Songklanakarin J. Sci. Technol.
36 (2), 249-254, Mar. - Apr. 2014

SJST SONGKLANAKARIN JOURNAL OF SCIENCE AND TECHNOLOGY
http://www.sjst.psu.ac.th

Original Article

Estimation of mortality with missing data using logistic regression

Amornrat Chutinantakul[1,2], Marzukee Mayeng[2], and Phattrawan Tongkumchum[2]*

[1] *Office of Disease Prevention and Control, Region 11 Nakhon Si Thammarat, Mueang, Nakhon Si Thammarat, 80000 Thailand.*

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

Received 16 June 2013; Accepted 22 January 2014

Abstract

The present study aims to improve estimation of mortality data with unknown demographic factors using logistic regression, based on inflation factors for distributing such cases. The method is illustrated on death proportions based on numbers of deaths reported in 1996-2009 classified by gender, age and province in the death registration (DR) system. The results indicate that cases with unknown province mostly occurred in ages 0-4 years and 15-44 years and cases with unknown age mostly occurred in the central and southern regions. The method is straightforward, provides confidence intervals, and can be generally used for eliminating biases due to cases with unknown values of demographic factors in DR data. The resulting estimated numbers of deaths were used to examine age-specific mortality using cubic spline function. Cubic spline interpolation is effective to smoothly interpolate non-negative mortality data. These methods provide valuable information to the Ministry of Public Health.

Keywords: estimation, mortality, unknown demographic factors, logistic regression, cubic spline

1. Introduction

In vital registration surveys it is quite common for reports to have missing or unknown values for important variables. Yearly death certificates in the death registration (DR) database for the years 1996-2009 from the Ministry of Interior in Thailand had between 0.01% and 6.82% of records with unknown province and between 0.05% and 4.78% with unknown age. Although these percentages are relatively small and have decreased substantially since 2005 to 2009, it is important to have appropriate methods to adjust for biases that arise in such cases. Note that if cases with missing values are omitted from analysis under the possibly valid assumption that they are representative of the population, consequent estimates of incidence rates will be biased downwards unless appropriate inflation factors are applied.

Gender and age-specific mortality rates are obtained by dividing numbers of deaths for each sex and age group by corresponding populations (usually 1000s or 100,000s). Death rates in two regions with different age patterns are then compared by applying these rates to the same population. Based on such understanding of how Thai population distributions vary geographically throughout Thailand, we consider variation in numbers of deaths (the numerator in the mortality rate formula). Before examining these patterns, it is important to focus on data quality. As pointed out by Carmichael (2011) "significant under-registration and age-selective under-enumeration" occur in Thailand.

Although there is a sizable literature on statistical methods for handling missing data (see Little and Rubin, 2002), these methods are not directly appropriate for correcting incidence rates as described above. For subjects with unknown age in cancer registries, Fallah and Kharazmi (2008) compared several methods, finding them to be superior to

* Corresponding author.
Email address: tphattra@bunga.pn.psu.ac.th

conventional methods of simply inflating by a constant or allocating the unknown ages to the oldest age group. Although similar methods could be applied to categorical factors other than age group, including nominal grouped data such as province or district (using the size of the group to rank the data), these methods are essentially *ad hoc*.

Logistic regression (McNeil, 1996; Woodward 1999) provides a more promising general method for handling predictors with unknown values when estimating incidence rates, because the problem can be formulated in terms of a binary outcome (unknown or known value of the predictor) with other known factors as predictors. This method has been used to address missing data in other studies, including case-control studies (see, for example, Li *et al.,* 2004). An extensive study by Williams *et al.* (2005) approached the problem using a Bayesian formulation of logistic regression to calculate the posterior distribution by integrating out missing data. In this paper we describe a simple method for correcting mortality data in the presence of cases with unknown factors, based on the proportions of missing data with known factors for other variables. It involves fitting a logistic regression model and thus provides estimates and confidence intervals for proportions. The method is illustrated using all-cause mortality data from Thailand in 1996-2009.

2. Data and Methods

The DR database in year 1996-2009 was provided by the Bureau of Registration Administration, Ministry of Interior and coded as cause-of death by the Bureau of Policy and Strategy (2010), Ministry of Public Health. The reported deaths were classified in 18 five-year age groups up to 85-89 and 90 or more by gender in 76 provinces in Thailand. The numbers of deaths reported were tabulated into cells based on all combinations of the two sexes, the 77 groups comprising the provinces and an additional "unknown province" group, and 20 groups comprising the age groups and a further "unknown age" group.

The basic method involves fitting the generalized linear model from the binomial family with logit link by maximum likelihood as described in Venables and Ripley (2002). Logistic regression formulates the logit of the probability that a person died with unknown province or unknown age group as a linear function of the determinant factors. For the data considered, the binary outcome was defined as either (a) unknown or known province, or (b) unknown or known age. In each case the model fitted was an additive combination of two factors. The models are formulated as

$$logit\left(P_{ij}\right) = ln\left(\frac{P_{ij}}{1-P_{ij}}\right) = \mu + \alpha_i + \beta_j, \qquad \text{(a)}$$

where P_{ij} is the probability of death with unknown province, μ is a constant, α_i and β_j are individual parameters specifying gender i and age group j (j=1,2,3,…,19), respectively.

$$logit\left(P_{ik}\right) = ln\left(\frac{P_{ik}}{1-P_{ik}}\right) = \mu + \alpha_i + \gamma_k, \qquad \text{(b)}$$

where P_{ik} is the probability of death from unknown age group, α_i and γ_k are individual parameters specifying gender i and province k (k=1,2,3,…,76), respectively. The data for each model comprised all cases with known age or province, and thus only those cases with both province and age unknown were omitted. The models were fitted sequentially with the total numbers of known deaths in provinces for model (b) inflated and rounded to integers using results from model (a). The adequacy of fit of each model was assessed by using a plot of deviance residuals versus theoretical quantiles.

The adjusted proportions from model (a) were used to correct deaths in each age group with unknown province by multiplying the reported numbers in gender i and age group j by $1/(1-P_{ij})$, where P_{ij} is the estimated proportion of deaths with unknown province in gender i (i =1 for males) and age group j. Thus, the estimated numbers of deaths were obtained.

In the same way, we corrected deaths in each province with unknown age group by multiplying the estimated numbers of deaths from model (b) in gender i and province k by $1/(1-P_{ik})$, where P_{ik} is the estimated proportion of deaths with unknown age group in gender i and province k. These procedures thus redistribute reported deaths with unknown province for each gender and age group to known province and redistribute reported deaths with unknown age group for each gender and province to known age groups.

The adjusted proportions of deaths with unknown province or unknown age were presented using graphs of confidence intervals. Since it is more appropriate to compare province effects with their overall mean, rather than with an arbitrary province, the standard errors for the estimated parameters in the model are based not on the conventional treatment contrasts but on weighted sum contrasts. A method for doing this is described by Tongkumchum and McNeil (2009), with thematic mapping of regions based on this method described by Odton *et al.* (2010a,b).

After allocation of deaths with unknown province or unknown age, natural cubic splines were used to interpolate age-specific demographic data to ensure that relevant boundary conditions on second derivatives are satisfied as described by McNeil *et al.* (2011). Age-specific demographic data functions are necessarily non-negative. This method can be used to smoothly interpolate non-negative mortality data. "Cubic spline interpolation is a useful technique to interpolate between known data points due to its stable and smooth characteristics" (Kruger, 2003).

All data analysis was undertaken and graphical displays created using basic R software (R Development Core Team, 2012).

Table 1. Number of deaths and percentage unknown of province and age, 1996-2009

year	number of deaths	percentage unknown			
		province	age	both	total
1996	342,643	0.44	4.78	0.00	5.22
1997	300,321	1.06	2.72	0.14	3.92
1998	310,535	1.60	0.20	0.00	1.80
1999	362,607	1.15	0.89	0.00	2.05
2000	365,741	2.03	0.42	0.00	2.45
2001	369,494	0.34	0.30	0.00	0.65
2002	380,364	3.53	0.25	0.07	3.86
2003	384,131	4.27	0.24	0.01	4.52
2004	393,592	6.82	0.19	0.01	7.02
2005	395,374	0.51	0.35	0.00	0.86
2006	391,126	0.39	0.32	0.00	0.71
2007	393,254	0.04	0.20	0.00	0.23
2008	397,327	0.02	0.06	0.00	0.08
2009	393,916	0.01	0.05	0.00	0.06

3. Results

Annual numbers of deaths from 1996 to 2009 varied from 300,321 to 397,327 cases. Percentages of DR data for numbers of deaths with an unknown province and age are shown in Table 1. The highest unknown province is in the year 2004 whereas unknown age is in the year 1996. The percentages of deaths with unknown province or age are negligible after 2006, but need correction in earlier years.

Logistic regression models provide an acceptable fit for both proportions of unknown province and age. Figure 1 shows logistic regression modeling results from model (a). The graphs show adjusted percentages of deaths with unknown provinces in 1997 (upper panel) and 2004 (lower panel). The 95% confidence intervals are plotted for comparing estimated percentages of deaths with unknown provinces from the overall mean by gender and age group (red line). Overall percentage of unknown province deaths increased from 1.1% to 6.8% from 1997 to 2004.

The pattern of unknown provinces in 1997 and 2004 are similar. Percentage of unknown provinces above average mostly occurred in the worker age groups, 15-44 years, including the highest in age group 0-4 years. Unknown provinces below average occurred in ages over 60 years. The graphs also indicate that the percentage of deaths with unknown province for males were slightly lower than that for females in 1997 and higher in 2004.

Figure 2 graphs similar results from the logistic regression model (b). The model estimates the percentages of deaths with unknown age by gender and province. The provinces with codes of 10-27 and 70-77 are in the Central region, 30-49 in the North-East, 50-67 in the North and 80-96 in the South. In 1997, over 15% of female unknown age deaths occurred in Rayong (code 21) and Chanthaburi (code 22) provinces (upper panel). Unknown-age deaths decreased from 2.7% to 0.2% from 1997 to 2004 (red line). In 2004, most provinces had unknown age around the mean and below 1%

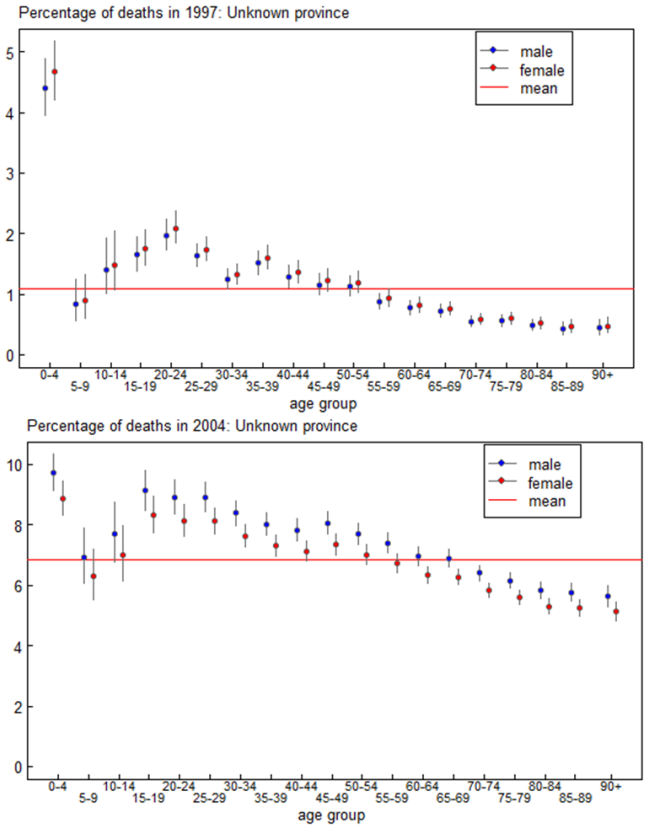

Figure 1. Estimates with 95% confidence intervals for percentages of deaths with unknown province in Thailand in 1997 (upper panel) and 2004 (lower panel)

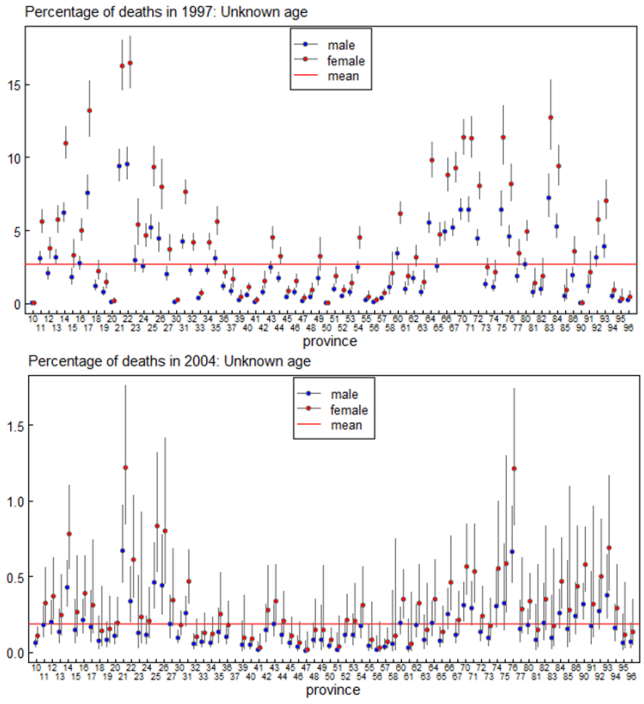

Figure 2. Estimates with 95% confidence intervals for percentages of deaths with unknown age in Thailand in 1997 (upper panel) and 2004 (lower panel)

with the exception of females in Rayong (code 21) and Phetchaburi (code 76) provinces (lower panel). Unknown-age deaths were reported to a greater extent for females.

The thematic maps clearly illustrate unknown age changed for both genders in each province from 1997 to 2004 as shown in Figure 3. They separate three levels of these percentages into groups where the confidence intervals are entirely above the mean (coloured in red), crossing the mean (orange) and entirely below the mean (cyan). In 1997 (left panel), nineteen provinces had above-average unknown age percentage occurring mainly in the provinces of the Central and the Southern regions. Lower than average percentages of reported deaths with unknown age mostly occurred in the North and North-East provinces. In 2004 (right panel), nine provinces had above-average unknown age percentages. Seven provinces in the central and Phatthalung remained unchanged from 1997 to 2004, whereas the percentage for Songkhla changed from below-average to above-average. Provinces below-average remained unchanged in Chiang Mai, Chiang Rai, Udon Thani, Sakon Nakhon, Khon Kaen, Bangkok and Amnat Charoen (no unknown age deaths in 2004).

Missing data cause biases and can lead to misunderstanding of information. The methods in this study applies to data with unknown province or age and thus provide estimated numbers of deaths for further investigation of mortality patterns and trends. This method allocates numbers of deaths and provides estimation with 95% confidence intervals.

We graphed the changes in smoothed age distributions of deaths for 76 provinces by fitting cubic spline to cumulative death counts after allocation of deaths with unknown province or age. It shows male peaks occurred around age 25 in several provinces in 1996 including Chiang Mai, Lumphun, Lampang, Payao, Chiang Rai, Phetchaburi and Ranong, but these peaks had disappeared by 2009.

Figure 4 shows three patterns of trends in spline-smoothed deaths from 1996-2009 in six provinces. First, Narathiwat and Phuket show increasingly high levels of infant mortality. Second, Bangkok and Khon Kaen have quite similar patterns, with minor peaks between ages 20 and 30. Third, Chiang Rai and Chiang Mai have very much higher peaks in this age range. Some anomalies are apparent. Infant mortality levels for Phuket in 1998 and 1999 were about three times higher than in other years, whereas no infant deaths were reported in Bangkok in 1997, probably due to incomplete data reporting.

We compared male and female patterns in Narathiwat, Bangkok and Chiang Rai provinces as shown in Figure 5. Numbers of female deaths are generally lower, but the comparison is distorted by the greater numbers of male deaths in ages 20-40 in Chiang Rai and to a lesser extent in Bangkok.

The total number of deaths in any year for a province is simply the area under its curve. For example, the excess number of male deaths in age group 20-40 years for Chiang Rai in 1996 is approximated by the triangle with area 240× 30/2=3600. The triangle is superimposed on the curve for male deaths in 2009 and shaded in yellow (top right panel). A triangle with similar area is also superimposed on the curve for female deaths in 2009 (bottom right panel). It shows that the number of male deaths in the age group 20-40 in 1996 is similar to the number of female deaths in all age groups in 2009.

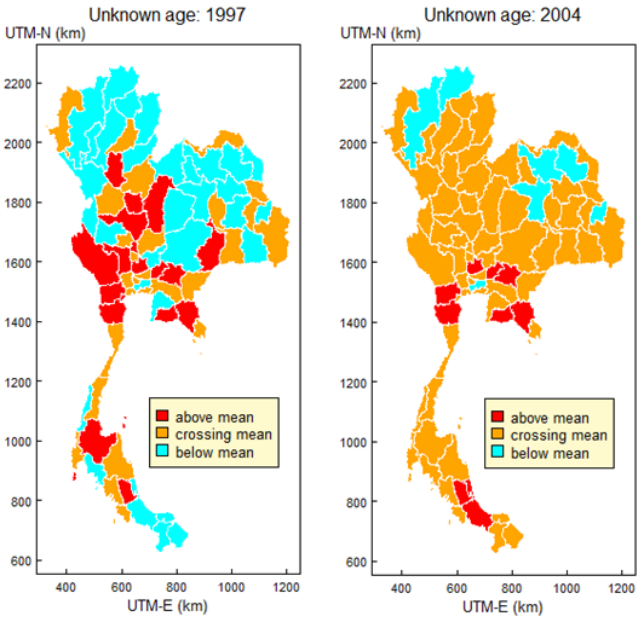

Figure 3. The thematic map of unknown age in 1997 and 2004

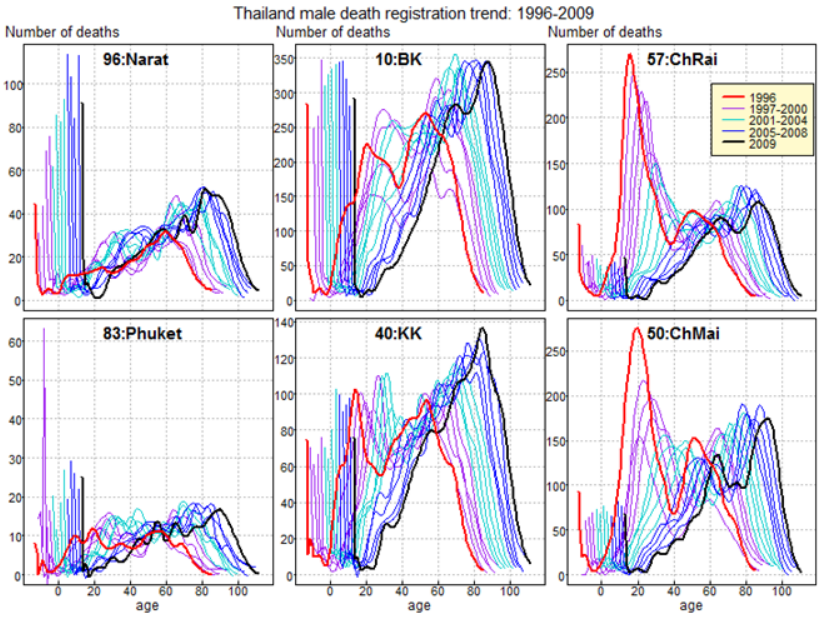

Figure 4. Trends in spline-smoothed deaths in six provinces from 1996-2009

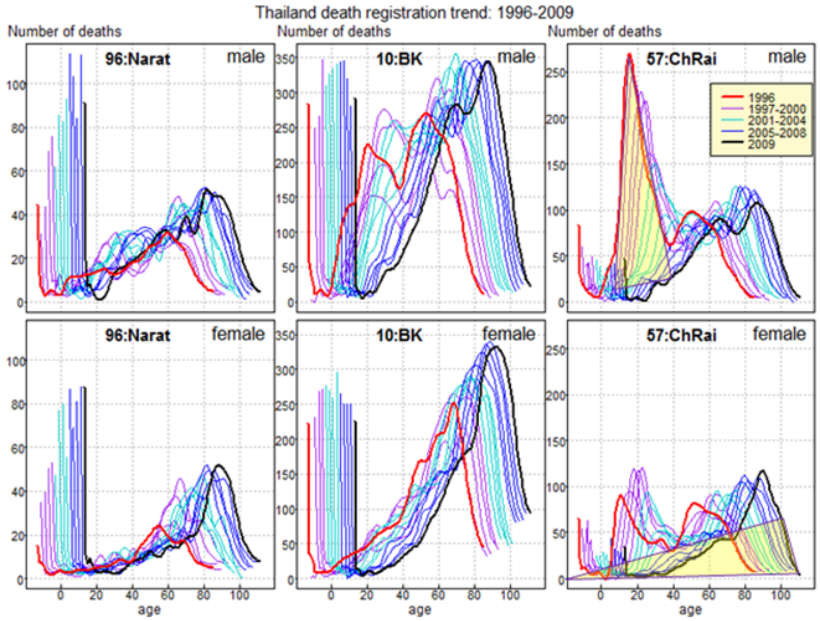

Figure 5. Trends in spline-smoothed deaths by gender from 1996-2009

4. Discussion

Reasons for the absence of demographic details on death certificates could be due to persons dying away from relatives and without ID cards. The age distribution of deaths with an unknown province of residence mostly occurred in worker age groups, suggesting that many such cases could be due to injuries occurring outside the province of residence or migration. The number of deaths with unknown province increased from 1996 to 2004, possibly partially due to migration (Vapattanawong and Saplon, 2011). Thus it would be informative to investigate such cases further. Our finding that deaths with unknown age were higher among females also needs further investigation.

Since death certificates in Thailand also provide information of the deceased person's age, gender and district (available from the Ministry of Interior), it is possible that further accuracy could be gained by using the 926 districts into which the nation is divided, rather than the 76 provinces. Faramnuayphol *et al.* (2008) used district as the location

factor in their regional comparison of cause-specific mortality, and Odton *et al.* (2010b) used 235 aggregated districts with similar populations in their study of regional variation of age-group and gender-adjusted all-cause mortality. However, neither of these studies made adjustments for deaths reported with unknown region or age. Although in a subsequent study Odton *et al.* (2010a) investigated regional variations in deaths reported with unknown cause, correcting for unknown cause in mortality data is essentially a different problem, because in this case there is no cause-specific population denominator.

Logistic regression is a method to correct a misclassification binary outcome. It is appropriate for dealing with missing data. Many researchers have used this method to correct for missing data in their studies (Duffy *et al.*, 2004; Li *et al.,* 2004; Lyles *et al.,* 2011). For simplicity and conciseness in smoothed age distributions of deaths, cubic spline interpolation is an efficient strategy for smoothing cumulative death counts (Wand, 2000; McNeil *et al.,* 2011). The methods in this study minimize the effects of less than perfect data and provide more accurate estimation of mortality with 95% confidence intervals. They also provide valuable information to the Ministry of Public Health for efficient health policy planning.

Acknowledgements

We thank Prof. Dr. Don McNeil, Professor Emeritus of Statistics at Macquarie University in Australia, for his helpful guidance. We also thank the Bureau of Registration Administration in the Ministry of Interior and the Bureau of Policy and Strategy in the Ministry of Public Health for providing the data.

References

Bureau of Policy and Strategy. 2010. Vital Registration Database, Ministry of Public Health, Thailand.

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

Duffy, S.W., Warwick, J., Williams, A.R.W., Keshavarz, H., Kaffashian, F., Rohan, T.E., Nili, F. and Sadeghi-Hassanabadi, A. 2004. A simple model for potential use with a misclassified binary outcome in epidemiology. Journal Epidemiology Community Health. 58, 712-717.

Fallah, M. and Kharazmi, E. 2008. New methods of handling cases of unknown age in cancer registry data. Asian Pacific Journal of Cancer Prevention. 9(2), 259-262.

Faramnuayphol, P., Chongsuvivatwong, V. and Pannarunothai, S. 2008. Geographical variation of mortality in Thailand. Journal of the Medical Association of Thailand. 91(9), 1455-1460.

Kruger, C.J.C. 2003. Constrained cubic spline interpolation for chemical engineering Applications. Available from: http://www.korf.co.uk/spline.pdf. [November 28, 2012]

Li, X., Song, X. and Gray, R.H. 2004. Comparison of the missing-indicator method and conditional logistic regression in 1:m matched case-control studies with missing exposure values. American Journal of Epidemiology. 159(6), 603-610.

Little, R.J.A. and Rubin, D.B. 2002. Statistical Analysis with Missing Data, Wiley-Interscience, New York, U.S.A.

Lyles, R.H., Tang, L., Superak, H.M., King, C.C., Celentano, D.D., Lo, Y. and Sobel, J.D. 2011. Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. Journal of Epidemiology. 22(4), 589-597.

McNeil, D. 1996. Epidemiological Research Methods, Wiley, Chichester, England, pp. 125-194.

McNeil, N., Odton, P. and Ueranantasun, A. 2011. Spline interpolation of demographic data revisited. Songklanakarin Journal of Science and Technology. 33(1), 117-120.

Odton, P., Bundhamcharoen, K. and Ueranantasun, A. 2010a. District-level variations in quality of mortality data in Thailand. Asia-Pacific Population Journal. 25(1), 79-91.

Odton, P., Choonpradub, C. and Bundhamcharoen, K. 2010b. Geographical variations in all-cause mortality in Thailand. Southeast Asian Journal of Tropical Medicine and Public Health. 41(5), 1209-1219.

R Development Core Team. 2012. The R project for statistical computing. Applications. Available from: http://www.r-project.org [September 9, 2012]

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

Vapattanawong, P. and Saplon, O. 2011. Deaths outside residential area of Thais: study from death registration, 1996-2009. Thai Population Journal. 3(1), 73-89.

Venables, W.N. and Ripley, BD. 2002. Modern Applied Statistics with S, Springer, New York, U.S.A., pp. 183-199.

Wand, M. P. 2000. A comparison of regression spline smoothing procedures. Computational Statistics. 15(4), 443-462.

Williams, D., Liao X., Ya, X. and Carin, L. 2005. Incomplete-data classification using logistic regression. Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, August 7-11, 2005, 972-979.

Woodward, M. 1999. Epidemiology: Study Design and Data Analysis, Chapman & Hall, U.S.A., pp. 448-481.