Analyzing a Statistical Method of Estimating Respiratory Deaths based on the Thailand Verbal Autopsy study

Pradthana Minsan

Department of Mathematics and Statistics, Faculty of Science and Technology, Chiang Mai Rajabhat University, Muang, Chiang Mai, Thailand, 50300

Corresponding author. E-mail address: pradthana.m@gmail.com

Received: 1 November 2016; Accepted: 7 June 2017

Abstract

The archives record of the causes of death in Thailand using the death registration (DR) system is important source of mortality data. Over the past two decades, Thailand has presented many formats to enhance civil registration and vial statistics systems. More than 30% of death unregistered and about 40% of deaths registered have given the cause of death as "ill-defined". The aim of this study was to propose a statistical model to estimate percentages of respiratory deaths in Thailand based on a sample of 9,644 deaths from the 2005 Verbal Autopsy (VA) study. Logistic regression was used to predict respiratory deaths classified by three factors, province, gender-age group and cause of each death. The receiver operating characteristic (ROC) curve was used to assess accuracy of the model prediction and the area under the ROC curve measures discrimination model which has ability to correctly predict those with and without respiratory deaths. Province, gender-age group and cause of death were statistically significant associated with respiratory deaths. The area under ROC curve was 0.7 with a false positive rate of 5.52% and sensitivity of 39.2%. Moreover, the results revealed that the under-reporting of respiratory deaths were those registered as tuberculosis, septicemia and other CVD cases. In conclusions, the logistic model in this study can be used for estimating the respiratory deaths from the DR database in Thailand or the DR system in other countries that are under-reporting the death rate.

Keywords: Respiratory Mortality, Verbal Autopsy study, Logistic Regression, Confidence Intervals, ROC curve

Introduction

Respiratory disease is caused by a malfunction of the tissues or organs in the respiratory tract, such as the nasal passages, the bronchial arteries and the lungs. In 2004, the cause of death from respiratory disease, approximately 2.8 million or one quarter of all deaths (estimated 12 million) in the Asia-Pacific region (defined by the WHO as the "Western Pacific" plus Thailand) (Jamrozik & Musk, 2011). Statistics on the Causes of Deaths were classified by Basic Tabular List of "International Statistical Classification of Diseases and Health Related Problems the Tenth Revision" (ICD-10) in Thailand that was published in Public Health statistics A.D.2014, 29,654 persons died of diseases of the respiratory system or the death rate was 46.6 per 100,000 population in 2010. However, the death rate increased to 39,638 persons or 61.0 per 100,000 population in 2014 (Bureau of Policy and Strategy, 2014). Moreover, morbidity rates of out-patients which were classified by disease groups in Thailand (Bangkok not included), respiratory disease was the highest rate in 2013 in other words that is 418.25 per 1,000 population (Strategy and Planning Division, 2013). Although over the past two decades, Thailand has presented many formats to enhance civil registration and vial statistics systems, but the efficiencies of mortality statistics continued to have a low capability (Tangcharoensathien, Faramnuayphol, Teokul, Bundhamcharoen, & Wibulpholprasert, 2006). However, the Death Registration (DR) in Thailand used have stated that more than 30% of deaths unregistered and about 40% of deaths registered have given the cause of death as "ill-defined" (Mathers, Ma Fat, Inoue, Rao, & Lopez, 2005). In contrast, Japan for example a country in Asia, which is

a high-income country, has less than 4% of deaths ill-defined (Rao et al., 2010). According to these DR figures, 65% of deaths in Thailand occur outside hospitals. The causes of these deaths are informed by civil registers that are based on a testimony from a cousin, frequently recorded by physical opinion that consider sickness that could have contributed to death (Tangcharoensathien et al., 2006; Polprasert et al., 2010). Hence, the Ministry of Public Health (MoPH) proposes two procedures to address this problem (Rao et al., 2010). Firstly, the deaths that occur in hospitals that have been verified by the physician are reported according to ICD-10 code. Secondly, for the deaths that have occurred outside the hospitals, local sanitarians should be trained for verifying more precisely the causes of death by using accessible medical records and verbal autopsy (VA) methods. After that, all reports of the causes of death registration could be transferred to the MoPH database in ICD-10 code pattern and tabulated by age, sex and cause for promulgation and analysis for improved results. Hence, in 2005, the MoPH offered a VA study that grouped Thai public health (doctors, medical staffs, biostatisticians and epidemiologists) together to evaluate the DR database seriously to obtain more accurate registered causes of death in Thailand (Waeto et al., 2014). Although the VA can be precise in estimating the true underlying cause of death, it can still be indecisive, so the probabilistic modeling addition should be in the research as suggested (Byass, 2010). As, many have studied, trying to use the 2005 VA data applying suitable statistical method for estimating the true cause of death. For instance, from the DR database in Thailand 2005, a capture-recapture method and matching deaths was used to estimate the total deaths by age and sex, and used proportional mortality rates distribution by age, sex and cause of the death for adjusting data (Porapakkham et al., 2010). This study focuses on respiratory mortality which the ICD-10 code are J00-J99 (World Health Organization, 2004) and the aims of this study were to: 1. Propose a statistical method to estimate percentages of respiratory deaths in Thailand base on the 2005 VA study. 2. Apply adjusted percentages confident intervals using weighted some contrasts to compare population proportions.

Methods and Materials

The data from a VA study that was used this research was selected from a sample of 9,644 cases of deaths from 5 years and older, from the nationally representative study based on a database for 2005 which using a stratified cluster sampling approach. The sampling units that were drawn from the DR database of Thai citizens, who were permanent resident in Thailand in 28 districts, 9 provinces (2 provinces in each 4 regions of Thailand including Bangkok) were 9,495 cases of deaths. For each case, the data collection, which were from 5 fields: a) the 9 provinces were used in collecting data as shown in Figure 1.; b) gender and age of death; c) the cause of death, which classified according ICD-10 code; d) the place of death (inside or outside the hospital); e) the VA method which assessed ICD-10 code.

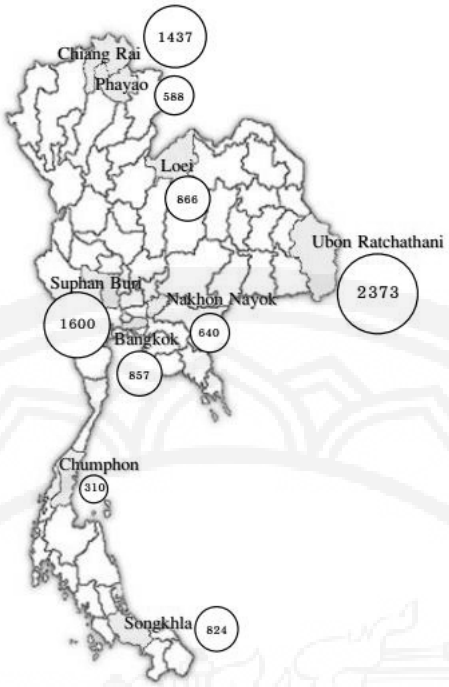

Figure 1 The geographical distribution of sample size in Thailand.

From the chapter-block classification of ICD-10 codes (Porapakkham et al., 2010), the 21 major cause groups for deaths at age 5 or more were based on the distribution of these reported with VA counts. Figure 2 summarizes the association between VA and DR cause groups. Accordingly, the sample of respiratory deaths that occurred in Thailand during 2005 was 603 cases. Meanwhile, the deceased from respiratory disease from the VA report were 801 cases. However, only 256 cases from the sample that VA matched DR were reported.

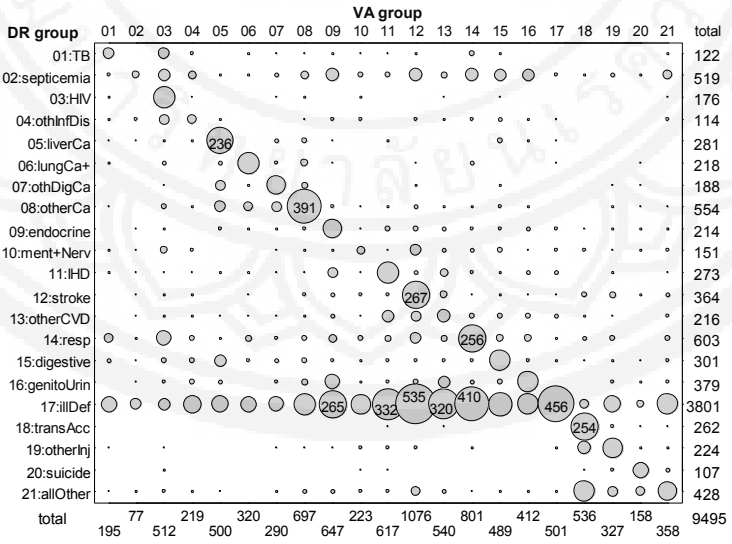

Figure 2 Association between VA and DR cause groups.

The associations between the deceased and the location of death, age group and reported cause from preliminary analysis are shown in Figure 3. Payao had the highest percentage of respiratory deaths followed by Chiang Rai and Nakhon Nayok. Percent of respiratory death vary by age groups for both males and females. Males had higher respiratory death than females for all age groups except aged 5-29 years. In terms of reported cause, more than 40% of reported respiratory deaths were really due to respiratory death in and outside hospital. However, the deaths were reported as ill-defined also accounted for around 10% of respiratory deaths. Additionally, septicemia (septic), other Cardiovascular disease death (other CVD) and Tuberculosis (TB) reported deaths accounted for the bulk of the remainder.

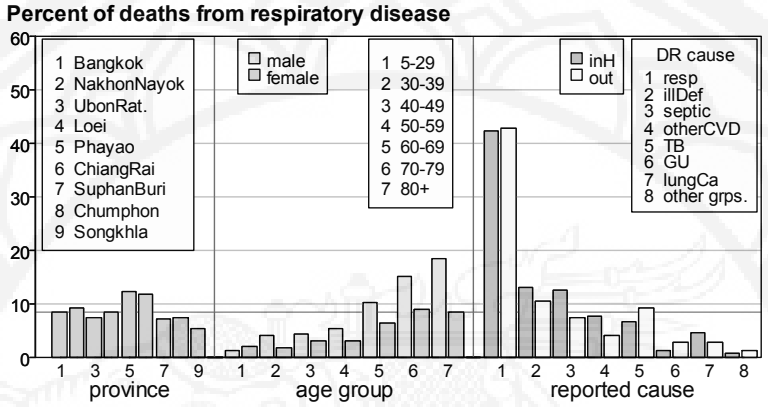

Figure 3 Percent of deaths from respiratory disease

Statistical Analysis

Since the response variable of these data was a dichotomous outcome (that is died from respiratory by ICD-10 code or not). And the objective of this study was to find the best fitting model to estimate percentages of respiratory deaths in Thailand base on the 2005 VA study. Then the appropriate model to fit this prediction is a logistic regression model. Let $\pi(\mathbf{x})$ is the probability that a person died from respiratory as an additive linear function of the three covariate factors denoted by the vector $\mathbf{x}' = (x_1, x_2, x_3)$. The logit of a multiple logistic regression model is given by the equation:

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \mu + \alpha_i + \beta_j + \gamma_k . \qquad (1)$$

In Eq(1), μ is a constant and the term α_i, β_j and γ_k are individual parameters that specify province, gender-age group and DR cause-location factor, respectively. Province factor has 9 levels, according to the nine provinces in the VA sample. The gender-age group factor has 14 levels, by dividing age into seven groups for males and females. The DR cause-location factor has 16 levels, according to DR cause group and location of death (in or outside hospital).

Hypothesis testing to determine whether the factors in model significantly, $H_0 : \alpha_i = \beta_j = \gamma_k = 0$ is Wald statistic or $P(\chi^2 > D)$ Chi-square distribution with df = m-1 where m is the number of levels and D is the reduction deviance (a measure of goodness of fit of the

logit model. For estimation of unknown parameter in Eq(1) is done with the Fisher Scoring method that is used to derive the maximum likelihood estimation (MLE).

In this study, the receiver operating characteristic (ROC) curve is used to assess the goodness of fit of the model prediction (Hosmer, Lemeshow, & Sturdivant, 2013), it plotted of the probability of sensitivity and specificity for an entire range of possible cut-points (as c varies). The area under the ROC curve provides a measure of the capacity of the model to predict a binary outcome. If $\pi(\mathbf{x}) \geq c$ the predicted outcome as 1 (death from respiratory), or 0 (other death) if $\pi(\mathbf{x}) < c$. In this study, the cut-off point, c, was chosen in order to predict respiratory deaths consistent with the respiratory deaths in the VA study, which were 801 cases (Figure 2). The 95% confident interval (CI) based on sum contrasts that are described by Tongkumchum and McNeil (2009) and Kongchouy and Sampantarak (2010) were used to estimate the 95% CI of an adjusted percentage for each level of the factor in VA and the DR database. Since this CI can be reduced confounding bias that cause of association of the factors with the binary outcome and the levels of the factors (McNeil, 1996).

Results

From the likelihood ratio test, the logistic regression model with three factors was created and gave the deviance reduction as shown in Table 1. All of the three factors had p-value less than 0.001 indicated that the three factors were statistically significant associated with respiratory deaths and the model yielded a χ^2 of 920.09 with 1,212 degrees of freedom.

Table 1 Likelihood Ratio Test for Logistic Regression Model of Respiratory Deaths by DR cause-location, Gender-age group and Province.

Factors	Deviance reduction	df	p-value
Province	965.91	8	0.0000002
Gender-age group	1103.92	13	< 0.001
DR cause-location	1785.97	15	< 0.001
Error	920.09	1212	

Figure 4 shows the ROC curve that was constructed from a logistic regression model. In this study, choosing a single cut-off point, c = 0.208, gave 792 predicted respiratory deaths with the specificity of 0.945 and the false positive of 0.055. The area under the ROC curve (AUC) of 0.7 showed the best precision model with the sensitivity of 0.390.

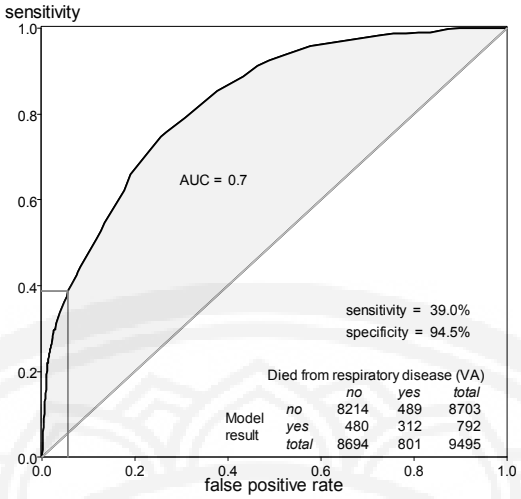

Figure 4 Receiver Operating Characteristic (ROC) curve and classifying observed and estimated respiratory disease

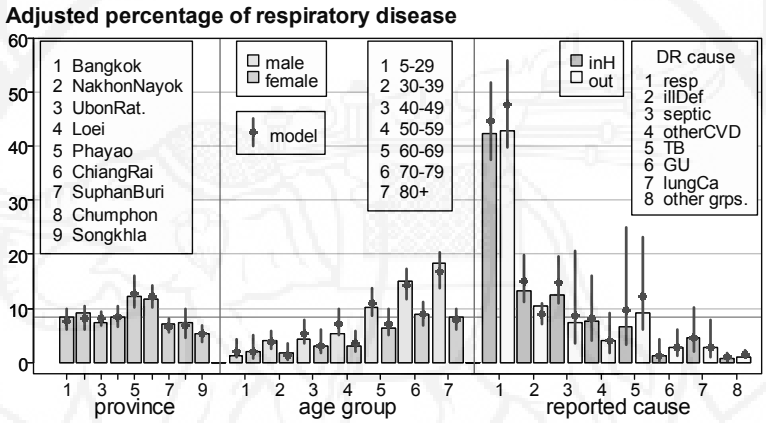

Figure 5 Bar chart of the percentage of respiratory death and 95% CI of adjusted percentages of the VA assessed estimated from respiratory death by province, age group and reported cause.

Figure 5 shows a bar chart of the percentage of respiratory deaths compared to the model – based on a 95% confidence interval of adjusted percentage of the VA it was assessed and estimated from respiratory death by province, age group and reported cause of death. From the graph, the red line denotes as overall mean and the blue vertical line segments is shown as 95% CI of the adjusted percentage of VA assessed estimated respiratory death for each of the three cause groups. For the province, the percentage of respiratory death in Phayao and Chiang Rai province in the north were higher than the average percentage. Meanwhile, the cause of gender- age groups, male aged between 60-69, 70-79 and 80+ were higher than the average percentage. For the DR cause-location, respiratory deaths in hospital corresponded to the DR-reported more than 40% same as deaths outside hospital. As more than 10% were reported as ill-defined and septicemia for both deaths in and outside hospital which corresponded to the reported DR.

Discussion

The accuracy of death registration of the DR system provides important information for public health policies. The 2005 VA study found that ill-defined cause of death was 40.03%, after applied a simple cross-referencing method on DR and VA data and reclassification in the VA study. The ill-defined causes of death decreased was at 5.28%. The description of

quality controlling measures for the 2005 VA in Thailand was explained by Rao et al. (2010). Out of the 9,495 deaths in the VA data showed that there were 8.47% of respiratory deaths due to the 2.70% that DR system matched by the VA data and the accuracy of the reported DR was 42.45%. Therefore, and the under-reporting of respiratory death in the DR was estimated by 1.98%.

In this study, we applied logistic regression model for the 2005 VA database to estimate the correct number of respiratory deaths. The model with three determinants, 9 provinces, 14 gender-age groups and 16 DR cause-location groups are created. The 95% confidence intervals plot for adjusted percentages of VA-assessed respiratory deaths was constructed. The mode showed that the three determinants are highly statistically significant associated with respiratory deaths. The respiratory deaths probably occurred in males more than females in all of the age groups according to Public Health Statistics 2007 (Bureau of Policy and Strategy, 2007). Meanwhile, the model showed that Payao Province in the North of Thailand had the highest respiratory death in agreement with respiratory morbidity rate in hospital Thailand morbidity report in 2005 (Strategy and Planning Division, 2005). The number of respiratory deaths in the DR system was underreported for all causes of deaths both in and outside hospital. Thus, applying the logistic regression model is accurate for estimating the respiratory deaths in the DR system in Thailand.

Conclusion and Suggestion

According to the model, we can conclude that, the number of respiratory deaths in the DR database is biased by under-reporting especially in the cause of specific groups in and outside hospital. Although, many methods have been used for estimating the correct causes of deaths like the logistic regression used in this study that is well-known in public health research. However, this methodology has its limitations. Since the sampling plan of the verbal autopsy is a stratified cluster sampling approach which the province effect had fixed, then the standard error of this approach has larger number than a simple random sampling (Lumley, 2010).

Acknowledgement

The author acknowledges the support received from Ms Amornrat Chutinantakul and Mr Nattakit Pipatjaturon toward data management and is grateful to Professor Don McNeil for his helpful advice and suggestion. Also, thanks to the Thai Ministry of Public Health for providing mortality data.

References

Bureau of Policy and Strategy, Ministry of Public Health (Thailand). (2007). Public Health statistics A.D.2007. Retrieved from http://bps.moph.go.th/new_bps/sites/default/files/statistic-50.pdf

Bureau of Policy and Strategy, Ministry of Public Health (Thailand). (2014). Public Health statistics A.D. 2014. Retrieved from http://bps.moph.go.th/new_bps/sites/default/files/health_statistics2557.pdf

Byass, P. (2010). Integrated multisource estimates of mortality for Thailand in 2005. *Population health metrics*, *8*(1), 10. https://doi.org/10.1186/1478-7954-8-10

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (Eds) (2013). *Applied Logistic Regression*. New York: Wiley.

Jamrozik, E., & Musk, A. W. (2011). Respiratory health issues in the Asia-Pacific region: An

overview. *Respirology*, *16*(1), 3-12. http://dx.doi.org/10.1111/j.1440-1843.2010.01844.x

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

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

Mathers, C. D., Ma Fat, D., Inoue, M., Rao, C., & 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), 171-177c.

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

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

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

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

Strategy and Planning Division, Ministry of Public Health (Thailand). (2005). Morbidity Report A.D. 2005. Retrieved from http://bps.moph.go.th/new_bps/%E0%B8%AA%E0%B8%A3%E0%B8%B8%E0%B8%9B%E0%B8%A3%E0%B8%B2%E0%B8%A2%E0%B8%87%E0%B8%B2%E0%B8%99%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B8%9B%E0%B9%88%E0%B8%A7%E0%B8%A2

Strategy and Planning Division, Ministry of Public Health (Thailand). (2013). Morbidity Report A.D. 2013. Retrieved from http://bps.moph.go.th/new_bps/%E0%B8%AA%E0%B8%A3%E0%B8%B8%E0%B8%9B%E0%B8%A3%E0%B8%B2%E0%B8%A2%E0%B8%87%E0%B8%B2%E0%B8%99%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B8%9B%E0%B9%88%E0%B8%A7%E0%B8%A2

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), 233-238.

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

Waeto, S., Pipatjaturon, N., Tongkumchum, P., Choonpradub, C., Saelim, R., & Makaje, N. (2013). 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). *International Statistical Classification of Diseases and Health Related Problems Tenth Revision Volume 1* (2nd ed., pp. 977-1066). World Health Organization, Geneva, Switzerland.