Abstract

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

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

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

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

Introduction

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

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

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

There are two kinds of liver cancer. Hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA). The ICD-10...
code for HCC is C22.0 and for CCA is C22.1. HCC and CCA have different etiology but Thai death certificates code both as C22.9 (unspecified liver cancer).

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

Methods

Data sources

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

The VA study was carried out in 2005 based on a sample of 3,316 in-hospital and 6,328 outside-hospital deaths from 28 selected districts in nine provinces, giving a data table with 5 fields: (a) the deceased person’s province; (b) the person’s gender and age; (c) the ICD-10 code reported on the death certificate; (d) the location of death (in hospital or outside hospital); (e) the VA-assessed ICD-10 code.

The VA data were separated by field (d), grouped fields (c) and (e) into the 21 leading causes of death for each location plus all other cause group, and thus found inflation factors for determining percentages of deaths in specific cause groups. The 22 groups were classified according to the ICD-10 Mortality Tabulation categories, giving each group had to be large enough for statistical analysis. The cause group based on the VA count ranged from 77 for septicemia (A40-41) to 1,076 for stroke (I60-69). There were 500 deaths for liver cancer (C22).

Statistical Methods

The outcome was liver cancer death (yes/no) and the determinants were province, gender-age group and VR cause location. The logistic regression model was used for describing the relation between the outcome and determinants. This model formulated the logit of the probability $p$ that a person died from liver cancer as an additive linear function of the three determinant factors as follows:

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

(1)

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

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

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

Sum contrasts

Sum contrasts was used to obtain confidence intervals for comparing means/proportions with the overall mean/proportion. An advantage of these confidence intervals is that they provide a simple criterion for classifying levels of the factor into three groups according to whether each corresponding confidence interval exceeds, crosses, or is below the overall mean. The confidence intervals based on sum contrasts are used because they are more appropriate compared to the corresponding confidence intervals based on the treatment contrasts. The confidence intervals compare percentage of liver cancer deaths in each category factor with the overall percentage. They applied equitably to each category, whereas the commonly used confidence intervals based on treatment contrasts measured the difference from a reference group that is taken to be fixed and thus does not have a confidence interval.

Triangulation Method

To predict results for provinces outside the VA study, we estimated provinces’ coefficients based on latitude and longitude of their central points. Triangles were drawn linking the nine VA provinces. These triangles were set at planes, like roofs on poles with heights corresponding to their model coefficients value at the vertices of the triangles. Coefficients for provinces inside triangles were obtained by solving three linear equations via linear algebra.

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

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

Results

Preliminary Results

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

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

Logistic Regression Model

The P-value for a factor in the regression model is the probability of \( P > 0.02 \), the tail area of a chi-squared distribution with \( k-1 \) degrees of freedom, where \( k \) is the number of levels and \( D \) is the reduction in deviance (a measure of lack of fit of the model) achieved by the factor. The three factors in the logistic regression model are highly statistically significant (\( P < 0.001 \)).

Figure 2 shows the ROC curve of logistic regression model. Choosing \( c = 0.216 \) gives 500 predicted liver cancer deaths, in agreement with the VA study, for which the sensitivity is 0.64 and the false positive rate is 0.02. Note that just using the reported cause to predict the true cause has sensitivity 0.47. Only 236 cases out of 500 liver cancer deaths were correctly reported.

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

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

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

![Figure 1: Percentage of liver cancer deaths by province, gender-age group and VR cause-location](image1)

![Figure 2: Receiver Operating Characteristic (ROC) curve and cross-classifying observed and estimated outcome](image2)
The estimated probabilities of liver cancer deaths from the model were applied to the VR data for males and females by age groups from 2000 to 2009. Over the decade 2000-2009, the estimated numbers of liver cancer deaths were 134,244 (males) and 58,548 (females). These are 56% and 64% higher than the reported totals of 85,873 and 35,643, respectively. Figure 4 compares numbers of liver cancer deaths between VA estimated and VR reported deaths using area plot.

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

Discussion

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

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

It is well known that liver cancer mortality varies with gender and age. It more pronounces among males and elderly. The results in this study were of high percentages in males. For age the percentages of liver cancer deaths were higher than average in ages 40-49, 50-59 and 60-69 for males and marginally higher in age 60-69 for females.

Therefore, estimating number of liver cancer deaths is necessary to take these demographic factors into account.

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

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

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

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

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

Conclusions

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

Acknowledgments

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

**Conflict of interest statement**

The authors have no conflict of interests to declare.

**Funding**

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

**References**


