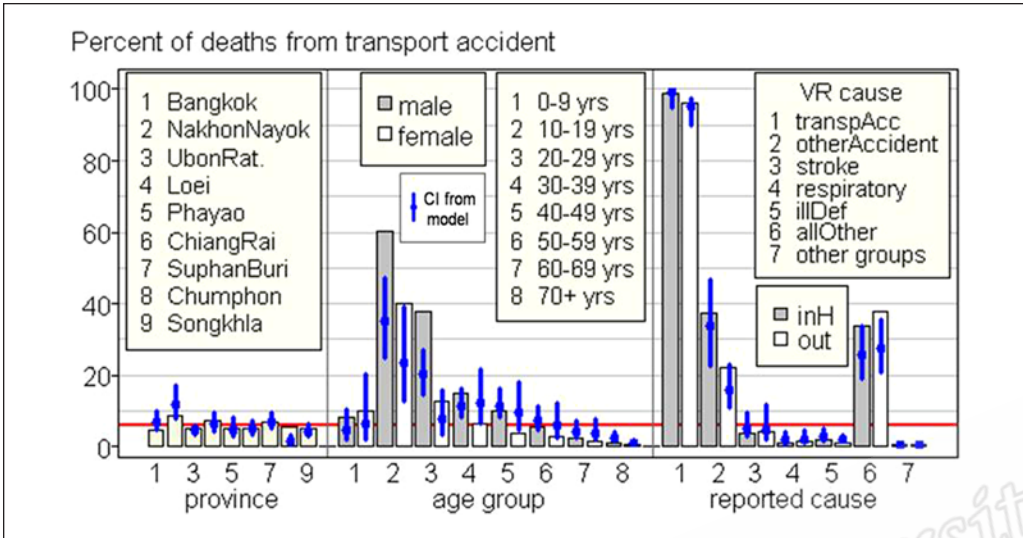
Figure 1. Bar chart for transport accident deaths and 95% confidence intervals for adjusted percentages of the verbal autopsy (VA) assessed estimate of transport accident deaths by province, age group, and reported cause of deaths.

each of the 3 determinants compared with the overall percentage of deaths from accidents (shown as horizontal line). From the graphs, it appears immediately evident that age group and gender have a major impact on the probability of being killed in a transport accident, but province play a lesser rate. More than 95% of deaths from transport accident were consistent with the VR report. The figure clearly shows that there is no confounding after adjusting for all factors.

The logistic regression model provided an accurate measure for predicting the possibility of transport accident deaths, with the sensitivity of 73.8% and for predicting the false positive rate of 1.6% (a specificity of 98.4%). The area under the ROC curve (AUC) of 0.97 indicated excellent model accuracy.

Results from the model revealed that transport accident deaths were underreported in VR data in 2005. The estimated deaths (VA) for both males and females were higher than the reported deaths in VR data by between 1.68 and 2.65 times, as shown in Table 2.

Discussion and Conclusion

Mortality statistics and causes of death are important for public health utility. Therefore, public health policies and planning for preventing transport accident mortalities should be based on correct data. A previous study on VA data by the SPICE project showed that there were 2.8% of transport accident deaths from VR system of which VR cause matched VA cause for 47.4% and the accuracy of VR was 97%. Thus, reported transport accident deaths in VR were underreported by an estimate of 52.6%, of which 6.1% were ill defined and 46.5% were other causes of deaths. Most of underlying reason of transport accident deaths was coded as "all other" group, which registered with unspecified event and undetermined intent (Y34; 27.1%). Hence, the use of the 2005 VA data should be substantially improved by applying appropriate biostatistical methods for estimating the true number of transport accident deaths from VR. Similarly, many studies tried to link VA and VR for providing the true causes of deaths.[18-20]

The issue of the quality and the availability of VR data is a global phenomenon.[8-10,13,18] Many methods have been used for estimating the correct causes of deaths.[10-13,18,19,21] In this study, we

Table 2. The Ratio of Estimated Deaths From the Model and Reported Deaths in Thailand in 2005.

Gender–Age Group	Estimated Deaths (VA)	Reported Deaths (VR)	Ratio (IF)
Male			
0-9 years	371	184	2.02
10-19 years	3621	1523	2.38
20-29 years	5228	2538	2.06
30-39 years	3404	1719	1.98
40-49 years	3057	1313	2.33
50-59 years	1735	823	2.11
60-69 years	947	485	1.95
≥70 years	771	291	2.65
Female			
0-9 years	281	120	2.34
10-19 years	801	344	2.33
20-29 years	605	359	1.68
30-39 years	862	379	2.27
40-49 years	772	324	2.38
50-59 years	544	229	2.37
60-69 years	433	171	2.53
≥70 years	258	112	2.30

Abbreviations: VA, verbal autopsy; VR, vital registration; IF, inflation factor.

used the correct causes of deaths from VA data to create the model for estimating the correct number of transport accident deaths. For VA study, all causes of deaths in nine provinces in Thailand were sampled. The logistic regression model used 95% confidence intervals for adjusted percentages of VA-assessed transport accident deaths. The unadjusted and adjusted percentage of transport accident deaths gave the same results, showing no confounding in the sample.

However, we described a suitable logistic regression model for estimating reliable transport accident deaths based on VA data covered with 3 determinants: 9 provinces, 16 gender–age groups, and 14 VR cause–location groups. Accuracy of the model was improved by using ROC curve, which denoted the predicted outcome as 1 (transport accident) if $P \geq c$, or 0 (other death) if $P < c$. We chose a cutoff point at $c = 0.333$, which gave 546 predicted transport accident deaths as the most accurate level based on the VA study. This model gave results for corrected causes of transport accident deaths with high sensitivity (73.8%), low false positive rate (1.6%), and high AUC (0.97). These methods have been used extensively for estimating and evaluating models in other studies.[22,23] The model used in our study was created from VA data that was conducted to correct for the proportion of transport accident deaths in VR data. According to the model, the number of deaths from unspecified causes was reduced whereas the number of deaths from specified causes was increased which adjusted the correct causes of deaths in the VR report. After applying the model to VR data for 2005, we confirmed that the number of transport accident deaths in the VR system was substantially underreported for all gender–age groups. Therefore, this model is accurate for estimating the proportion of transport accident deaths in VR system in Thailand.

Acknowledgments

Declaration of Conflicting Interests

Funding

References

1. World Health Organization. *Strategic Approaches for Injury Prevention and Control in the South-East Asia Region*. New Delhi, India: World Health Organization; 2011.
2. Arokiasamy JT, Krishnan R. Some epidemiological aspects and economic costs of injuries in Malaysia. *Asia Pac J Public Health*. 1994;7:16-20.
3. World Health Organization. *Regional Report on Status of Road Safety: The South-East Asia Region*. New Delhi, India: World Health Organization; 2009.
4. Tanaboriboon Y, Satiennam T. Traffic accidents in Thailand. *IATSS Res*. 2005;29:88-100.
5. Koedklai S, Dumnakeaw K, Kumwongsa A. Epidemiology of injuries and deaths caused by road traffic accidents in Thailand [in Thai]. *J Health Res Syst*. 2009;3:598-605.
6. World Health Organization. *Intercountry Workshop on Injury Surveillance*. New Delhi, India: World Health Organization; 2009.
7. Mathers CD, Ma Fat D, 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;83:171-177.
8. Tangcharoensathien V, Faramnuayphol P, Teokul W, Bundhamcharoen K, Wibulpholprasert S. A critical assessment of mortality statistics in Thailand: potential for improvements. *Bull World Health Organ*. 2006;84:233-238.
9. Rukumnuaykit P. Mortality and causes of death in Thailand: evidence from the survey of population change and death registration. *Asia Pac Popul J*. 2006;21:67-84.
10. 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 Metrics*. 2010;8:13.
11. Rao C, Porapakkham Y, Pattaraarchachai J, Polprasert W, Swampunyalert N, Lopez AD. Verifying causes of death in Thailand: rationale and methods for empirical investigation. *Popul Health Metrics*. 2010;8:11.
12. Pattaraarchachai J, Rao C, Polprasert W, et al. Cause-specific mortality patterns among hospital deaths in Thailand: validating routine death certification. *Popul Health Metrics*. 2010;8:12.
13. Porapakkham Y, Rao C, Pattaraarchachai J, et al. Estimated causes of death in Thailand, 2005: implication for health policy. *Popul Health Metrics*. 2010;8:14.
14. World Health Organization. Special tabulation lists for mortality and morbidity: mortality tabulation list 1. In *International Statistical Classification of Diseases and Health Related Problems–Tenth Revision* (Vol 1). Geneva, Switzerland: World Health Organization; 2004;1163-1175.
15. Westin LK. *Receiver Operating Characteristic (ROC) Analysis: Evaluating Discrimance Effects Among Decision Support Systems*. Umeå, Sweden: Umeå University, Department of Computer Science; 2001.
16. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. Hoboken, NJ: Wiley; 2000.
17. R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2011. http://www.r-project.org. Accessed January 20, 2012.
18. Setel PW, Whiting DR, Hemed Y, et al. Validity of verbal autopsy procedures for determining cause of death in Tanzania. *Trop Med Int Health*. 2006;11:681-696.
19. Setel PW, Sankoh O, Rao C, et al. Sample registration of vital events with verbal autopsy: a renewed commitment to measuring and monitoring vital statistics. *Bull World Health Organ*. 2006;83:611-617.

20. Mudenda SS, Kamocha S, Mswia R, et al. Feasibility of using a World Health Organization standard methodology for Sample Vital Registration with Verbal Autopsy (SAVVY) to report leading causes of death in Zambia: results of a pilot in four provinces, 2010. *Popul Health Metrics*. 2011;9:40.
21. Jacobs G, Aeron-Thomas A, Astrop A. *Estimating Global Road Fatalities*. Berkshire, UK: Global Road Safety Partnership; 2000.
22. Wang JC, Huan SK, Kuo JR, Lu CL, Lin H, Shen KH. A multivariable logistic regression equation to evaluate prostate cancer. *J Formos Med Assoc*. 2011;110:695-700.
23. Kononen DW, Flannagan CAC, Wang SC. Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. *Accid Anal Prev*. 2011;43:112-122.

Appendix II Article

"Estimating external causes of death in Thailand 1996-2009 based on the

2005 Verbal Autopsy study"

Original Article

Estimating external causes of death in Thailand 1996-2009 based on the 2005 Verbal Autopsy study

Nuntaporn Klinjun[1]*, Apiradee Lim[1], and Kanitta Bundhamcharoen[2]

*[1] Department of Mathematics and Computer Science, Faculty of Science and Technology,
Prince of Songkla University, Mueang, Pattani, 94000 Thailand.*

*[2]International Health Policy Program, Ministry of Public Health,
Mueang, Nonthaburi, 11000 Thailand.*

Abstract

This study aimed to develop models based on Verbal Autopsy (VA) data and to estimate correct number of deaths from external causes in Thailand from 1996 to 2009. Logistic regression was used to create models of the three external causes of death classified by province, gender-age group and Vital registration (VR) cause-location group. Receiver operating characteristic (ROC) curves were used to validate the models by matching the number of reported deaths to the number of deaths predicted by the models. The models provided accurate prediction results, with false positive error rates 1.6%, 2.0% and 0.6% and sensitivities 73.8%, 46.3% and 62.0%, respectively. The results reveal that under-reporting of external causes of death increased over the 14-year period. Our statistical method confirms that the Thai 2005 VA data can be used to estimate external causes of death from VR report in Thailand to allow for the under-reporting rate.

Keywords: external causes of death, verbal Autopsy, vital registration, logistic regression, ROC

1. Introduction

Within recent decades, external causes of death, deaths due to accident and violence (World Health Organization, 2004), have had a large impact in terms of premature deaths and potential years of life lost in all regions of the world. The Global Burden of Disease (GBD) database in 2000 reported that injuries accounted for 9% of global deaths and comprised 12% of the GBD. The injury-related disease burden is expected to rise in rank in 2020, particularly for deaths from road traffic accidents, homicide, war and suicide (Peden *et al.*, 2002). The GBD in 2004 reported that injuries accounted for 10% of global deaths and is expected to rank among the 20 leading causes of death in the world by 2030, particularly

in deaths from road traffic accidents, suicide and homicide (World Health Organization, 2010). Road traffic accident deaths are expected to increase to be the fifth-leading cause of death in 2030, with suicide and homicide rising to be the 12th and 18th leading causes of death, respectively (World Health Organization, 2010).

In Thailand in 2004, external causes of death due to road traffic accidents, homicide and violence, suicide and drowning among males caused disability-adjusted life year (DALY) loss in the 2nd, 12th, 14th and 20th rank, respectively, whereas road traffic accidents were the 7th leading cause of DALY loss for females. Furthermore, road traffic accidents in both males and females caused the burden of diseases to increase DALYs lost in 2004 over and above the 1999 level (Bundhamcharoen *et al.*, 2011). The VR data in 2005 showed that road traffic accidents and suicide were among the 20 leading causes of all deaths (Porapakkham *et al.*, 2010). In 2008, Thailand was in the first four countries in Asia with

* Corresponding author.
  Email address:

high mortality rates from injuries. Over 40% of all injury deaths were due to road traffic accidents, followed by all other injuries and suicide (OECD/ World Health Organization, 2012). Many deaths from external causes are controllable and largely preventable. However, a major obstacle to these issues is incorrect data on cause of death recorded in the VR system in Thailand. The VR system is the important source of mortality data for the whole country but the problem of the quality of data on specific causes of death has been regularly acknowledged (Mathers *et al.*, 2005; Tangcharoensathien *et al.*, 2006; Rao *et al.*, 2010; Porapakkham *et al.*, 2010). In Thailand, about 40-49% of death certificates reported the cause of death as "ill-defined", which includes unknown or unspecified causes and senility, and this significantly limits their public utility (Mathers *et al.*, 2005; Tangcharoensathien *et al.*, 2006; Polprasert *et al.*, 2010). To address this problem, the Thai Ministry of Public Health (MOPH) used the VA method in 2005 to critically assess vital statistics data and improve the quality of cause of death recorded at registration in Thailand (Rao *et al.*, 2010).

VA is a tool for obtaining cause of death when medical diagnoses are not available (Maude and Ross, 1997; Setel *et al.*, 2006; Mudenda *et al.*, 2011). Thus, the VA study in Thailand in 2005 conducted by the Setting Priorities using Information on Cost-Effectiveness analysis (SPICE) project was used for verifying causes of deaths and to determine the actual causes of deaths recorded in the VR system in Thailand by gender, age group and province (Rao *et al.*, 2010; Porapakkham *et al.*, 2010; Polprasert *et al.*, 2010; Pattaraarchachai *et al.*, 2010). The results from this study showed that Transport accident deaths of VR increased from 2.8% to 6.4% for deaths in hospital (Pattaraarchachai *et al.*, 2010) and increased from 2.7% to 5.2% for deaths outside hospital (Polprasert et al., 2010). Transport accident was the second leading cause of death in males (8.1%) (Rao *et al.*, 2010). Suicide deaths of VR increased from 1.3% to 1.8% for deaths outside hospital (Polprasert *et al.*, 2010).

This study used the number of deaths from the Thai 2005 VA study to create models for estimating external causes of death from the VR database in Thailand, and thus to systematically estimate the corresponding number of external causes of death in Thailand from 1996 to 2009.

2. Methods

2.1 Data sources and management

External causes of death in this study focused on transport accidents (V01-V99), other accidents (W, X00-X59) and suicide (X60-X84) as classified by the International Statistical Classification of Diseases and Related Health Problems (ICD-10) (World Health Organization, 2004).

Data from the Thai 2005 VA study and the VR databases from 1996 to 2009 were obtained from the Bureau of Health Policy and Strategy, MOPH. The Thai 2005 VA study data had 9,644 samples of deaths recorded with the deceased

person's gender and age, province, cause of death (ICD-10 coded in death certificate and VA-assessed code) and place of death (in hospital or outside hospital) (Rao *et al.*, 2010). The data were a clustered sample taken from 28 districts from nine provinces in Thailand 2005. The provinces include Chiang Rai, Phayao, Loei, Ubon Ratchathani, Nakhon Nayok, Bangkok, Suphanburi, Chumphon and Songkhla (Figure 1). The nine provinces were selected from Bangkok and two provinces from each of the four broad regions: Northeast, North, Central and South, by ranking provinces by numbers of reported deaths and selecting one province above and one below the median similar to the method for selecting the 28 districts. Finally, 50% of reported deaths were selected from urban areas and 50% from rural areas within these districts using simple random sampling (Rao *et al.*, 2010). Grouping of the gender-age group depended on the age distribution of deaths from each injury cause of death. Transport accident deaths and other accident deaths were divided into eight groups: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69 and 70 years and over. Suicide was divided into six groups: 5-29, 30-39, 40-49, 50-59, 60-69 and 70 years and over. These were combined with gender to form gender-age groups.

Data management for systematic analysis of the VA data began with the use of the classification of ICD-10 for grouping causes of death after consultation with the Epidemiology Unit, Faculty of Medicine, Prince of Songkla University. Thus, we created 22 major cause groups with VA counts. To ensure statistical accuracy, we structured the groups so that, as far as possible, there were at least 200 deaths in each group, except for septicemia (over-reported), perinatal and congenital child death (confined to ages 0-4) and suicide as groups deserving special attention, as shown in Table 1.

Of these major cause groups, we found that the VR cause of other accidents deaths could be attributed to 16 VA causes: "ill-defined" (97 deaths), "all other" (41 deaths),

Figure 1. Map showing the nine provinces of the Thai 2005 VA study

Table 1. VA counts for 22 major causes of deaths in Thailand, 2005: The VA study

No	Cause groups		VA data	No	Cause groups		VA data
1	TB	(A15-19)	195	12	Stroke	(I60-69)	1,076
2	Septicemia	(A40-41)	77	13	Other CVD	(I⁻)	544
3	HIV	(B20-24)	514	14	Respiratory	(J)	802
4	Other Infectious	(A⁻, B⁻)	222	15	Digestive	(K)	490
5	Liver Cancer	(C22)	500	16	Genitourinary	(N)	412
6	Lung Cancer+	(C30-39)	320	17	Child Mortality	(P, Q≤5yrs)	97
7	Other Digestive	(C15-26⁻)	290	18	Ill-defined	(R)	502
8	Other Cancer	(C⁻, D0-48)	704	19	Transport Accident	(V)	546
9	Endocrine	(E)	649	20	Other Accident	(W, X0-59)	341
10	Mental, Nervous	(F, G)	228	21	Suicide	(X60-84)	158
11	Ischemic	(I20-25)	617	22	All other*		360

+ all respiratory and thoracic, ⁻ excluding cause groups listed above

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

"stroke" (14 deaths), "respiratory" (11 deaths), "septicemia" (8 deaths), "genitourinary" (5 deaths), "suicide" (5 deaths), "transport accident" (4 deaths), "digestive" (3 deaths), "other cancer" (2 deaths), "endocrine" (2 deaths), "IHD" (2 deaths), "other infectious diseases" (1 death), "other digestive cancer" (1 death), "mental and nervous" (1 death) and "other CVD" (1 death). We then classified causes of deaths from these 16 groups into 8 groups: the 7 most likely risk groups, and other groups combined: "other accidents", "ill-defined", "stroke", "respiratory", "septicemia", "genitourinary", "all other" and "other groups" (the remaining cause groups). A new variable was created by combining VR prediction groups with the two places of death (in hospital and outside hospital) which we named VR cause-location group. In the same way, VR causes of suicide and transport accident deaths could be attributed to 13 VA causes and 11 VA causes, respectively, Then we classified the cause of death into 4 groups for suicide and 7 groups for transport accident.

2.2 Statistical methods

From 9,644 deaths in the Thai 2005 VA study, the determinants were province, gender-age group and VR cause-location group. The outcomes of applying the models were binary values which were coded 1 and 0 to represent "the cause of death of interest" and other (for example, transport accident, other accidents and suicide), respectively. Logistic regression was used to create three models for estimating external causes of death: transport accident, other accidents and suicide with regional (province), demographic (gender-age group) and medical component (VR cause-location group) from the VR database in Thailand. Each model formulates the logit of the probability P that a person died from each external causes of death as an additive linear function of the determinant factor. The simple logistic regression model is formulated as

$$In\left[\frac{P_k}{1-P_k}\right] = m + g_k \qquad (1)$$

In each model P_k is the probability of transport accident death, m is a constant and g_k specified the VR cause-location group (k). The simple model (1) is compared with the full model (2) as an additive linear function of the three determinant factors as follows:

$$In\left[\frac{P_{ijk}}{1-P_{ijk}}\right] = m + a_i + b_j + g_k \qquad (2)$$

In each model P_{ijk} is the probability of transport accident death and the terms a_i, b_j and g_k refer to province (i), gender/age group (j) and VR cause-location group (k), respectively.

The p-value for a variable in a logistic regression model is Prob[$\chi^2 > D$], the tail area of a chi-squared distribution with k-1 degrees of freedom (df), where k is the number of levels and D is the reduction in deviance (a measure of lack of fit of the model) achieved by the variable. The models were then scaled by appropriate constants to yield 546 transport accident deaths, 341 other accidents deaths and 158 suicide deaths, corresponding to the totals in the VA study.

The 95% confidence interval (CI) of adjusted percentages for each variable was used for adjusting fitted outcome values to reduce the effects of confounding bias arising from covariates associated with both the binary outcome and the variables (McNeil, 1996). The confidence intervals provide standard errors for the differences between each variable level and their overall mean. A method for doing this is described by Tongkumchum and McNeil (2009) and by Kongchouy and Sampantarak (2010).

The ROC curve is used to validate the logistic regression model (Hosmer and Lemeshow, 2000; Westin, 2001). This method is an important step to evaluate the validity of the

model that estimates the same number of each external causes of death as in the VA study. In this study, the ROC methods present a statistical analysis of how well a model predicts a binary outcome.

According to the VA study, the province variable of the logistic regression model contains only nine of Thailand's 76 provinces. Therefore, a method is needed to extend the model-estimated province coefficients to the other 67 provinces. The coefficients of the nine provinces in the models were used for interpolation and triangulation method (Denzin, 2010; Waeto *et al*., 2014) for the coefficients of the other provinces. Triangles were used to predict results for provinces outside the VA study. Their coefficients were estimated, based on the latitude (*lat*) and longitude (*long*) of their central points. The nine VA provinces were linked using triangles; the co-ordinates at the vertices were set using Bangkok as a reference point. Each of these triangles creates an area covering multiple provinces. The coefficient for each point within each triangle can be calculated from its relationship to the values of each of its vertices.

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

$$a + long P_1 \times b + lat\ P_1 \times c = \beta_1 P_1 \quad (3)$$

$$a + long P_2 \times b + lat\ P_2 \times c = \beta_2 P_2 \quad (4)$$

$$a + long P_3 \times b + lat\ P_3 \times c = \beta_3 P_3 \quad (5)$$

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

$$a + long P_j \times b + lat\ P_j \times c = \beta_j P_j$$

In each triangle *P* is province and *β* is coefficient. Coefficients for provinces outside triangles are obtained similarly by extrapolation. To compare differences in coefficients of cause of death across the province thematic maps were used. Provinces were classified into three groups, according to whether the coefficient was (1) totally below the lower quartile, (2) crossing the median, or (3) totally above the upper quartile. The thematic map was used to display this information using corresponding colours, (1) white (2) pink (3) red.

After obtaining the appropriate model, the logistic regression models were applied to estimate the number of deaths for each of the three cause groups: transport accidents, other accidents and suicide from the VR database during 1996 to 2009. Data management and statistical analysis were performed using R statistical software (R Development Core Team, 2011).

3. Results

In the 9,644 VA study transport accident, other accidents and suicide accounted for 546 (5.7%), 341 (3.5%)

and 158 (1.6%) deaths, respectively. Of these, the VR cause of death matched the VA cause for 259 (47.4%), 143 (41.9%) and 90 (57.0%), respectively. The accuracy of death registries were 259 deaths of 267 reported deaths (97.0%) in transport accident, 143 deaths of 237 reported deaths (60.3%) in other accidents and 90 deaths of 107 reported deaths (84.1%) in suicide. Therefore, the death registry in VR reports were under-reported for transport accidents by 52.6%, of which 11.7% were other accidents, 6.0% were ill-defined and 34.8% were other causes of death. Reported other accident deaths in the VR were under-reported by an estimated 58.1%, of which 28.5% were ill-defined and 29.6% were other causes of death. Reported suicide deaths in the VR were under-reported by an estimated 43.0%, of which 11.4% were ill-defined and 31.7% were other causes of death, as shown in Figure 2.

3.1 Estimation of models

Logistic regression models, based on the Thai 2005 VA study with the three determinants were created. The determinants had a statistically significant p-value less than 0.05 for associations with the three cause groups. However, the p-value for the province determinant associations with other accidents exceeded 0.05, indicating no evidence for any difference in the proportion of other accidents among the nine provinces, as shown in Table 2.

Figure 3 shows bar charts of deaths for each cause and model-based 95% confidence intervals of VA-assessed estimates for the three cause group deaths by province, gender-age group and reported cause and location of death. The overall mean is shown as a red line. The bar charts show percentages of the three cause group deaths corresponding to each determinant. The figure shows relatively small province effects. The highest proportion of transport accident deaths was in Nakhon Nayok province, males aged 10-19
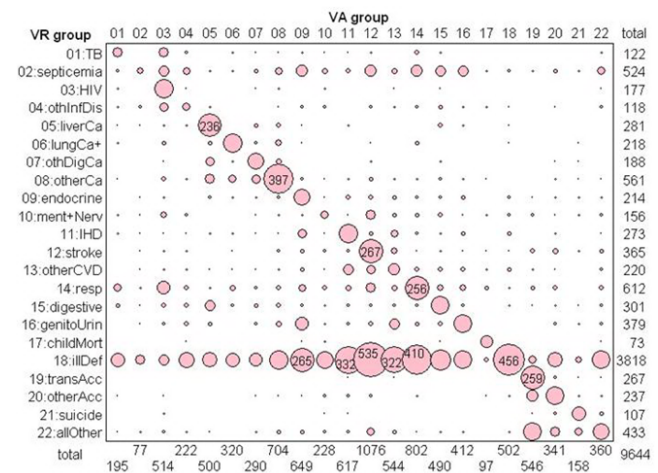

Figure 2. Association between the VA counts and the VR counts from 22 causes of death. Areas of bubbles denote numbers of deaths

Table 2. Likelihood-ratio test for three logistic regression models from deaths of the three cause groups by province, gender-age group and VR cause-location group

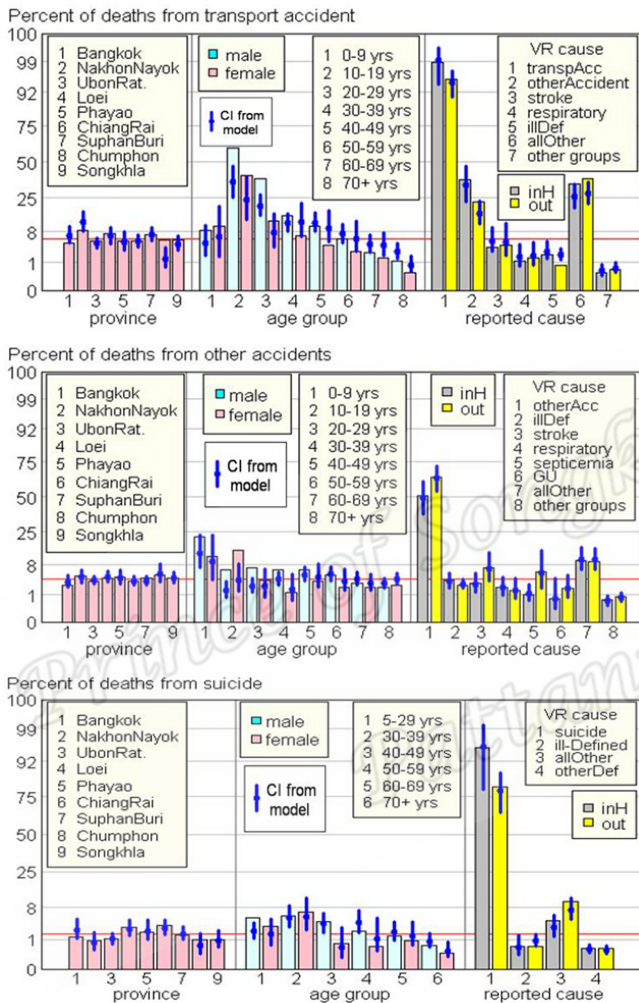| Variables	Transport Accident	Other Accidents	Suicide
Province	0.0003***	0.8430	0.0324*
Gender-age group	<0.001***	0.0011**	<0.001***
VR cause-location group	<0.001***	<0.001***	<0.001***

***p-value ≤ 0.001, **p-value ≤ 0.01, *p-value ≤ 0.05

Figure 3. Bar charts of deaths for each cause and 95% CI of adjusted percentages of the three VA-assessed estimates of cause of death by province, gender-age group and reported cause and location of death. To improve clarity a folded cube root scale is used on the vertical axis to expand tick marks close to 0 and 100

years had the highest percent of transport accident deaths, and more than 95% of transport accident deaths both in and outside hospital were consistent with the VR reported. The highest proportion of other accident deaths was in Chumphon province, males aged 0-9 years had the highest

percent of other accident deaths, and more than 50% of other accident deaths both in and outside hospital were consistent with the VR reported. The highest proportion of suicide deaths was in Chiang Rai province. Females aged 30-39 years had the highest percent of suicide deaths. More than 80% of suicide deaths both in and outside hospital were consistent with those VR reported.

The blue vertical line segments denote 95% CI of the adjusted percentage of VA-assessed external causes of death by province, gender-age group and reported cause of death, showing some confounding due to correlation between determinants.

Figures 4 and 5 show ROC curves for the three logistic regression models comparing the simple models with the full models. The full logistic regression models with three determinants predicted the proportions of deaths in three cause groups substantially better than the simple logistic regression model, with areas under the ROC curve (AUC) of 0.965, 0.832 and 0.935, sensitivity of 73.8%, 46.3% and 62.0% and false positive rates of 1.6%, 2.0% and 0.6%, respectively.
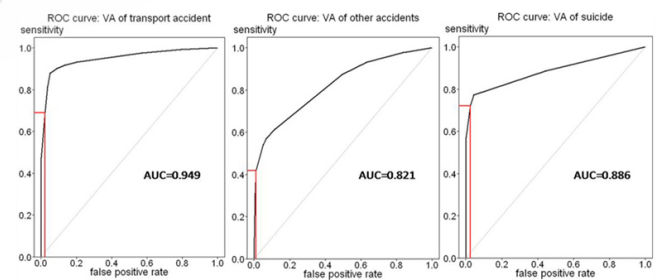

Figure 4. ROC curves for the simple models of the three cause groups
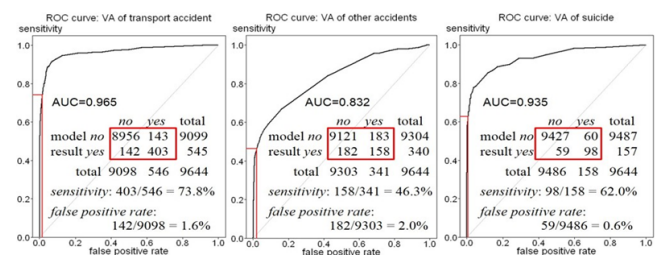

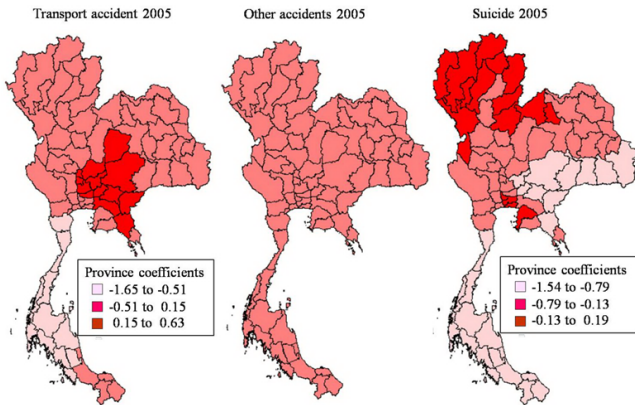Figure 5. ROC curves for the full logistic regression models of the three cause groups

Figure 6. Thematic map of values interpolated the coefficients from the logistic regression model for all 76 provinces for the three cause groups in 2005. No evidence of a regional effect was found for other accident deaths

Figure 6 shows thematic maps of values interpolated for all 76 provinces. The three cause groups had different regional patterns. Transport accidents and suicide in 2005 show highly statistical significant differences between provinces. The highest proportions of transport accident deaths occurred in most of the Eastern and Central region, two provinces of the North-Eastern region and one province of the Bangkok Metropolitan area. The highest proportions of suicide deaths occurred in most of the Northern provinces, three provinces of the Bangkok Metropolitan area, two provinces of the North-Eastern region and one province of the Eastern region. There is no evidence for any differences between other accident death proportions in 2005.

Results from the models revealed that reported deaths from VR data from 1996 to 2009 in both males and females were under-reported. The under-reporting of the three cause groups increased over the 14-year period, as shown in Table 3.

4. Discussion and Conclusions

The quality of data on cause of death of the VR system is important information for public health utility for predicting, controlling and preventing external causes of death. Thus, valid solutions to problems of external causes of death

should be based on correct data.

The Thai 2005 VA study found that a number of deaths for unspecified cause such as ill-defined/unknown and unspecified septicemia required adjustment to achieve the correct levels of cause of death in the VR report (Rao *et al.*, 2010). A method for quality controlling measures in the Thai 2005 VA study is described by Rao *et al.* (2010). The application of a simple cross-referencing method on VR and VA data revealed that after reclassification in the VA study, ill-defined deaths and unspecified septicemia deaths of VR decreased from 39.6% to 5.2% and from 5.4% to 0.8%, respectively. Validation and validity of VA has two major categories: physician-certified verbal autopsy (PCVA) and computerized coding of verbal autopsy (CCVA) (Chandramohan, 2011). Each method found that the VA method can be a useful tool for generating the actual causes of death in a population without a complete VR system, and an estimate based on machine learning yield that VA is 75% accurate when measured against clinical gold standards (Flaxman *et al.*, 2011; Fligner *et al.*, 2011; Riley, 2011). However, physician-certified deaths in natural disease processes which have the same signs, symptoms and assumptions found inaccurate cause of death in up to 40% of VA data (Riley, 2011). These results suggest that the VA study can be used for improving cause of death and verifying accuracy of causes of deaths from VR databases. Similarly, other studies were linked to VA and VR for improving accuracy of cause of death (Setel *et al.*, 2005; Setel *et al.*, 2006; Porapakkham *et al.*, 2010; Mudenda *et al.*, 2011; França *et al.*, 2011). Porapakkham *et al.* (2010) used VA data for estimating causes of death from the VR database in Thailand, 2005. Their study used a capture-recapture method and matching deaths to estimate the total deaths by age and sex, and used proportionate mortality distribution by age, sex and cause of death for adjusting data from the VR database (Porapakkham *et al.*, 2010).

In this study we fitted logistic models to the three outcome cause groups. The full logistic regression models with province, gender- age group and VR cause-location group were more accurate for predicting cause of death corresponding to the determinants of the Thai VA study in 2005 than the simple logistic regression models with VR cause-location groups. The p-value for the province determinant showed no differences between proportions of non-transport accident deaths between the nine provinces consistent with the thematic map. These models accurately allocate deaths

Table 3. Ratios of estimated deaths from models and reported deaths in Thailand in 2005

Period	Transport Accident		Other Accident		Suicide	
	Male	Female	Male	Female	Male	Female
1996-2004	2.03	2.13	1.13	1.78	1.56	1.38
2005	2.16	2.23	1.28	2.10	1.62	1.34
2006-2009	2.27	2.26	1.32	2.22	1.66	1.37

into cause groups by gender age-group and province with high sensitivity and low false positive rate in transport accident and suicide, acceptable sensitivity and low false positive rate in other accidents. In our study, the results of logistic regression models showed that the full logistic regression model can be used for estimating the number of deaths from the three cause groups of external cause of injuries using the same information as provided by the Thai VA study in 2005.

This study predicted numbers of deaths only in year 2005, and the model gave a good fit. This method can be applied for predicting the number of true deaths before and after 2005, provided it is assumed that the patterns of misreporting of deaths in these years are the same as in 2005 when the VA study was undertaken. According to the three models, the VA/VR inflation factor for each cause group is the ratio of the estimated number of deaths to the number of reported deaths in the VR database, and thus is a measure of under-reporting bias in the VR data for each cause group. Transport accident deaths have higher under-reporting biases and smaller gender differences than other accident and suicide deaths. Transport accident deaths and other accident deaths in males were higher than in females over the three periods except suicide deaths in females were higher than in males over the three periods. The under-reporting bias of the three cause groups increased over the 14-year period. Under-reporting of transport accidents, other accident and suicide deaths could occur from misclassification of the causes of death (Pattaraarchachai et al., 2010; Polprasert et al., 2010). These findings are in agreement with the previous studies conducted by Tangcharoensathien et al. (2006) and Prasart-kul and Vapattanawong (2006). These studies reported that deaths outside hospital were recorded by village heads with limited knowledge and expertise in reporting the cause of death and some deaths in hospital were recorded by physicians from their extensive use of vague codes. The study conducted by Mathers et al. (2005) revealed that the highest percentage of using vague codes such as ill-defined deaths form 115 countries was found in Thailand during 1950-2000. These results show that the Thai 2005 VA study is another death data source for verifying the actual external causes of death from VR data in Thailand.

This paper provides an example for improved utility of the Thai 2005 VA data by applying appropriate biostatistical methods. The results from this study should be useful for investigating external causes mortality spatial patterns by gender-age groups, province and year in Thailand from 1996 to 2009.

Acknowledgements

References

Bundhamcharoen, K., Odton, P., Phulkerd, S. and Tang-charoensathien, V. 2011. Burden of disease in Thailand: changes in health gap between 1999 and 2004. BMC Public Health. 11, 53-61.

Chandramohan, D. 2011. Validation and validity of verbal autopsy procedures. Population Health Metrics. 9 (22), doi:10.1186/1478-7954-9-22.

Denzin, N. 2010. The fundamentals. In An Introduction to Triangulation, D. Rugg, editor. UNAIDS Monitoring and Evaluation Fundamentals, Geneva, Switzerland., pp 12.

Flaxman, A.D., Vahdatpour, A., Green, S., James, S.L. and Murray, C.J.L. 2011. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. Population Health Metrics. 9(29), doi:10.1186/1478-7954-9-29.

Fligner, C.L., Murray, J. and Roberts, D.J. 2011. Synergism of verbal autopsy and diagnostic pathology autopsy for improved accuracy of mortality data. Population Health Metrics. 9(25), doi:10.1186/1478-7954-9-25.

França, E., Campos, D., Guimarães, M.D. and Marinho, M.F. 2011. Use of verbal autopsy in a national health information system: Effects of the investigation of ill-defined causes of death on proportional mortality due to injury in small municipalities in Brazil. Population Health Metrics. 9(39), doi:10.1186/1478-7954-9-39.

Hosmer, D.W. and Lemeshow, S. 2000. Applied Logistic Regression, 2nd ed, John Wiley & Sons, Inc., Hoboken, New Jersey, U.S.A., pp. 162.

Kongchouy, N. and Sampantarak, U. 2010. Confidence intervals for adjusted proportions using logistic regression. Modern Applied Science. 4, 2-7.

Maude, G.H. and Ross, D.A. 1997. The effect of different sensitivity, specificity and cause-specific mortality fractions on the estimation of differences in cause-specific mortality rates in children from studies using verbal autopsies. International Journal of Epidemiology. 26, 1097-1106.

Mathers, C.D., Ma Fat, D., 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, 171-177.

McNeil, D. 1996. Epidemiological Research Methods, John Wiley and Sons Ltd, England, pp.305.

Mudenda, S.S., Kamocha, S., Mswia, R., Conkling, M., Sikanyiti, P., Potter, D., Mayaka, W.C. and Marx, M.A. 2011. Feasibility of using a World Health Organization-standard methodology for sample vital registration with verbal autopsy (SAVVY) to report leading

causes of death in Zambia: results of a pilot in four provinces, 2010. Population Health Metrics. 9(40), doi:10.1186/1478-7954-9-40.

OECD/ World Health Organization. 2012. Health at a Glance: Asia/Pacific 2012. Retrieved June 20, 2013, from http://dx.doi.org/10.1787/9789264183902-en.

Pattaraarchachai, J., Rao, C., Polprasert, W., Porapakkham, Y., Pao-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), doi:10.1186/1478-7954-8-12.

Peden, M., McGee, K.and Sharma, G. 2002. The injury Chart Book: a graphical overview of the global burden of injuries, World Health Organization, Geneva, Switzerland. pp. 7-18.

Polprasert, W., Rao, C., Adair, T., Pattaraarchachai, 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. doi:10.1186/1478-7954-8-13.

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

Prasartkul, P. and Vapattanawong, P. 2006. The completeness of death registration in Thailand: evidence from demographic surveillance systrm of the Kanchanaburi project. World Health and Population. 8, 43-51.

R Development Core Team. 2011. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for statistical Computing. Available from: http://www.r-project.org [September, 6 2012]

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), doi:10.1186/1478-7954-8-11.

Riley, L. 2011. Computer-based analysis of verbal autopsies: revolution or evolution?. Population Health Metrics. 9(26), doi:10.1186/1478-7954-9-26.

Setel, P.W., Sankoh, O., Rao, C., Velkoff, V.A., Mathers, C., Gonghuan, Y., Hemed, Y., Jha, P. and Lopez, A.D. 2005. Sample registration of vital events with verbal autopsy: a renewed commitment to measuring and monitoring vital statistics. Bulletin of the World Health Organization. 83, 611-617.

Setel, P.W., Whiting, D.R., Hemed, Y., Chandramohan, D., Wolfson, L.J., Alberti, K.G.M.M and Lopez, A.D. 2006. Validity of verbal autopsy procedures for determining cause of death in Tanzania. Tropical Medicine and International Health. 11, 681-696.

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

Tongkumchum, P. and McNeil, D. 2009. Confidence intervals using contrasts for regression model. Songklanakarin Journal of Science and Technology. 31, 151-156.

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, 18-22.

Westin, LK. 2001. Receiver operating characteristic (ROC) analysis: evaluating discriminance effects among decision support systems. UMINF report, Department of Computer Science, Umea University, Sweden: Available from: http://www.cs.umu.se/research/reports/2001/018/part1.pdf [June, 11 2012]

World Health Organization. 2004. International Statistical Classification of Diseases and Health Related Problems Tenth Revision Volume 1, 2nd ed. World Health Organization, Geneva, Switzerland, pp. 977-1066.

World Health Organization. 2010. Injuries and violence: the facts. World Health Organization, Geneva, Switzerland, pp. 2-6.