

Correcting Misreported Multinomial Outcome Data Based on Logistic Regression Model with Application to Stroke Mortality in Thailand

Arinda MA-A-LEE^{1,2}, Nattakit PIPATJATURON³ and Phattrawan TONGKUMCHUM^{1,*}

¹*Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand*

²*Centre of Excellence in Mathematics, Commission on Higher Education, Bangkok 10400, Thailand*

³*The Office of Diseases Prevention and Control 2nd Phitsanulok, Phitsanulok 65000, Thailand*

(* Corresponding author's e-mail: phattrawan.t@psu.ac.th)

Received: 22 July 2016, Revised: 28 January 2017, Accepted: 16 February 2017

Abstract

Causes of death in Thailand are misreported; about 40 % of deaths have been recorded as “ill-defined”. This study aims to describe statistical methods to correct misreported multinomial outcome by using verbal autopsy (VA) data. Since the outcome is a nominal variable, with 21 levels, the appropriate model for systematic analysis of death by ICD-10 code is multinomial regression. Moreover, it is simpler and more informative to separately fit logistic regression models to the 21 outcome cause groups, and then rescale the results to ensure that the total number of estimated deaths for each group match those reported in the corresponding populations. This method also gives confidence intervals for percentages of deaths in cause groups for levels of each risk factor, adjusted for other risk factors. These confidence intervals are compared with bar charts of sample percentages, to assess evidence of confounding bias. The methods were illustrated using stroke deaths. Area plots are used to show results by gender, age group, and year. The most misclassified stroke deaths were ill-defined, other cardio vascular disease, mental and nerve (outside-hospital), septicemia, and respiratory disease (in-hospital).

Keyword: Logistic regression, misreported deaths, multinomial regression, stroke deaths, Thailand

Introduction

Mortality statistics are important for measuring health status at global and national levels. The World Health Organization (WHO) has used mortality statistics as indicators for global health [1]. National mortality statistics reflect the quality of life of people, and provide useful information for priority setting and resource allocation in a country.

Data quality is a main issue to provide accurate mortality statistics, especially in developing countries. In Thailand, under-registration and ill-defined causes of death are often reported [2-4], which lead to some degree of misclassification of causes of death. For example, septicemia is over-reported, while stroke and others are under-reported [5].

A verbal autopsy (VA) study has been widely used for the assessment of cause of death in countries where death registration (DR) systems are fragile and most people die at home without medical certification of cause of death. 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 causes of death collected are of high quality [6-9].

Estimated causes of death based on the VA data have been published. The data were analyzed using simple cross-referencing tabulation [7,8] and statistical modeling [10-12]. A benefit of using statistical

models is that several factors can be investigated, and the effect of one factor can be adjusted for other factors. Moreover, the statistical models can detect confounders that bias the results. The logistic regression model has been used for correcting mortality in recent studies. They reported that HIV [10], transport accidents [11], and liver cancer deaths [12] were substantially under-reported. The main focus of these studies was on medical aspects, rather than on describing the methods. This study aims to describe systematic methods for correcting misreported deaths. The methods involve analysis of a 2005 VA sample. The multinomial outcome is of causes of death comprising 21 categories. Since the outcome is nominal variable with 21 levels, the appropriate model for systematic analysis of death by ICD-10 code is multinomial regression [13]. Multinomial regression is an extension of binary logistic regression. It can be used when the outcome has more than 2 categories. For outcomes with 21 categories, estimating a multinomial model will have 20 logit equations. It is undesirable to estimate a model as complex as this. Moreover, the interpretation of results from such a complex model is not straightforward [14].

However, it is simpler, and more informative, to separately fit logistic regression models to the 21 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. The estimates from logistic regression are less efficient. The standard errors of the estimates are larger than estimates from the multinomial model. However, it has been shown that the efficiency loss is minor [13,15]. Therefore, the systematic methods used for correcting misreported deaths in this study were based on logistic regression. The methods were illustrated using stroke mortality. A previous cross-reference analysis of the 2005 VA data reported that stroke was a leading cause of death in Thailand, causing 10.7 % of deaths [8]. Moreover, stroke was reported as a leading cause of disability and death among people aged 45 years and older [16]. Stroke prevalence also varies with demographic and geographic factors [17].

Materials and methods

Verbal autopsy (VA) and death registration (DR) data were used. These data were obtained from the Bureau of Health Policy and Strategy, Ministry of Public Health. The VA survey was carried out in 2005 by the Setting Priorities using Information on Cost-Effectiveness (SPICE) analysis project 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 9 provinces. Districts were selected by 2-stage stratification sampling, where Bangkok and pairs of provinces from 4 regions were first randomly selected. Stratification was based on the number of deaths in the regional province or district. Then, a number of death certificates to be assessed were randomly selected from the 28 districts using Probability Proportional to Size (PPS) method.

For this study, the sample was reduced to 9,495 deaths aged 5 years and older. We used the chapter-block classification of ICD-10 codes based on mortality tabulation [18], creating 21 major cause groups. These groups required adequate sample sizes for statistical analysis. Groups with small counts were combined into a larger group identified as "All other". The sample sizes varied from 77 (0.8 %) for septicemia (ICD code A40-A41) to 1,076 (11.3 %) for stroke (ICD code I60-I69).

The cause of death, the outcome, was a nominal variable, with 21 levels, as shown in **Table 1**. It could be analyzed using either a multinomial regression model or the logistic regression model. Since logistic regression was appropriate for binary outcome, the cause of death was considered as 21 binary variables. We created only 2 cause groups, where one group was the cause of interest, and the other group was an aggregation of deaths from all other causes. For example, consider the outcome of interest as TB. There were 195 deaths due to TB, coded as 1, and 9,300 deaths due to all other causes, coded as 0. We next consider the outcome of interest as septicemia. There were 77 deaths due to septicemia, coded as 1, and 9,418 deaths due to all other causes, coded as 0. Next, the outcome of interest is stroke, with 1,076 cases, coded as 1, and 8,419 deaths due to all other causes, coded as 0; and so on.

Table 1 Cause of deaths comprising 21 major cause groups.

Cause of death	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

The determinants were gender-age group, location of death, province, and registered cause on the death certificates for a particular outcome. Gender and age group were combined. The number of levels in the gender-age group factor depended on the age distribution of deaths from the selected outcome. For stroke, we chose 12 levels of gender-age group factors, with 2 gender and 6 age groups (5 - 39, 40 - 49, 50 - 59, 60 - 69, 70 - 79, and 80+).

Similarly, the number of levels in the DR-cause location factor depended on the number of such reported cause groups that affected the outcome cause group. For stroke, we choose 18 levels of a DR-cause location factor, with 2 locations (in or outside hospitals), and 9 DR-reported ICD-10 coded cause groups most likely to be reported as stroke (stroke, ill-defined, septicemia, respiratory disease, mental and nervous, other cardiovascular diseases, endocrine, all other, and other groups).

The province factor had 9 levels, corresponding to the 9 provinces in the VA sample (Chiang Rai, Phayao, Loei, Ubon Ratchathani, Supan Buri, Nakhon Nayok, Bangkok, Chumpon, and Songkla).

Simple logistic regression is a model with one determinant. The simple model with only DR-cause location group as a determinant was first fitted to the data. The full model [19] formulated the logit of the probability of death due to the specified cause group as an additive linear function of the 3 determinant factors, as follows.

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

where, p_{ijk} is the probability of the specified cause group in each of the i, j and k groups of determinant factors, μ is a constant, and $\alpha_i, \beta_j,$ and γ_k refer to effects of the province, gender-age group, and DR-cause location, respectively.

Eq. (1) can be inverted to give an expression for the probability, as follows.

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

The model was fitted to the data using weighted sum contrasts [20-23]. This model also gave confidence intervals for percentages of stroke deaths for levels of each risk factor, adjusted for other risk factors. The confidence intervals, based on weighted sum contrasts, had an advantage, in that they provided a simple criterion for classifying levels of the factor into 3 groups, according to whether each corresponding confidence interval exceeded, crossed, or was below the overall percentage. They were more appropriate compared to the corresponding confidence intervals based on the treatment contrasts. The confidence intervals compared the percentage of the specified cause group in each category with the overall percentage. Each of the categories was equitably applied, whereas the commonly-used confidence intervals, based on treatment contrasts, measured the difference from a reference group that was taken to be fixed, which did not have a confidence interval. These confidence intervals were compared with bar charts of sample percentages to assess evidence of confounding bias.

A Receiver Operating Characteristic (ROC) curve was used to assess the model's ability to distinguish between adverse outcome (stroke deaths) and others [24]. The ROC curve gives error rates. It plots sensitivity against the false positive rate to show how well a model predicts a binary outcome. It gives the proportion of positive outcomes correctly and incorrectly predicted by the model. Area under the curve (AUC) is considered to be the standard method to assess the accuracy of the model. There are several scales for the AUC value interpretation but, in general, ROC curves with AUC close to 1 indicates well fit [25]. The ROC curves of simple and full models were compared. The ability of the full model to predict an adverse outcome reflected the effects of gender-age group and province on reducing error rates in predicting the adverse outcome.

Since only 9 of 76 provinces were in the VA survey, a spatial triangulation method [10-12,26] was used to interpolate the province effects outside the VA study.

The model results were applied to number of deaths in the DR data and, thus, the estimated numbers of stroke deaths were obtained. VA/DR inflation factors (IF) greater than one reflected under-reporting. Area plots were used to show results by gender, age group, and year, as well as DR-cause location. In summary, steps of estimating number of deaths for the whole of Thailand were as follows.

- Step 1: Gather VA data in 2005 with 9,495 deaths aged 5 years and older
- Step 2: Identify 21 major cause groups (21 binary outcomes)
- Step 3: Perform cross tabulation between VA- and DR-cause groups
- Step 4: Separately fit logistic regression models to the 21 outcome cause groups
- Step 5: Estimate coefficients for provinces outside the VA study
- Step 6: Estimate probabilities of the specified cause group for 76 provinces
- Step 7: Apply the probabilities to the DR data in 1996 - 2009
- Step 8: Obtain VA-assessed deaths in 1996 - 2009

To illustrate these methods, we applied them to stroke deaths. Graphical displays and statistical analyses were performed using the R program [27].

Results and discussion

Results

The VA data comprised 9,495 deaths aged 5 year and older, of which 1,076 deaths (386 in-hospital and 690 outside hospital deaths) verified stroke as cause of death. The 9 DR-cause groups most likely to be stroke were stroke (267), ill-defined (535), septicemia (60), respiratory disease (45), mental and nerve (42), other CVD (35), all other (31), endocrine (13), and the rest, with a small number of cases aggregated to other groups (48).

The p-values based on the logistic regression model were statistically significant for DR-cause location, gender-age group, and province. **Figure 1** shows bar charts of percentages of stroke deaths, superimposed with adjusted percentages and their corresponding 95 % confidence intervals. The horizontal red line is the average percentage of stroke deaths (11.3 %).

The 95 % confidence intervals of percentages of stroke deaths in Bangkok, Suphan Buri, and Songkhla were higher than the average. The adjusted percentages of stroke deaths for males in age groups 60 - 79 years, and for females in age groups 70 years and older, were above the average.

The misreported causes of stroke outside hospital were mental and nerve, ill-defined, and other CVD. For inside hospital, the misreported cause was all other groups. For outside hospital, 37.1 % of mental and nerve, and 25.7 % of other cardiovascular diseases, were deaths due to stroke. On the other hand, in hospital, 22.5 % of all other causes were deaths due to stroke.

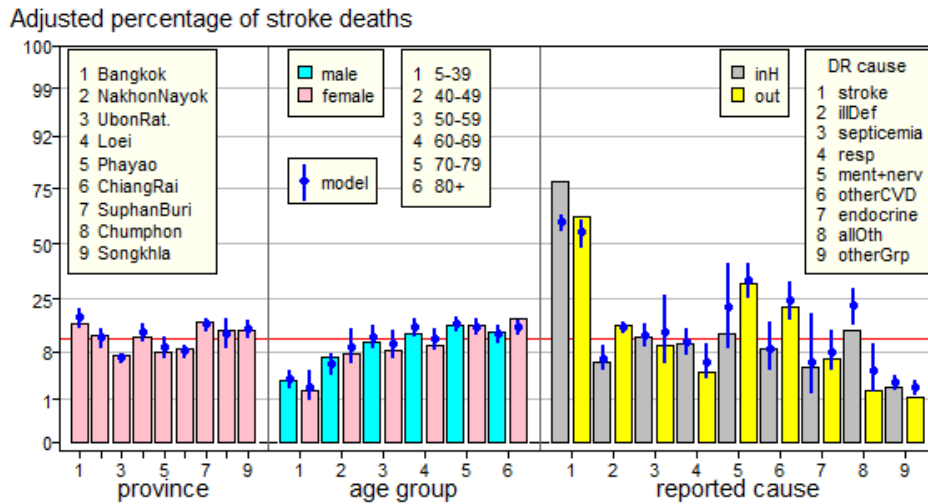


Figure 1 Adjusted percentage of stroke death by province, gender-age, and DR-cause location.

Figure 2 shows the ROC curve. A cut-off point is presented in the curve, where the predicted number of stroke deaths (1,072) is close to the observed value (1,076) in the VA data. The sensitivity is 41.6 %, and the specificity is 92.6 %. The area under the curve (AUC) is 0.67, indicating that the model performance is moderate predicting. The model estimated the proportion of other deaths (non-stroke) that were classified as stroke (false-positive) to be 7.4 %, and the proportion of stroke deaths that were misclassified as non-stroke deaths (false-negative) to be 58.4 %. The ROC curve for the simple model, with only DR-cause location group as a determinant, was also given. It shows that adding province and age-group into the model decreased error rate by 10 %.

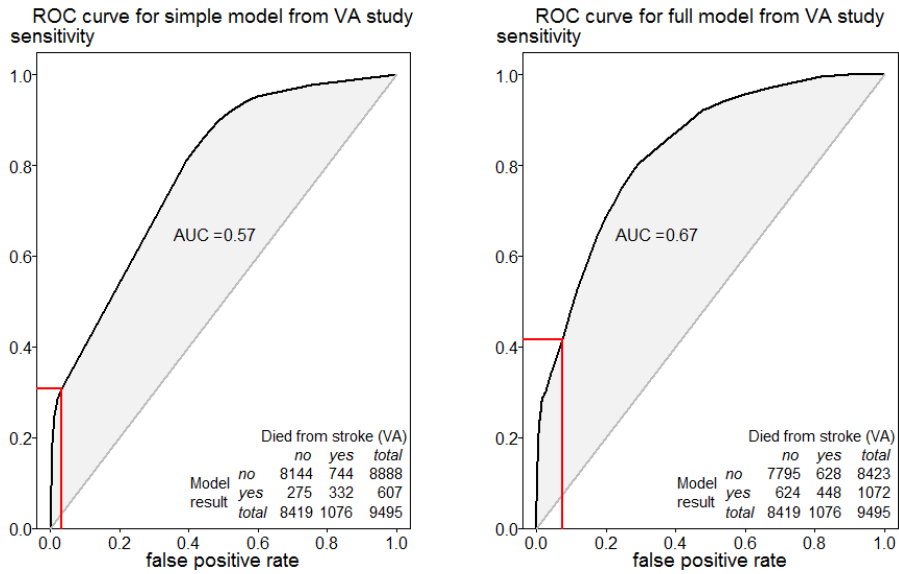


Figure 2 ROC curve of stroke model.

The result from the full model, developed from the sample data, was then applied to the DR data, to estimate the number of stroke deaths in 2005. **Figure 3** shows estimated percentages of stroke deaths. The percentage of VA-estimated stroke death was higher in the central region (Sing Buri, Chai Nat, and Samut Songkhram), and lower in the upper north (Chiang Mai, Chiang Rai, Mae Hong Son, and Phayao) and north east regions (Udon Thani, Sakon Nakhon, Nakhon Phanom, Kalasin, Mukdahan, Maha Sarakham, Roi Et, Yasothon, Amnat Charoen, Surin, Si Sa Ket, and Ubon Ratchathani).

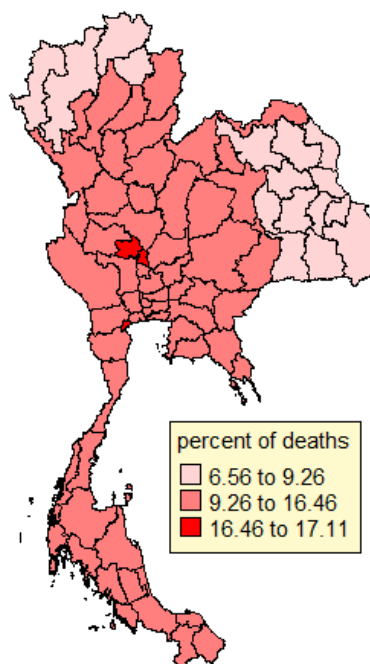


Figure 3 Percentages of VA-estimated strokes in 2005.

Table 2 shows DR-reported and VA-estimated stroke deaths. The model estimated stroke deaths as 24,110 (11.0 %) in males, and 21,913 (13.3 %) in females. The reported stroke deaths were 9,105 (4.2 %) in males, and 6,470 (3.9 %) in females. The inflation factors (IFs) showed that deaths in all gender-age groups were substantially under-reported, especially for the age group of 60 years and older.

Table 2 DR-reported and VA-estimated stroke deaths in 2005.

Gender age group	DR stroke	% DR stroke	VA-estimated stroke	% VA-estimated stroke	IF
<i>male</i>	9,105	4.2	24,110	11.0	2.7
m:5-39	1,145	2.5	1,388	3.0	1.2
m:40-49	1,579	5.5	1,878	6.5	1.2
m:50-59	1,964	6.3	3,707	11.8	1.9
m:60-69	1,886	5.1	5,158	13.9	2.7
m:70-79	1,701	3.9	7,107	16.5	4.2
m:80+	830	2.5	4,872	14.8	5.9
<i>female</i>	6,470	3.9	21,913	13.3	3.4
f:5-39	361	2.0	362	2.0	1.0
f:40-49	617	4.7	1,096	8.4	1.8
f:50-59	1,035	5.5	1,821	9.7	1.8
f:60-69	1,339	4.9	3,083	11.2	2.3
f:70-79	1,805	4.5	6,773	16.7	3.8
f:80+	1,313	2.8	8,778	18.6	6.7

Figure 4 shows area plots of DR-reported, simple model estimated, and full model estimated number of deaths from stroke by gender-age group in 1996 - 2009. The area of each color strip denotes the number of deaths in each age group.

The total number of deaths reported for 14 years were 157,537. The estimated total numbers of stroke deaths from the simple and full models were 549,653 and 570,245, respectively. The total number of DR-reported stroke deaths were lower than those estimated by the 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 the full model was used. For the older age groups, causes 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 the 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 the simple model and full model by factors of 3.78 and 4.10.

The area plots for stroke deaths clearly revealed that numbers of deaths were under-reported, especially in earlier years. Similar patterns were seen with number of deaths increasing in recent years.

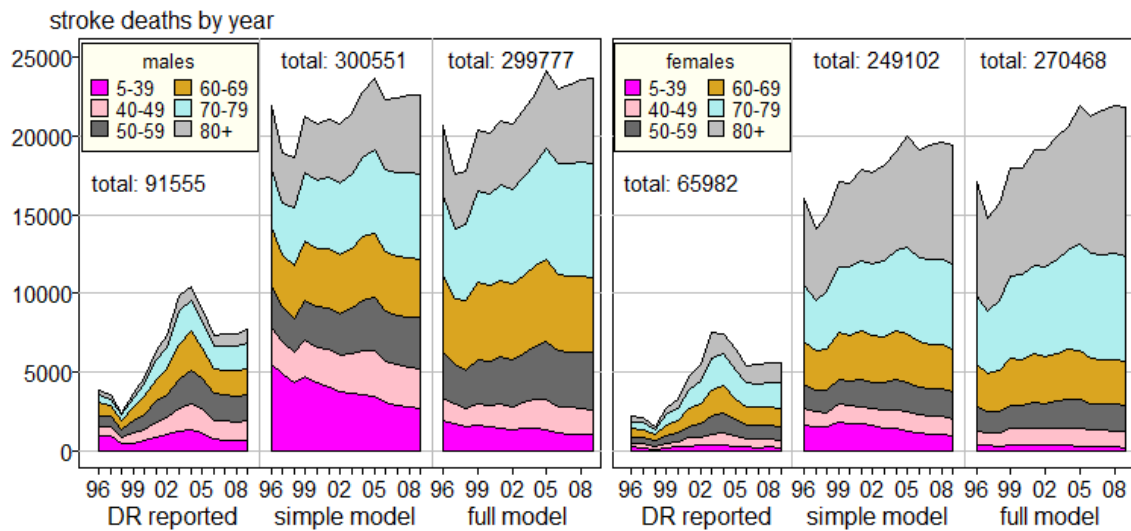


Figure 4 Area plot of DR-reported and VA estimated stroke deaths.

Figure 5 shows area plots of DR-reported and VA-estimated deaths by DR-cause location groups in 1996 – 2009, based on the simple and full model. The model estimated 382,784 for outside-hospital, and 166,869 for inside-hospital, stroke deaths. The reported deaths for outside-hospital were lower than the estimated by simple and full model by factors of 6.40 and 6.67, respectively. The reported deaths for inside-hospital were lower than the estimated by simple and full model by factors of 1.71 and 1.75, respectively.

Most stroke deaths outside hospital were misclassified as ill-defined (66 %), other CVD (9.4 %), and mental and nerves (6.6 %), respectively. In contrast, most of stroke deaths in hospital were misclassified as septicemia (12.8 %) and respiratory (10.1 %) causes, respectively.

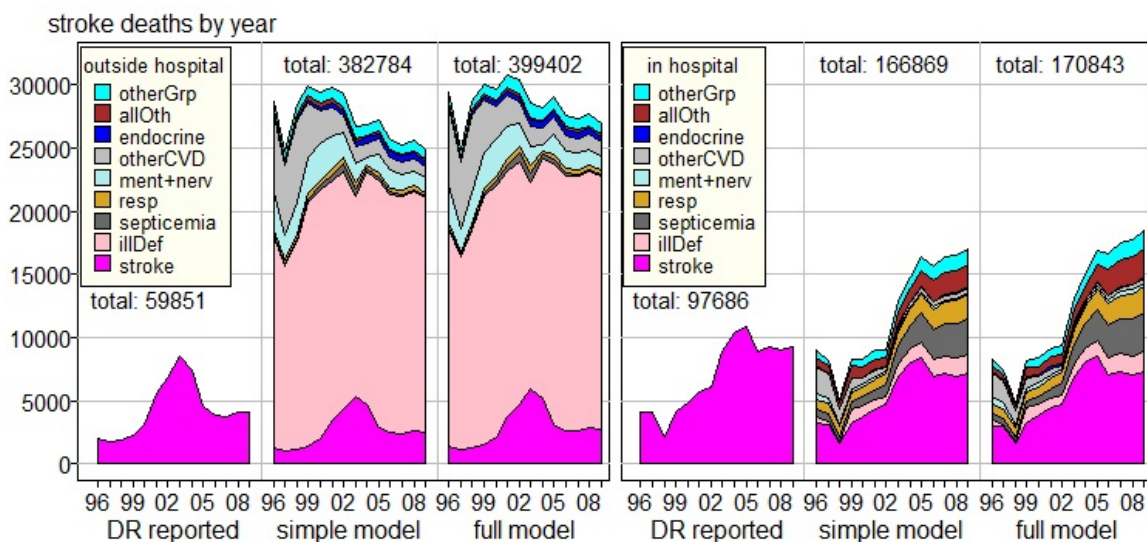


Figure 5 Area plot of DR-reported and VA-estimated stroke deaths by DR-cause location group.

Discussion

This study described the methods based on logistic regression for correcting misreported causes of deaths, and illustrated the methods using deaths from stroke. The data were from a 2005 VA survey. The logistic regression model of VA-assessed causes of deaths with demographic factors was found to be appropriate to use. It analyzed gender-age group, province, and DR-cause location, and predicted cause-specific deaths with higher sensitivity and specificity, compared with those derived from a simple model (the simple cross-referencing method) [8]. The methods can be simply used to estimate the number of deaths for another 20 causes. The models do not ensure that adjusted death counts, in each year, match reported totals, because they aggregate results from separate logistic regression models.

The multinomial model ensures that adjusted and reported totals match with the DR-reported deaths, and is also similar to totals from separate logistic models [13]. Fitting one model to an outcome with 21 levels gives complicated results. The simpler logistic regression model was chosen, because the interpretation and generalization of the results are more straightforward. The estimates from fitting binary logistic regression models separately for the pairings of responses differ from the estimates from the multinomial logistic regression model. They are less efficient, tending to have larger standard errors. However, it has been shown that the efficiency loss is minor when the response category having highest prevalence is used as the baseline [13,15].

The advantage of using the logistic regression model to analyze data is that it can simply handle general determinants. The logistic function has many desirable properties. Its range is between 0 and 1 when the independent variable varies from $-\infty$ 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 can assess interaction very effectively when there are several confounders or the confounder is a continuous variable [19]. 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 from a specific disease can also be calculated.

To reduce costs from conducting a VA study for the whole country, we proposed appropriate statistical methods to be applied to a large-scale VA study, for example, in the case of HIV [10], transport accidents [11], and liver cancer [12]. This method enables public health researchers to estimate percentages of specific causes of deaths in countries where there are death registration records of low quality but where reliable sample data, such as VA studies, should be available.

The importance of evaluating the reliability and validity of causes of death in mortality statistics has long been recognized in public health [28]. 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 [29]. VA surveys are generally the most reliable method to determine causes of death [7,30]. However, conducting a survey is expensive and time consuming. It is important for public authorities to pay attention to the quality of death registry, rather than conducting verbal autopsies.

The analysis of VA data with stroke deaths with adverse outcome showed that the full model reduced error rates. The full model has the ability to allocate misclassification of stroke deaths better than the simple model. The simple model ignored the effect of gender-age groups and province, which could give incorrect estimates due to confounding [13,19]. In order to observe the effects of demographic factors (gender-age groups and province) on stroke deaths clearly, we compared the simple and full models. An AUC from the full model indicated that the model had moderate predicting stroke deaths, and was greater than the AUC from the simple model [24]. This finding also reflects that gender-age group and province improved the prediction of stroke deaths [25,31].

The estimated numbers of deaths were higher than the DR-reported deaths. The percentage of deaths was slightly higher than those from a previous publication which used the same data set, but they used a cross-referencing method based on proportionate mortality distributions [8]. A model-based method can include gender-age group and province effects, and can predict stroke death.

Most of the stroke deaths in hospital were misclassified as septicemia and respiratory causes. This might be partly due to insufficient medical history in cases where death is immediate [9]. Some physicians have inadequate skills to define the cause of deaths [5,32]. In contrast, stroke deaths outside hospital were misclassified as ill-defined, other CVD, and mental and nerves causes. Ill-defined causes

were recorded mainly due to the absence of, or incomplete, medical opinion, and insufficient information of final hospital discharge record [33].

Stroke is a cause of death among the elderly. However, in our study sample, there were 5 cases with ages of less than 19, so we had an age group of 5 to 39 in our analysis. Although this age group is wide, it was retained in the model as a basis for comparing stroke deaths with other causes of deaths. It is not surprising that a high percent of stroke death were observed among people aged 60 and older. This agrees with a previous study. A high prevalence of stroke deaths was found among adults aged 75 years and older [17]. Another study also supported that the average age of stroke onset was 65 years [34]. These might be reflective of poorly controlled risk factors of stroke, such as total cholesterol, blood pressure, smoking, and obesity among the adult population [8,34,35].

Geographical distribution of stroke deaths was higher in the central region, and lower in the north-eastern region. The geographical variation on stroke deaths in 2000 has been reported by using Standardized Mortality Ratio (SMR) to be higher in Bangkok and lowest in upper north-eastern region [36]. Yet, in other studies, the prevalence of stroke deaths in the central region was the highest, and the lowest in the north-eastern region [17,37]. These might be related to a prevalence of risk factors of stroke, such as hypertension, diabetes, and smoking habits [38]. A high prevalence of hypertension and diabetes was reported in Bangkok and the central region [35].

Our analysis has a few limitations. The survey design has not considered a strength predictor, such as the registered cause during stratification. Thus, the study sample may not adequately cover all of the population at risk, for example, the Muslim majority districts.

Conclusions

The methods enable health professionals to estimate any specific causes of deaths in countries where causes of death are of low quality and where reliable cause of death from other sources are available. In addition, there is still a substantial misclassification of stroke mortality, according to our model.

Acknowledgements

This study was funded by the Program Strategic Scholarships Fellowships Frontier Research Networks (Specific for Southern Region) for the Ph.D. Program, Thai Doctoral degree, from the Office of the Higher Education Commission, Thailand, and also by the 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 with the data.

References

- [1] World Health Organization. *World Health Statistics 2012*. Geneva, Switzerland, 2012.
- [2] CD Mathers, DM Fat, M Inoue, C Rao and AD Lopez. 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**, 171-7.
- [3] V Tangcharoensathien, P Faramnuayphol, W Teukul, K Bundhamcharoen and S Wibulpholprasert. A critical assessment of mortality statistics in Thailand: Potential for improvements. *Bull. World Health Organ.* 2006; **84**, 233-9.
- [4] P Vapattanawong and P Prasartkul. Under-registration of deaths in Thailand in 2005-2006: Results of cross-matching data from two sources. *Bull. World Health Organ.* 2011; **89**, 806-12.
- [5] DL Brown, F Al-Senani, LD Lisabeth, MA Farnie, LA Colletti, KM Langa, AM Fendrik, NM Garcia, MA Smith and LM Morgenrtern. Defining cause of death in stroke patients: The brain attack surveillance in Corpus Christi project. *Am. J. Epidemiol.* 2007; **165**, 591-6.

- [6] J Pattaraarchachai, C Rao, W Polprasert, Y Porapakkham, W Poa-in, S Nophcha and AD Lopez. Cause-specific mortality patterns among hospital deaths in Thailand: Validating routine death certification. *Popul. Health Metr.* 2010; **8**, 12.
- [7] C Rao, Y Porapakkham, J Pattaraarchachai, W Polprasert, N Swampunyaalert and AD Lopez. Verifying causes of death in Thailand: Rationale and methods for empirical investigation. *Popul. Health Metr.* 2010; **8**, 11.
- [8] Y Porapakkham, C Rao, J Pattaraarchachai, W Polprasert, T Vos, T Adair and AD Lopez. Estimated causes of death in Thailand, 2005: Implications for health policy. *Popul. Health Metr.* 2010; **8**, 14.
- [9] W Polprasert, C Rao, T Adair, J Pattaraarchachai, Y Porapakkham and AD Lopez. Cause-of-death ascertainment for deaths that occur outside hospitals in Thailand: Application of verbal autopsy methods. *Popul. Health Metr.* 2010; **8**, 13.
- [10] A Chutinantakul, P Tongkumchum, K Bundhamcharoe and V Chongsuvivatwong. Correcting and estimating HIV mortality in Thailand based on 2005 verbal autopsy data focusing on demographic factors, 1996-2009. *Popul. Health Metr.* 2014; **12**, 25.
- [11] N Kinjun, A Lim and K Bundhamcharoen. A logistic regression model for estimating transport accident deaths using verbal autopsy data. *Asia Pac. J. Public Health* 2015; **27**, 286-92.
- [12] S Waeto, N Pipatjaturon, P Tongkumchum, C Choonpradub, R Saelim and N Makaje. Estimating liver cancer deaths in Thailand based on verbal autopsy study. *J. Res. Health Sci.* 2014; **14**, 18-22.
- [13] A Agresti. *Categorical Data Analysis*. 2nd ed. John Wiley & Sons, New Jersey, 2002.
- [14] B Huitema. *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-experiments, and Single-case Studies*. 2nd ed. John Wiley & Sons, New Jersey, 2011.
- [15] CB Begg and R Gray. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika* 1984, **71**, 11-8.
- [16] C Jalayondeja, J kaewkungwal, PE Sullivan, S Nidhinandana and S Pichaiyongwongdee. Factors related to community participation by stroke victims six month post-stroke. *Southeast Asian J. Trop. Med. Public Health* 2011; **42**, 1005-13.
- [17] S Hanchaiphiboolkul, N Pongvarin, S Nidhinandana, NC Suwanwela, P Puthkhao, S Towanabut, J Suwantamee and M Samsen. Prevalence of stroke and stroke risk factors in Thailand: Thai Epidemiologic Stroke (TES) study. *J. Med. Assoc. Thai.* 2011; **94**, 427-36.
- [18] World Health Organization. *ICD-10 International Statistical Classification of Diseases and Related Health Problems*. Geneva, Switzerland, 2004.
- [19] DW Hosmer and S Lemeshow. *Applied Logistic Regression*. 2nd ed. John Wiley & Sons, New Jersey, 2000.
- [20] WN Venables and BD Ripley. *Modern Applied Statistics with S*. 4th ed. Springer-Verlag, New York, 2002.
- [21] P Tongkumchum and D McNeil. Confidence interval using contrasts for regression model. *Songklanakarin J. Sci. Tech.* 2009; **31**, 151-6.
- [22] N Kongchouy and U Sampantarak. Confidence intervals for adjusted proportions using logistic regression. *Mod. Appl. Sci.* 2010; **4**, 2-7.
- [23] U Sampantarak, N Kongchouy and M Kuning. Democratic confidence intervals for adjusted means and incidence rates. *Am. Int. J. Contemp. Res.* 2011; **1**, 38-43.
- [24] SK Sarkar and H Midi. Importance of assessing the model adequacy of binary logistic regression. *J. Appl. Sci.* 2010; **10**, 479-86.
- [25] J Fan, S Upadhye and A Worster. Understanding receiver operating characteristic (ROC) curves. *Can. J. Emerg. Med.* 2006; **8**, 19-20.
- [26] GF Bonham-Carter. *Geographic Information Systems for Geoscientists: Modelling with GIS*. Pergamon, Oxford, 1994.
- [27] R Core Team. A Language and Environment for Statistical Computing, Available from: <http://www.R-project.org>, accessed August 2014.
- [28] IM Moriyama. Problems in measurement of accuracy of cause-of-death statistics. *Am. J. Public Health* 1989; **79**, 1349-50.

- [29] A Khosravi, C Rao, M Naghavi, R Taylor and N Jafari. Impact of misclassification on measure of cardiovascular disease mortality in the Islamic Republic of Iran: a cross-sectional study. *Bull. World Health Organ.* 2008; **86**, 688-96.
- [30] RA Lahti and A Penttilä. The validity of death certificates: Routine validation of death certification and its effects on mortality statistics. *Forensic Sci. Int.* 2001; **115**, 15-32.
- [31] G Vanagas. Receiver operating characteristic curves and comparison of cardiac surgery risk stratification systems. *Interact. Cardiovasc. Thorac. Surg.* 2004; **3**, 319-22.
- [32] DR Lakkireddy, MS Gowda, CW Murray, KR Basarakodu and JL Vacek. Death certificate completion: How well are physicians trained and are cardiovascular causes overstated? *Am. J. Med.* 2004; **117**, 492-8.
- [33] LA Johansson and R Westerling. Comparing hospital discharge records with death certificates: Can the differences be explained? *J. Epidemiol. Comm. Health* 2002; **56**, 301-8.
- [34] NC Suwanwela. Stroke epidemiology in Thailand. *J. Stroke* 2014; **16**, 1-7.
- [35] V Chongsuvivatwong, T Yipintsoi, P Suriyawongpaisal, S Cheepudomwit, W Aekplakorn, P Faramnuayphol, P Tatsanavivat, V Kosulwat, S Thamthitiwat and C Nuntawan. Comparison of cardiovascular risk factors in five regions of Thailand: InterASIA data. *J. Med. Assoc. Thai.* 2010; **93**, 17-26.
- [36] P Faramnuayphol, V Chongsuvivatwong and S Panarunothai. Geographical variation of mortality in Thailand. *J. Med. Assoc. Thai.* 2008; **91**, 1455-60.
- [37] K Kongbunkiat, N Kaemsap, K Thepsuthammarat, S Tiamkao and K Sawanyawisuth. National data on stroke outcomes in Thailand. *J. Clin. Neurosci.* 2015; **22**, 493-497.
- [38] DG Hoy, C Rao, NP Hoa, S Suhardi and AM Lwin. Stroke mortality variations in South-East Asia: Empirical evidence from the field. *Int. J. Stroke* 2013; **8**, 21-7.