STATISTICAL MODELLING OF CHILDHOOD DIARRHEA IN NORTHEASTERN THAILAND

Jurairat Ardkaew[1] and Phattrawan Tongkumchum[2]

[1]Department of Sciences, Faculty of Science and Technology, Loei Rajabhat University, Loei; [2]Department of Mathematics and Computer Sciences, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand

Abstract. Diarrhea remains an important cause of morbidity in Thailand, particularly for children below age five. This study identified the patterns of diarrhea incidence in children below age five in northeastern provinces of Thailand along the border with Lao PDR, based on the individual hospital case records of patients with diarrhea routinely reported from 1999 to 2004. Linear regression models containing the district, season and year as factors were fitted to the log-transformed disease incidences, with generalized estimating equations used to account for spatial correlation between districts. Low disease counts suggesting under-reporting were handled by imputation based on these models. This study found a seasonal pattern higher in January to March and April to June. Higher rates occurred in most districts of Loei and Amnat Chroen Provinces. Using a thematic map to display the level of diarrhea incidence by district can provide useful information for health authorities to direct their intervention plans more effectively and to set up health policies for prevention of disease.

INTRODUCTION

Diarrhea is an infectious disease that is the third highest listed cause of mortality at 3.2% of total deaths worldwide. It is estimated that 1.8 million people die each year due to diarrhea, most of them are children below age five (Brownlie *et al*, 2006). Intestinal infection casing diarrhea may be caused by bacterial, viral or parasitic organisms, most of which are spread by contaminated water. It is more common where there is a shortage of clean water for drinking, cooking and sanitation. Diarrhea can also spread from person to person, aggravated by poor personal hygiene (WHO, 2009). It continues to be a major cause of morbidity and mortality among children in developing countries. Kosek *et al* (2003) reported the diarrhea incidence among children below age five in studied countries was on average 3.2 episodes per child per year. In Thailand, the mortality of children below age five due to diarrhea has decreased over recent years, but its annual incidence rate has increased from 3,031 cases per 100,000 population in 1984 to 7,243 cases per 100,000 in 2003 (Wibulpolprasert *et al*, 2004). Furthermore, according to hospital records in the Annual Epidemiological Surveillance Report 2006 of Thailand, diarrhea is the disease with the highest incidence of morbidity, with approximately 1.24 million hospital-reported cases, 33.9% of them involving children less than five years of age. This age group had the highest morbidity rate, namely, 10,610 per 100,000 (Bureau of Epidemiology, 2006).

Many people do not seek medical atten-

Correspondence: Phattrawan Tongkumchum, Department of Mathematics and Computer Sciences, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus 94000, Thailand.
E-mail: tphattra@bunga.pn.psu.ac.th

tion for diarrhea, thus hospital records may not be adequate for estimating the incidence of diarrhea to be used for the development of health programs (Woldemicael, 2001). Successful health care intervention requires an accurate knowledge of the disease (Kandala *et al*, 2006), including its distribution by geographical region and its seasonal pattern.

The aim of our study was to identify the patterns of diarrhea incidence in the northeastern border provinces of Thailand. The study used statistical models that combined linear regression and generalized estimating equations (GEE).

MATERIALS AND METHODS

Study area and data

This study focuses on the regions of the northeast border provinces of Thailand, located along the Mekong River, which forms the boundary between Thailand and Lao PDR. The five border provinces in this study were: Loei, Nong Khai, Nakhon Phanom, Mukdahan and Amnat Charoen, containing 57 districts (Fig 3).

There are three seasons during the year in these provinces: rainy, winter and summer. The rainy season starts around the middle of May and ends around mid-October. The winter begins in mid-October and ends around mid-February. This is followed by the summer from mid-February to mid-May (Prukpitikul *et al*, 2007). Data regarding the hospital diagnosed cases of diarrhea for 1999 were obtained from the provincial offices of the Ministry of Public Health. Data in this study cover the six-year period of 1999-2004 for the five border provinces. Reported cases for each year were available in computer files with individual records for disease cases and fields comprising characteristics of the subject and the disease, including dates of illness, the diagnosis, subject's age, gender, address, and severity of illness, including date of death for mortality cases. The resident population denominator used to compute the incidence rates was obtained from the Population and Housing Census of 2000 undertaken by the National Statistics Office of Thailand. Preliminary analysis showed little evidence of a gender effect so data for the two sexes were combined. The study periods are defined as January-March, April-June, July-September and October-December, giving 24 quarter periods for the six years, effectively eliminating correlation between disease rates in successive periods.

Statistical methods

First, we defined the outcome variable as the disease rate in a cell indexed by district, month and year. We estimated the reported disease incidence in the population by converting this to an annual rate per 1,000. Rates of disease generally have positively skewed distributions so it is conventional to transform them by taking logarithms and the simplest model is the additive combination

$$\ln\left(\frac{n_{ijt}}{P_i}\right) = y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_t \quad (1)$$

In this model, P_i is the population in 1000s in district *i,* and n_{ijt} is the corresponding number of reported cases in district *i* and month *j* of year *t*. Since some cells had no reported cases preventing log-transformation, we first omitted these cells from the model and subsequently replaced the zeros by fitted values based on the model. After examining plots of standardized residuals, the model was refitted after further omission of cells corresponding to residuals below a specified cutoff value, with the object of improving the fit to normal distribution, and the omitted occurrences were again imputed using the model. Since infectious disease hospital cases are believed to be not fully reported in many districts of Thailand and the data contain some low counts that are statistically improbable, this method also reduces bias due to such occurrences. To reduce or eliminate the correlations that also occur in disease counts in successive months, we then aggregated the monthly data (including the imputed counts) into quarterly incidence rates and fitted a model similar to (1) with *j* now representing a quarter instead of a month.

We used the generalised estimating equation (GEE) method (Liang and Zeger, 1986; Yan and Fine, 2004) with fixed correlation structure to account for spatial correlation between districts. In this structured correlation matrix, correlations between different districts within a given province were specified as the common mean of the corresponding residual correlation coefficients after fitting the linear model, unless this value was less than 0.1, in which case it was rounded down to 0. Correlations between districts in pairs of different provinces were similarly fixed at the means of the corresponding residual correlations, provided these means exceeded 0.1 in magnitude. The use of the GEE method for modelling spatial correlation is discussed in a recent review by Dormann *et al* (2007).

Since residuals obtained from the GEE model remain correlated even though the method gives estimates and standard errors of parameters adjusted for the assumed spatial correlations, before plotting them we removed their correlation using a linear filter as follows. Suppose that Z is the n x k matrix of residuals after fitting the GEE model, where n is the total number of periods and k is the number of districts, and C is their correlation matrix. Define $A = ED^{1/2}E^T$, where E is the matrix of eigenvectors of C and D is the matrix containing the corresponding eigenvalues. Then, $W = ZA^{-1}$ is the matrix of residuals transformed to have uncorrelated columns.

To obtain unbiased estimates of incidence rates in cells we used the formula

$$\hat{r}_{ijt} = \exp(\hat{y}_{ijt} + c) \ , \qquad (2)$$

where $\hat{y}_{ijt}$ is the fitted value of y_{ijt} and c is a constant chosen to match the total number of observed cases with the total given by the model. After fitting the model, unbiased incidence rates for levels of each factor adjusted for other factors were calculated similarly. Standard errors for these adjusted incidence rates were obtained by using sum contrasts (Venables and Ripley, 2002) to compare the incidence rates for each level of a factor with the overall mean incidence rate. By using this method, the pattern of diarrhea was identified for each factor.

Since the confidence intervals for factor-specific incidence rates obtained from model divide naturally into three groups according to their location entirely above the mean, around the mean, or entirely below the mean, we used this trichotomy to create thematic maps of districts according to their estimated diarrhea annual incidence rates.

The R program (R Development Core Team, 2007) was used for all statistical analysis, graphs and maps.

RESULTS

Preliminary results

In the five northeast border provinces studied, 121,615 cases of diarrhea from 1,470,924 children below age five were diagnosed at district hospitals from 1 January 1999 to 31 December 2004. The maximum number of cases reported by month and district was 289, which occurred in Loei City in January 1999. The highest and lowest annual disease incidences by province were 101.0 per 1,000 in Amnat Charoen and 67.2 per 1,000 in Nong Khai, respectively (Table 1).

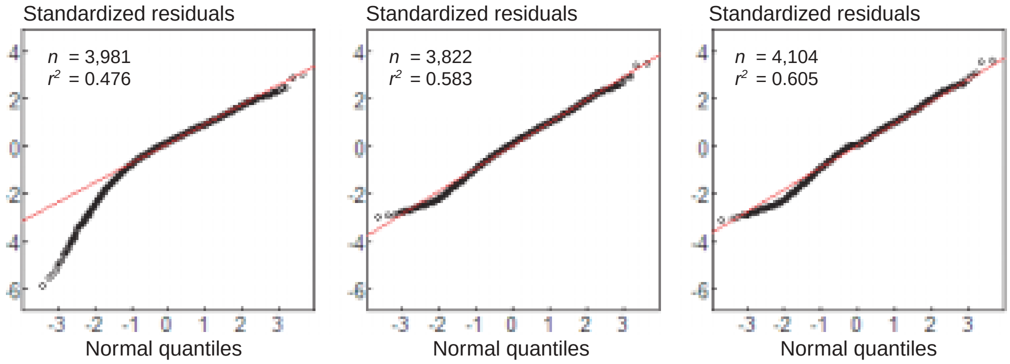


Fig 1–Residuals plots after fitting the model to the diarrhea incidence rate after omitting zero counts (left panel), omitting further counts with low residuals (middle panel), and imputing counts omitted by the model (right panel).

Table 1
Average annual diarrhea incidence per 1,000.

Province	Number of districts	Children below age five		
		Cases	Population	Incidence
Loei	12	26,702	275,448	96.9
Nong Khai	17	31,490	468,348	67.2
Nakhon Phanom	14	29,240	369,330	79.2
Mukdahan	7	14,816	166,014	89.2
Amnat Charoen	7	19,367	191,784	101.0

Statistical analysis

The results are presented in Figs 1, 2 and 3. Fig 1 shows the standardized residuals after fitting model (1) to the monthly data with three different methods for handling low disease counts. In the left panel the 123 cells with zero counts were simply omitted, whereas the middle panel shows the same graph after omitting a further 159 cells with residuals below -2 and refitting the model, and the right panel shows the result of using this model to impute the counts for all 282 omitted cells and refitting the model to all 4,104 cells. Including the imputed counts, this method gave a total disease count of 126,799 cases, 4.1% more than the number reported.

Table 2 shows common means and standard errors of spatial residual correlations among all districts in each province after fitting the model. It shows no evidence of a spatial auto-correlation in the districts of Mukdahan Province, but the four other provinces show auto-correlations among districts ranging from 0.1 in Loei to 0.46 in Amnat Charoen. There is no evidence of correlation among districts in different provinces, with the single exception of a negative cross-correlation between districts in Amnat Charoen and Nong Khai.

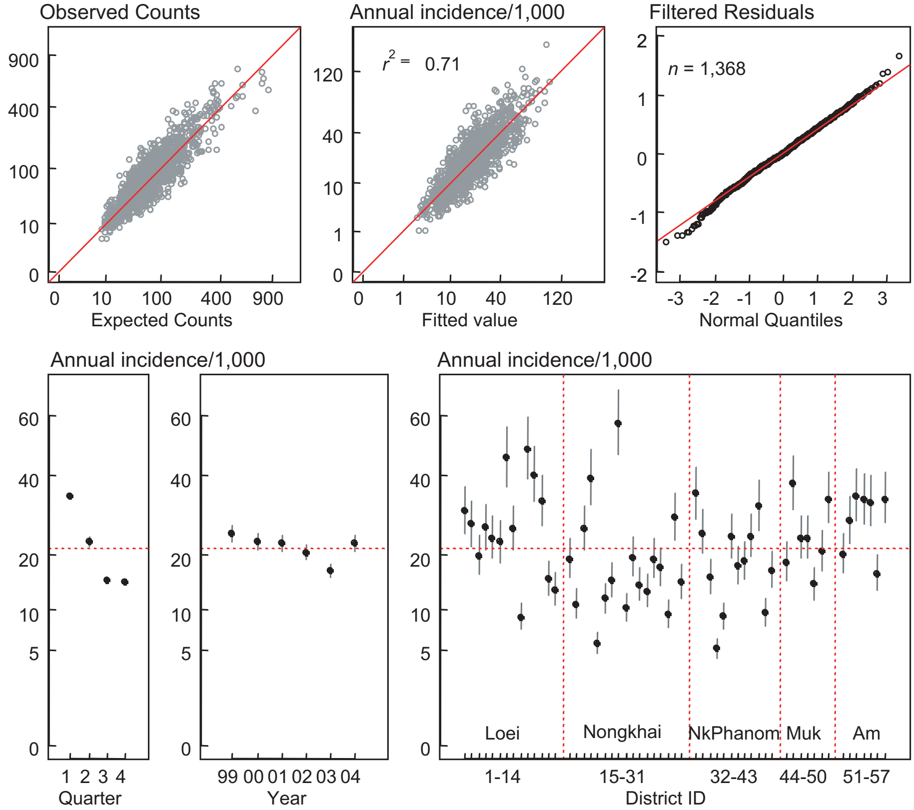
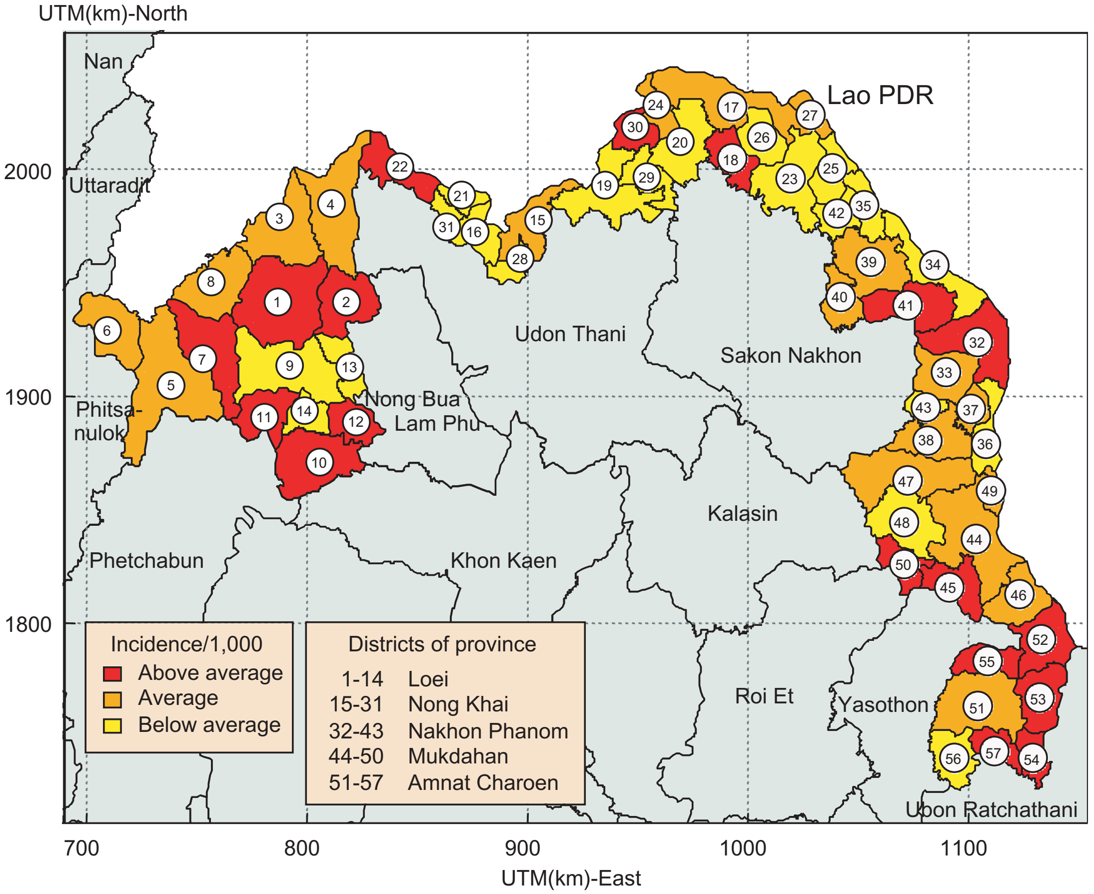


Fig 2–Results of fitting model.

Table 2
Spatial residual correlation within districts by provinces.

Province	Loei	Nong Khai	Nakhon Phanom	Mukdahan	Amnat Charoen
	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)	Mean (SE)
Loei	**0.10** (0.03)	0.06 (0.03)	0.20 (0.03)	-0.03 (0.03)	-0.02 (0.04)
Nong Khai		**0.17** (0.02)	0.02 (0.02)	0.02 (0.03)	**-0.12** (0.03)
Nakhon Phanom			**0.14** (0.03)	0.06 (0.03)	0.03 (0.03)
Mukdahan				0.00 (0.04)	0.02 (0.04)
Amnat Charoen					**0.46** (0.04)

Fig 2 shows the results after fitting model (1) with quarterly periods using the GEE method with the fixed correlation structure given in Table 2. The left and middle top panels show plots of observed counts and observed annual incidence rates per 1,000 versus corresponding fitted values, respectively. The top right plot shows filtered residuals versus normal quantiles and indicates a satisfactory fit.



1 Loei City	13 Erawan	25 Bueng Khong Long	37 Renu Nakhon	49 Wan Yai
2 Na Duang	14 Nong Hin	26 Si Wilai	38 Na Kae	50 Nong Sung
3 Chiang Khan	15 Nong Khai City	27 Bueng Khla	39 Si Songkhram	51 Amnat Charoen City
4 Pak Chom	16 Tha Bo	28 Sakhrai	40 Na Wa	52 Chanuman
5 Dan Sai	17 Bueng Kan	29 Fao Rai	41 Phon Sawan	53 Pathum Ratchawongsa
6 Na Haeo	18 Phon Charoen	30 Rattanawapi	42 Na Thom	54 Phana
7 Phu Ruea	19 Phon Phisai	31 PhoTak	43 Wang Yang	55 Senangkhanikhom
8 Tha Li	20 So Phisai	32 Nakhon Phanom City	44 Mukdahan City	56 Hua Taphan
9 Wang Saphung	21 Si Chiang Mai	33 Pla Pak	45 Ni Khom Kham Soi	57 Lue Amnat
10 Phu Kradueng	22 Sangkhom	34 Tha Uthen	46 Don Tan	
11 Phu Luang	23 Seka	35 Ban Phaeng	47 Dong Luang	
12 Pha Khao	24 Pak Khat	36 That Phanom	48 Khamcha-i	

Fig 3–Thematic map of diarrhea incidence: (1999-2004).

The graphs in the lower panels show 95% confidence intervals of annual diarrhea incidence/1,000 by quarter (left panel), year (middle panel) and district (right panel), each adjusted for the effects of the other two factors in the model. The dotted horizontal lines represent the overall mean annual incidence rate (22.9 per 1,000).

The seasonal pattern of diarrhea incidence clearly indicates a peak in the January to March quarter of each year (33.8 per 1,000) with much lower rates in the second half of each year (14.9 and 14.6 per 1,000, respectively). The incidence decreased from 24.6 in 1999 to 16.6 in 2003 but increased to 22.5 in 2004. The variation between districts was greater, ranging from 5.1 (95% CI 4.2 - 6.3) in Sangkhom District of Nong Khai Province to 56.9 (95% CI 46.4 - 69.9) in Ban Phang District of Nakhon Phanom Province. The incidence of diarrhea was higher than average in districts of Amnat Charoen and Loei Provinces, lower than average in Nong Khai, and average in the districts of the other three provinces.

Fig 3 maps the spatial distribution in districts classified according to whether their confidence intervals exceed, cross, or are below the overall mean. Higher disease incidence occurred mainly in Loei (6 of 12 districts) in the northwest and Amnat Chroen (5 of 7 districts) in the southeast. In contrast, the middle region showed mostly lower disease incidence rates.

DISCUSSION

In this study, we analyzed the patterns of childhood diarrhea incidence at the district level in five border provinces of northeastern Thailand using linear regression models containing districts, seasons, and years as factors with generalized estimating equations used to account for spatial correlation between districts.

It is also important to note that although many studies have used negative binomial regression models to analyse disease counts (Lindsey and Jones, 1998; Lim and Choonpradub, 2007; GschlöBl and Czado, 2008), this study fitted a linear regression model of the log-transformed incidence rate because it is simpler and software is readily available for handling spatial correlation. This model provided reasonable and explainable results of patterns of disease incidence and provides a good fit. We found diarrhea annual incidence varied by season, year, and district.

The patterns of diarrhea incidence decreased from 1999 to 2003 followed by a sharp increase in 2004 to a level just above the overall mean of annual incidence rate for the period.

Seasonal patterns of childhood diarrhea have been noted in many tropical areas, where there are two definite seasonal peaks: the summer is associated with bacterial infections, and the winter is related to viruses (Gracey, 1996). In our study, the diarrhea incidence rates were higher in January to March and April to June. That period overlaps the winter, summer, and part of the rainy season, so it is likely that both bacterial and viral factors may be associated with the high incidence of diarrhea. Similar trends were reported by Pinfold *et al* (1991) in their study of diarrhea in northeastern Thailand, and in other studies from Korea and Vietnam (Cho *et al*, 2006; Hien *et al*, 2008). Rotavirus was found to be the major cause of winter diarrhea in children (Maneekarn and Ushijima, 2000). Water used by households during the rainy season can become contaminated with pathogens due to excessive rainfall, particularly in rural areas where sanitation is below acceptable standards (Asian Development Bank, 1999; Boonprakrob and Hattirat, 2006; WHO, 2009).

The areas with the highest incidence of diarrhea were the poorer districts of Loei and Amnat Charoen Provinces, particularly in Amnat Charoen, which is the second poorest province in northeastern Thailand (Statistical Forecasting Bureau, 2007). People in rural areas often migrate to large cities seeking work, leaving their children with their parents without providing adequate resources. As a result, the children grow up in poverty with poor sanitation and low quality food.

The higher incidence of diarrhea in Sangkhom District of Nongkhai Province could be due to flooding from the Mekong River during the rainy season.

This model can be used to identify unusually high incidence rates of diarrhea among young children where and when they occur, and thus enable health authorities to reduce the severity of ensuing epidemics by putting preventative measures in place.

ACKNOWLEDGEMENTS

We thank the provincial officers of the Ministry of Public Health, for providing the data. This study was funded by the Graduate School, Prince of Songkla University. We are grateful to Professor Don McNeil for his assistance.

REFERENCES

Asia Development Bank. Policy for the health sector. Manila: Asia Development Bank Publication, 1999.

Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health. Annual epidemiological surveillance report 2006. Bangkok: The War Veterans Organization of Thailand, 2006.

Brownlie J, Peckham C, Waage J, *et al*. Foresight. infectious diseases: preparing for the future threats. London: Office of Science and Innovation, 2006.

Boonprakrob K, Hattirat S. Crisis of opportunity: Climate change and Thailand [press release]. Bangkok: Greenpeace Southeast Asia 2006 Jul 16.

Cho SH, Kim JH, Kim JC, Shin HH, Kang YH, Lee BK. Surveillance of bacterial pathogens associated with acute diarrheal disease in the Republic of Korea during one year, 2003. *J Microbiol* 2006; 44: 327-35.

Dormann CF, McPherson JM, Araujo MB, *et al.* Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 2007; 30: 609-28.

Gracey M. Diarrhea and malnutrition: a challenge for pediatricians. *J Pediatr Gastroenterol Nutr* 1996; 22: 6-16.

GschlöBl S, Czado C. Modelling count data with overdispersion and spatial effects. *Stat Papers* 2008; 49: 531-52.

Hien BTT, Scheutz F, Cam PD, *et al*. Diarrheagenic *Escherichia coli* and *Shigella* strains isolated from children in a hospital case-control study in Hanoi, Vietnam. *J Clin Microbiol* 2008; 46: 996-1004.

Kandala NB, Magadi MA, Madise NJ. An investigation of district spatial variations of childhood diarrhoea and fever morbidity in Malawi. *Soc Sci Med* 2006; 62: 1138-52.

Kosek M, Bern C, Guerrant RL. The global burden of diarrhoeal disease, as estimated from studied published between 1992 and 2000. *Bull World Health Organ* 2003; 81: 197-204.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.

Lim A, Choonpradub C. A statistical method for forecasting demographic time series counts, with application to HIV/AIDS and other infectious disease mortality in Southern Thailand. *Southeast Asian J Trop Med Public Health* 2007; 38: 1029-40.

Lindsey JK, Jones B. Choosing among generalized linear models applied to medical data. *Stat Med* 1998; 17: 59-68.

Maneekarn N, Ushijima H. Epidemiology of rotavirus infection in Thailand. *Pediatr Int* 2000; 42: 415-21.

Pinfold JV, Horan NJ, Mara DD. Seasonal effects on the reported incidence of acute diarrhoeal disease in Northeast Thailand. *Int J Epidemiol* 1991; 20: 777-86.

Prukpitikul S, Buakaew V, Kesdech W. Application of RS/GIS for monitoring typhoon and storm in Thailand. Proceeding of the Asian Conference on Remote Sensing (ACRC); 2006 Oct 9-13. Ulaanbaatar,Mongolia. Curran Associates Inc, 2007.

R Development Core Team. R: A language and environment for statistical computing.

Vienna, Austria: R Foundation for Statistical Computing, 2007. [Cited 2007 Jan 8]. Available from: URL: http://www.R-project.org

Statistical Forecasting Bureau, National Statistical Office. Statistical yearbook Thailand 2007 (Special edition). Bangkok: Text and Journal Publication, 2007.

Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer-Verlag, 2002: 146.

Wibulpolprasert S, Ekachampaka P, Wattanamano S, Taverat R. Thailand health profile 2001-2004. Bangkok: Printing Press Express Transportation Organization, 2004.

Woldemicael G. Diarrhoeal morbidity among young children in Eritrea: Environmental and socioeconomic determinants. *J Health Popul Nutr* 2001; 19: 83-90.

World Health Organization. Water-related disease, 2009. [Cited 2009 Jan 3]. Available from: URL: http://www.who.int/water_sanitation_health/diseases/diarrhoea/en/

Yan J, Fine J. Estimating equations for association structures. *Stat Med* 2004; 23: 859-80.