



**Statistical Methods for Modeling Incidence Rates with Application to
Diarrhea in Thailand, Tuberculosis in Nepal and Injury in Australia**

Sulawan Yotthanoo

**A Thesis Submitted in Fulfillment of the Requirements for the
Degree of Doctor of Philosophy in Research Methodology**

Prince of Songkla University

2010

Copyright of Prince of Songkla University

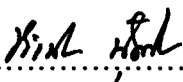
เลขทมิ	DA276	S94	2010
Bib Key	520153		
	7 ค.ย. 2553		

Thesis Title Statistical Methods for Modeling Incidence Rates with Application to
Diarrhea in Thailand, Tuberculosis in Nepal and Injury in Australia


Author Miss Sulawan Yotthanoo

Major Program Research Methodology

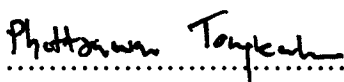
Major Advisor

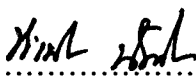

.....
(Asst. Prof. Dr. Chamnein Choonpradub)


Co-advisor

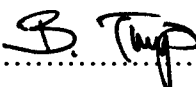

.....
(Emeritus Prof. Dr. Don McNeil)

Examining Committee:


 Chairperson
.....
(Dr. Phattrawan Tongkumchum)


.....
(Asst. Prof. Dr. Chamnein Choonpradub)


.....
(Emeritus Prof. Dr. Don McNeil)


.....
(Assoc. Prof. Dr. Bandit Thinkhamrop)

The Graduate School, Prince of Songkla University, has approved this thesis as fulfillment of the requirements for the Doctor of Philosophy in Research Methodology.


.....
(Assoc. Prof. Dr. Krerchai Thongnoo)
Dean of Graduate School

ชื่อวิทยานิพนธ์	วิธีการทางสถิติสำหรับการสร้างตัวแบบอัตราอุบัติการณ์ ประยุกต์ใช้กับข้อมูลโรคท้องร่วงในประเทศไทย วัณโรคในประเทศเนปาล และการบาดเจ็บในประเทศออสเตรเลีย
ผู้เขียน	นางสาวสุลาวัลย์ ยศธนู
สาขาวิชา	วิธีวิทยาการวิจัย
ปีการศึกษา	2552

บทคัดย่อ

ตัวแบบอัตราอุบัติการณ์สามารถอธิบายด้วยวิธีการทางสถิติ ได้หลายวิธี ซึ่งตัวแบบที่เหมาะสมจะสามารถอธิบายความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระได้

โดยปกติ นักสถิตินิยมใช้ ตัวแบบการถดถอยทั่วไป (Generalized Linear Models (GLMs)) ในการอธิบายตัวแบบอัตราอุบัติการณ์และใช้ตัวแบบการถดถอยทั่วไปที่ปรับขยาย เป็นตัวแบบสำหรับข้อมูลที่มีศูนย์จำนวนมาก (Zero-inflated GLMs) แต่อย่างไรก็ตามการแปลงค่าอัตราอุบัติการณ์ด้วยลอการิทึม หลังจากนั้นสร้างตัวแบบการถดถอยเชิงเส้นของลอการิทึม (Log-transformed linear regression model) จะได้ตัวแบบที่ดีเช่นกัน

วิทยานิพนธ์ฉบับนี้ ได้ทำการหาตัวแบบที่เหมาะสมของอัตราอุบัติการณ์ จากตัวแบบการถดถอยทั่วไปและตัวแบบการถดถอยเชิงเส้นของลอการิทึม โดยพิจารณาจากการแจกแจงปกติของค่าเศษเหลือ

ตัวแบบการถดถอยเชิงเส้นของลอการิทึมถูกนำมาประยุกต์ใช้กับข้อมูลโรคท้องร่วงในประเทศไทย เพื่อประมาณค่าจำนวนข้อมูลที่มีการรายงานต่ำกว่าปกติและใช้ตัวแบบสมการประมาณค่าโดยนัยทั่วไป (Generalized Estimating Equations (GEE) model) เพื่ออธิบายรูปแบบของข้อมูลเชิงพื้นที่และเวลา ตัวแบบการถดถอยทวินามนิเสธที่มีผลคูณของสององค์ประกอบหลัก (Negative binomial GLM with two multiplicative components) เป็นตัวแบบที่เหมาะสม สำหรับการประยุกต์ใช้กับข้อมูลวัณโรคในประเทศเนปาล นอกจากนี้ยังพบว่าตัวแบบการถดถอยทั่วไปมีความเหมาะสมเช่นเดียวกับตัวแบบเบย์เซียน (Bayesian model) สำหรับการวิเคราะห์ข้อมูลอัตราการบาดเจ็บในรัฐนิวเซาท์เวลส์ ประเทศออสเตรเลีย

Thesis Title Statistical Methods for Modeling Incidence Rates with Application to Diarrhea in Thailand, Tuberculosis in Nepal and Injury in Australia

Author Miss Sulawan Yotthanoo

Major Program Research Methodology

Academic Year 2009

ABSTRACT

There are several statistical methods for modeling incidence rates. The association patterns of outcome and determinant variables are identified by fitting the appropriate model.

Typically, generalized linear models (GLMs) are preferred by statisticians for modeling incidence rates, often with extensions to zero-inflated GLMs when the proportion of zero counts is large. However, log-transformed incidence rates can also be fitted well by a linear regression model.

In this thesis, suitable models for incidence rates were found after fitting GLMs and the log-transformed linear regression model after comparing appropriate residuals against the normal quantiles.

For application to diarrhea in Thailand, the log-transformed linear regression model was used to impute under-reported data with the generalized estimating equations (GEE) model to investigate regional and temporal patterns. The negative binomial GLM with two multiplicative components was preferred for modeling TB in Nepal. Additionally, the GLMs can fit the data just as well as Bayesian model for fitting injury incidence rates in NSW, Australia.

Acknowledgements

I am grateful to the Ministry of Public Health, Bangkok Thailand and the Nepal Tuberculosis Center (NTC), Bhaktapur Nepal for providing the data.

I would like to acknowledge the Graduate School and Faculty of Science and Technology, Prince of Songkla University for funding this study.

I would also like to express my sincerest gratitude and deepest appreciation to my supervisors, Professor Dr. Don McNeil and Assistant Professor Dr. Chamnein Choonpradub, for their helpful guidance, support and invaluable assistance throughout the completion of the thesis. I am also greatly indebted to Dr. Phattawan Tongkumchum and Assistant Professor Dr. Apiradee Lim for their helpful advice, to Greig Rundle for help with English corrections, to Sampurna Kakchapati for sharing his data from Nepal, to Dr. Shanley Chong for giving knowledge and sharing her data, and my good friend, Noodchanath Kongchouy, for recommending this program to study.

Last, but not least, I would like to thank my grandparents, my parents, my sister, and Naphongphot Supharp who tolerantly listened to my complaints and frustrations, for their encouragement and support throughout my study.

Sulawan Yotthanoo

Contents

	Page
Abstract	iii
Acknowledgements	v
Contents	vi
List of Table	viii
List of Figure	ix
Chapter	
1. Introduction	
1.1 Rationale for study	1
1.2 Literature review	4
1.3 The studies	9
1.4 Data collection	11
1.5 Objectives and plan of thesis	12
2. Methodology	
2.1 Data management	14
2.2 Poisson regression model	15
2.3 Negative binomial regression model	18
2.4 Log-transformed linear regression model	18
2.5 Logistic regression model	21

	Page
3. Modeling incidence rates with applications	
3.1 Studies completed	22
3.2 Preliminary analysis	23
3.3 Article 1	27
3.4 Manuscript 2	39
3.5 Manuscript 3	57
4. Summary and conclusions	
4.1 Summary of study results	69
4.2 Conclusions	77
References	79
Vitae	85

List of Table

Table	Page
Article 1	
Table 1: Number of reported diarrhea cases and population in each province	31
Table 2: Number of monthly reported diarrhea cases in ten selected districts	32
Table 3: Estimates of under-reporting percentages by quarter and year	33
Table 4: Means and standard errors of residual correlations between districts within and between provinces	34
Manuscript 2	
Table 1: Definitions and populations of super-districts	42
Table 2: TB incidence rates by year	46
Table 3: Analysis of Deviance for Poisson and Negative Binomial models	47
Manuscript 3	
Table 1: Hospitalisation injury incidence rates per 100,000 in NSW by age-gender group, quarter and statistical division, 2000-01 to 2004-05	61
Table 2: Comparison of results in LGAs for Bayesian and negative binomial Models	65
Table 4.1 Summary of under-reported cases categorized by district and year	70

List of Figure

Figure	Page
Figure 1.1 Path diagram for study	10
Figure 3.1: Numbers of monthly reported diarrhea cases in selected districts	24
Figure 3.2: Tuberculosis incidence rates by year in Nepal	25
Figure 3.3: Injury incidence rates by gender, age-group, and age-group-gender	26
 Article 1	
Figure 1: Districts of Thai Provinces bordering Cambodia (excluding Trad)	30
Figure 2: Plots of Standardized residuals against normal quantiles with zero cell counts omitted (left panel), further cells with low residuals omitted (middle panel), and imputed using the fitted model (right panel)	31
Figure 3: Plots of observed versus fitted counts and incidence rates (left panels) and residuals versus normal quantiles after fitting the GEE model	35
Figure 4: Confidence interval plots of annual incidence rates for each factor (quarter, year and district)	35
Figure 5: Schematic map of under-reporting (left panel) and annual diarrheal incidence rates (right panel) in districts of Thailand bordering Cambodia	36
 Manuscript 2	
Figure 1: Plot of two eigenvectors and corresponding basis functions (dotted)	46
Figure 2: Diagnostic plots for Poisson and negative binomial models	47
Figure 3: Plots of observed counts and observed incidence against fitted values	48
Figure 4: Annual TB incidence/1000 for males and females	49

	Page
Figure 5: Schematic map of annual tuberculosis incidence rates for males and females	50
Figure 6: Plots of component coefficients and trends of extreme super-districts for males	51
 Manuscript 3	
Figure 1: Plots of observed counts against fitted values (left) and deviance residuals versus normal quantiles (right)	62
Figure 2: Injury-related hospitalisation incidence rate/100,000 by super LGA (top panel), and sex and age group (bottom left panel), and year (right bottom panel), each adjusted for the effects of the other terms in the model	63
Figure 3: Thematic maps of adjusted annual injury-related hospitalisation rates in NSW with the insert of Sydney metropolitan areas. The upper panel shows rates for super-LGAs based on confidence intervals plotted in the upper panel of Figure 2, and the lower panel shows rates for LGAs based on an alternative Bayesian method	64
Figure 4.1: Results of fitting logistic regression model	73

Chapter 1

Introduction

1.1 Rationale for study

The incidence rate of an adverse outcome such as disease or injury is the number of new cases per unit population at risk in a specific period of time. In epidemiology, it can be used to compare the relative risks of the outcome for different factors such as age groups, regions, periods of time, exposures to occupations hazards, and other demographic and health status determinants.

Statistical models of varying complexity are used for analyzing incidence rates. Since the incidence rates in the cells comprising the sample are ratios of non-negative valued integer counts (the number of new cases observed) and population denominators that are regarded as fixed for purposes of statistical analysis, a Poisson generalized linear model (see, for example, Venables and Ripley 2002, Chapter 7) is frequently assumed as the basic model.

However, many studies (see, for example, Maul et al 1991, Lambert and Roeder 1995, Jansakul and Hinde 2004, Kaewsompak et al 2005, Paul and Saha 2007, Lim and Choonpradub 2007, Sriwattanapongse et al 2008, Sriwattanapongse and Kuning 2009, Kongchouy et al 2010) have shown that the Poisson distribution often does not fit incidence data in practice because it assumes that the variance is equal to the mean, and in many situations the variance is substantially greater than the mean. Thus the standard negative binomial generalized linear model (Venables and Ripley 2002

pages 206-208), for which the variance to mean (λ) ratio takes the form $1 + \lambda / \theta$, where the Poisson distribution arises in the limit as θ tends to infinity so the over-dispersion parameter is actually $1/\theta$, is usually preferred for analyzing incidence rates. This over-dispersion is often the result of clustering (see, for example, Demidenko 2007).

Another feature of incidence rates is that zero cell counts occur very frequently in practice, particularly for rare outcomes, and as a result statisticians have invented more complex *zero-inflated* models to account for this preponderance of zeroes (see, for example, Ridout et al 2001, Cheung 2002, Poston and McKibben 2003, Lewsey and Thomson 2004 and Ugarte et al 2004). However, Warton (2005) showed that data with many zeros does not necessarily mean zero inflation, on the grounds that the model itself can reduce the effect of the zero-inflation on the distribution assumed by the model.

Another problem that arises with incidence rates in practice is the presence of both time series and geospatial correlations. For linear regression models with normally distributed errors, these correlations can be handled using methods such as the generalized estimating equations (GEE) approach of Zeger and Liang (1986), but such methods have not yet been fully extended to generalized linear models (GLMs), although Yan and Fine (2004), Evans and Li (2005), Dormann (2007) and Faraway (2006) have developed methods for specified distributions. Other popular methods for handling spatial correlation in incidence rates are reviewed by Dormann et al (2007) and a Bayesian methodology described in Lawson et al (2003) is also widely used.

Generalized linear models have been further extended to handle incidence rates where the effects of predictors are modeled as unspecified smooth functions rather than as fixed functions, and the resulting models are called *generalized additive models* (see, for example, Hastie and Tibshirani 1990, Thurston et al 2000).

Although these and other complex statistical models are now the preferred methods used by biostatisticians for analyzing incidence rates, their advantages are offset by (a) the difficulties of understanding and correctly applying the methods experienced by scientists who lack an adequate knowledge of statistical theory, (b) the lack of availability of software packages and the associated difficulties in using these packages where they exist, and (c) the possible higher risk of bias associated with the use of complex models rather than simpler stratified analyses (Greenland 1989).

These considerations justify an investigation of the possibility of using simpler methods based on models for transformed incidence rates with normally distributed errors, and this is the main focus of the thesis. Some part of this thesis reports on the comparison of these simpler methods with those based on more complex generalized linear models, using data from three studies. In the first study we apply the methods to child diarrhea incidence rates in Thai provinces bordering Cambodia, where probable under-reporting of hospital cases is an important issue that needs to be considered. In the second study we investigate the spatial and temporal patterns of male and female tuberculosis incidence rates in districts of Nepal, while the third study is concerned with quarterly hospital admission rates for injuries to children aged under 15 year in local government areas of New South Wales (NSW), Australia.

1.2 Literature review

This review covers selected samples of relevant material from the statistical literature classified by topic (modeling over-dispersion, log-linear models, zero-inflated models, handling correlated data, generalized additive models, and Gaussian models for log-transformed incidence rates) in chronological order.

Modeling over-dispersion

Although the Poisson distribution continues to hold a central place in the analysis of incidence rates, Lawless (1987) was one of the first to investigate in detail extensions of the Poisson regression model that take account of extra-Poisson variation. This paper investigated negative binomial regression models and the efficiency and robustness properties of their properties.

Maul et al (1991) applied both the Poisson model and the more general negative binomial model to analyse quantal assays such as the toxic effect of sodium bromide on reproduction of the plankton species *daphnia magna*.

Lambert and Roeder (1995) introduced a convexity plot for graphing over-dispersion and relative variance curves, and developed relative variance tests that help to understand the nature of the data. In this paper they claimed that their convexity plots are superior to the score tests commonly used to detect over-dispersion.

Lee (1996) analyzed over-dispersed paired count data for comparing two treatments, using specific Poisson, mixed, and semi-parametric models. The conclusion from this investigation was that the semi-parametric model gives larger standard errors than the other two models.

Jansakul and Hinde (2004) used negative binomial models to analyze the number of embryos from an orange tissue culture experiment. They found that the negative binomial regression model with a cubic response functions over the dose levels was consistent with these data. They considered both the standard form of the negative binomial model with variance-mean ratio of the form $1 + \lambda / \theta$ and an alternative model with ratio $1 + \alpha$ for $\alpha \geq 0$ that requires use of the Newton-Raphson algorithm to obtain maximum likelihood estimates.

Kaewsompak et al (2005) studied epidemic patterns of dengue hemorrhagic fever and other acute febrile illnesses in Yala province in southern Thailand. They used Poisson and negative binomial distribution models to investigate relations between the incidence rates in terms of geographical patterns. They then developed a methodology that may be applied routinely to geographical epidemiologic research for the spatio-temporal mapping of disease. Schematic range maps and statistical models were used to investigate their distribution by year and location.

Log-linear models

Traditional statistical log-linear models describe patterns in contingency tables of cross-classified counts and are described in detail by Bishop et al (1975) and by Fienberg (1980). Since the population denominator is not taken into account in these models they are not directly appropriate for the analysis of incidence rates with variable populations at risk. Moreover, because these log-linear models focus on the associations between factors without necessarily focusing on the dependence of the outcome on its determinants, alternative generalized linear models based on the multinomial distribution (Venables and Ripley 2002, pages 199-205) are more

appropriate in many situations. However, log-linear models can still be used to examine risk factors for adverse outcomes.

For example, Tiensuwan et al (2000) used a log-linear model to identify the risk factors causing malaria in Tak province in northern Thailand in the rainy season for a case study of 1067 malaria patients, from whom 12 variables were recorded. They concluded that lack of knowledge of prevention was the main risk factor for disease in this population. Tiensuwan et al (2005) subsequently used a similar method to analyse risk factors for cancer incidence among patients admitted to National Cancer Institute in Thailand.

Zero-inflated models

Cheung (2002) investigated the impact of foetal growth and postnatal somatic growth on the ability of children aged 22 months to build a tower with three cubes, using data from a birth cohort of 16,955 in Britain. A negative binomial distribution fitted to the numbers of cubes built in cells with demographic determinants was fitted by a negative binomial distribution but found to contain an excessive proportion of zeros. However, the zero-inflated negative binomial distributed provided a satisfactory fit.

Poston and McKibben (2003) analyzed the average number of children ever born to women in the US, concluding that zero-inflated Poisson and negative binomial regression models are statistically appropriate for modeling fertility in low fertility populations, especially when there are many women in the society with no children.

In a cohort study of four data sets with dental caries outcomes in New Zealand, Lewsey and Thomson (2004) also used zero-inflated Poisson and negative binomial

regression models. The zero-inflated Poisson model was a poor fit for all four data sets, whereas the zero-inflated negative binomial model fitted well in each case.

In a study of regional patterns of brain cancer deaths in the Navarra region of Spain, Ugarte et al (2004) used a Poisson model to estimate the relative risks in different areas of the region, and also found justification for a zero-inflated Poisson model.

Despite the popularity and claimed benefits of zero-inflated regression model, Warton (2005) provided evidence that their appropriateness is over-rated. The negative binomial model was the best fitting of the count distributions, without zero-inflation, and Gaussian models based on log-transformed were found to fit surprisingly well to both simulated data and data from ecological studies.

Handling correlated data

Generalized Estimating Equations (GEE) have been developed to extend generalized linear model to accommodate correlated data.

Yan and Fine (2004) investigated GEEs for estimating association parameters, using data from a study on the genetics of alcoholism to illustrate the importance of reliable variance estimation in biomedical applications.

Demidenko (2007) did a simulation study to compare five methods for parameter estimation of a Poisson regression model for clustered data, including Poisson regression with and without fixed cluster-specific intercepts, GEE with exchangeable correlation structure, GEE with an exact covariance matrix, and maximum likelihood. All five methods gave consistent estimates of slopes but different efficiencies, and the conclusion was that both the simple Poisson and GEE methods were outperformed by the three alternatives, with exact GEE recommended for its simplicity.

Dormann et al (2007) compared various methods for taking account of spatial autocorrelation. The preliminary tests confirmed that most of the spatial modeling techniques give a good type I error control and precise parameter estimates. They also found that autocovariate methods consistently underestimate the effects of environmental controls of species distributions, and that the Bayesian approaches developed by Besag et al (1991) and advocated in Lawson et al (2003) are computationally intensive and of questionable benefit in comparison with more straightforward non-Bayesian methods.

Davis et al (2003) proposed lagged observation-driven models for Poisson counts that take account of time series autocorrelations. However, in their study of regional and temporal patterns of infectious disease mortality in provinces of southern Thailand over the period 1999-2004, Lim and Choonpradub (2007) found that these models are numerically unstable because the lagged terms arise as exponential functions in the mean, and need to be log-transformed to achieve stationarity.

Sriwattanapongse et al (2008) also used observation-driven negative binomial regression models to forecast monthly incidence rates of hospital-diagnosed malaria by district and age-group in two North-western border provinces of Thailand.

Generalized additive models

Thurston et al (2000) applied the generalized additive model is extended to handle negative binomial responses to analyze data involving DNA adducts counts and smoking variables among ex-smokers with lung cancer. This study included a detailed investigation of the parametric relationship between the number of addicts and years

since quitting while retaining a smooth relationship between addicts and the other covariates in the model.

Gaussian models for log-transformed incidence rates

Despite the complexity of generalized linear models and their extensions, Paul and Saha (2007) questioned their suitability for analyzing over-dispersed data, claiming that in many real life applications the distributional assumptions of such models cannot be justified.

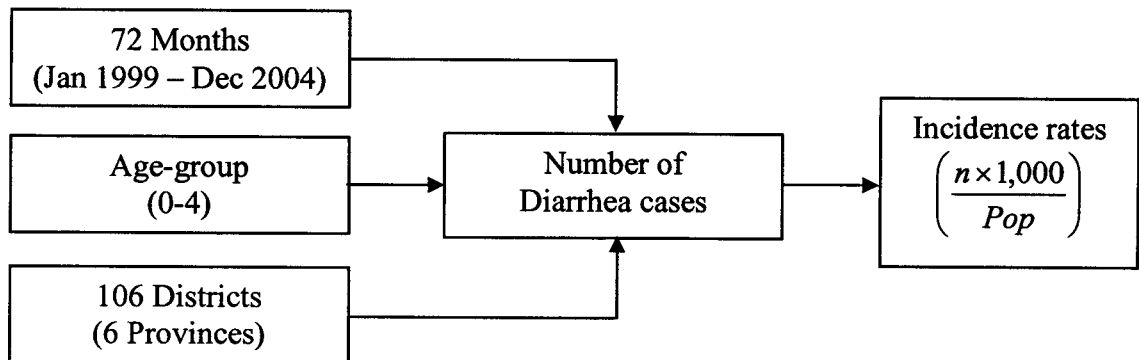
Sriwattanapongse and Kuning (2009) overcame lack of fit problems by including multiplicative interactions based on principal components of residuals from linear models to analyse the patterns of hospital diagnosed malaria incidences in districts and quarterly periods in the North-western region of Thailand in 1999-2004.

For modeling pneumonia incidence rates among children under 5 in districts of SuratThani province in southern Thailand, Kongchouy et al (2010) compared the negative binomial generalized linear model with a log transformed linear model after replacing zeros by a constant between 0 and 1. They suggested that the standard negative binomial models fail to cover the range of over-dispersion situations that commonly occur in practice, particularly for biological data.

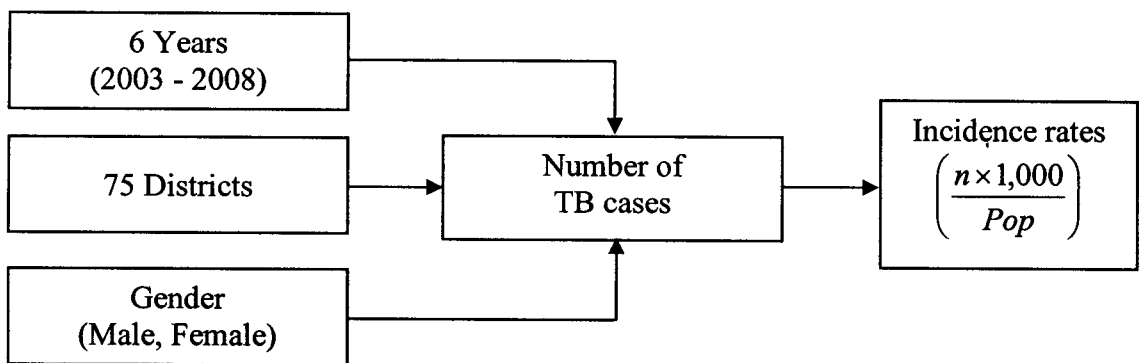
1.3 The studies

Although the data for our studies arose from three quite different sources, the outcome variable was essentially the same in each case, namely an incidence rate of an adverse event based on cells classified by demographic variables including gender and/or age group, location, and period of time (month, quarter or year). The path diagrams for the three studies are shown in Figure 1.1.

Study 1: Child diarrhea in Thai Provinces Bordering Cambodia: 1999-2004



Study 2: Tuberculosis (TB) in Nepal: 2003-2008



Study 3: Injuries in NSW, Australia: July 2000- June 2005

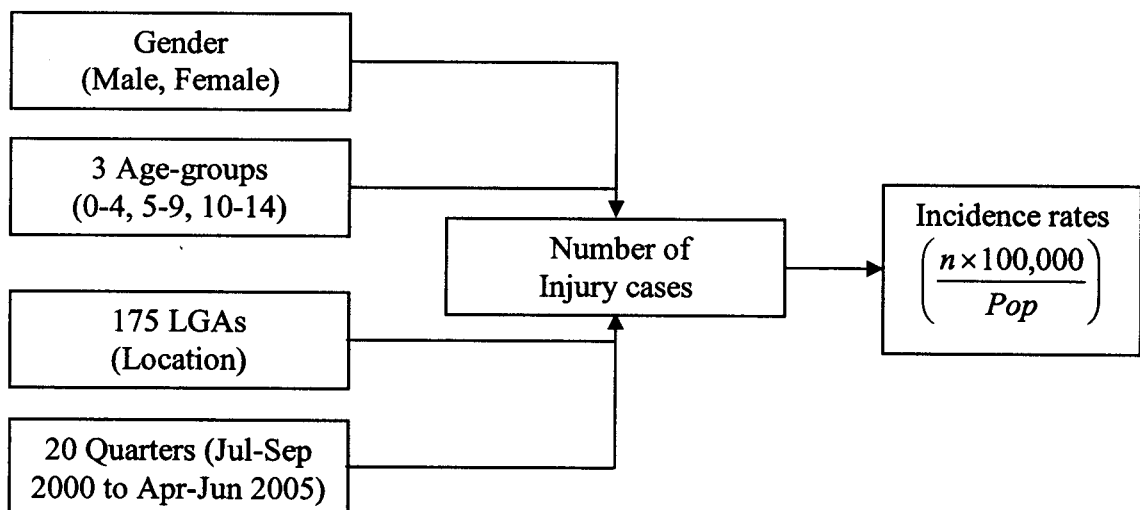


Figure 1.1: Path diagram for study

1.4 Data collection

The first study is the number of diarrhea cases from 1999 to 2004 in provinces of Thailand bordering Cambodia, based on a registry of hospital-diagnosed infectious disease cases collected routinely in each of Thailand's 76 provinces. These data are compiled from surveillance forms by the Ministry of Public Health. Nearly all subjects were Thai citizens, but all were treated equally in the study. Most cases were suspected but not lab-confirmed. The denominator used as the computed incidence rate was obtained from the Population and Housing Census of 2000 which was undertaken by the National Statistics Office of Thailand. The data were registered by the village of residence, the day of hospitalization of the patient and districts. The use of districts rather than smaller regions such as villages or sub-districts can substantially eliminate correlation between annual incidences outcomes in successive periods of time and neighbouring locations, while still enabling trends in both place and time to be identified.

The second study contains the cases of TB reported by the National Tuberculosis Control Program (NTC) in Nepal. The information used, regarding cases notified between 2003 to 2008, was provided by STAC (SAARC Tuberculosis and HIV/AIDS centre), the specific NTC information system managed by the NTC coordination team, working for prevention and control of TB and HIV/AIDS in the Region. The reported cases for each year were available in computer files comprising characteristics of the disease, gender, address, and the severity of the illness.

The third study was based on all hospital separations of New South Wales (NSW) residents aged below 15 years who were unintentionally injured from 1 July 2000 to

30 June 2005. This data include information on inpatient separations of NSW residents from public and private hospitals, private day procedures, and public psychiatric hospitals. They include data on episodes of care in hospital, which end with the discharge, transfer, or death of the patient, or when the service category for the admitted patient changes. The hospitalisation data were coded using ICD-10-AM (National Centre for Classification in Health 1998, National Centre for Classification in Health 2000, 2002, 2004).

1.5 Objectives and plan of thesis

Appropriate statistical models are used to model incidence rates. These models attempted to identify the associations between demographic factors (location, season, age, and gender) and longitudinal data outcomes (incidence rates), so linear regression, Poisson regression, and negative binomial regression models were applied to fit these data.

The objectives of studies were thus as follows.

1. To develop statistical methods for modeling incidence rates.
2. To investigate the epidemic patterns between demographic factors and outcomes variable (incidence rates).

This thesis contains four chapters. The introductory chapter discusses the rationale, the scope and the aim of the study, and also includes a review of some relevant literature.

Chapter 2 provides a description of the methodology including an overview of the statistical methods for data analysis aligned to the statistical models.

Chapter 3 shows original article and manuscripts that were written for child diarrhea in Thailand, TB incidence rates in Nepal and children's injury incidence rates in NSW, Australia.

The last chapter states the summaries and general conclusions. Suggestions for further research are also provided in this chapter.

Chapter 2

Methodology

This chapter describes the methodology including an overview of the statistical methods for data analysis aligned to the statistical models. Graphical and statistical analyses were carried out using R program (R Development Core Team 2008).

This presents statistical methods adopted in the three papers contained in Chapter 3. These methods include Poisson regression, negative binomial regression, log-transformed linear regression modeling, and adjustment for correlated residuals using the generalized estimating equations method. Statistical procedures in these methods are described and highlighted with respect to statistical models and relevant assumptions.

2.1 Data management

Infectious disease cases from the Ministry of Public Health are kept in files comprising individual records. The records contain many errors and omissions, which were corrected using purpose-written SQL programs in the SQL Server database system, which was also used to create tables of disease count data aggregated by the demographic risk determinants. Data were thus converted to a flat-file format for calculating descriptive statistics and modeling.

The tuberculosis data from Nepal were obtained as aggregated lists in computer files from the National Tuberculosis Control Program authorities and entered into computer text files suitable for data cleaning and analysis.

Childhood injury data were processed in the Injury Risk Management Research Centre at the University of NSW by Dr Shanley Chong, who used R programs developed by the author to analyse these data and produce the graphs and maps.

2.2 Poisson regression model

Poisson regression is appropriate for fitting models with count data (non-negative integer-values). A random variable Y is said to have a Poisson distribution with parameter $\lambda > 0$ if it takes integer values $y = 0, 1, 2, \dots$ with probabilities

$$\text{Prob}(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}. \quad (2.1)$$

The mean and variance of this distribution can be shown to be

$$E(Y) = \text{var}(Y) = \lambda. \quad (2.2)$$

Since the mean is equal to the variance, any factor that affects one will also affect the other.

Poisson regression model can be fitted by using the generalized linear models method with the log link function (McCullagh and Nelder 1989).

Poisson regression is commonly used for modeling the number of cases of disease in a specific population within a certain time period. Suppose that n_{ijt} is the number of observed cases in cells defined by demographic group (gender and/or age group) i , geographical location j and period of time t and P_{ij} is the corresponding population at risk. If λ_{ijt} denotes the mean incidence rate, an additive model with this distribution is expressed as

$$\ln(\lambda_{ijt}) = \ln(P_{ij}) + \mu + \alpha_i + \beta_j + \gamma_t. \quad (2.3)$$

The terms α_i , β_j and γ_t represent demographic group, location and period of time effects which sum to zero so that μ is a constant encapsulating the overall incidence.

Adjusted Incidence Rates

After fitting the model, adjusted incidence rates for each factor of interest are obtained by suppressing the subscripts in Equation (2.3) corresponding to the other factors and replacing these terms with a constant satisfying the condition that the sum of the counts based on the adjusted incidence rates matches the total (Swennen et al 2009). For Poisson regression, this is achieved simply by multiplying the incidence rates for the specified factor of interest by a scale constant specific to the factor.

Sum Contrasts

Sum contrasts (Venables and Ripley 2002, Tongkumchum and McNeil 2009) are used to obtain confidence intervals for comparing adjusted incidence rates within each factor with the overall incidence rate. An advantage of these confidence intervals is that they provide a simple criterion for classifying levels of a factor into three groups according to whether each corresponding confidence interval exceeds, crosses, or is below the overall mean.

Methods for creating geographical maps

A thematic map is a type of map that uses different colours or shades to graphically display information about the underlying data representing estimated values of a variable at different locations on the map. The thematic map using data in regions might show one region in dark red to indicate that the region has high values, while showing another region in very pale red to indicate that the region has low values. A range map is a type of thematic map that displays data according to ranges set by the

users. The ranges are shaded using colors or patterns. These types of maps are used to show the geographical distribution of the adverse outcome and to identify areas of high risk. Appropriate graphs are used for exploratory data analysis, visualizing the pattern of the data and highlighting possible errors in the data that could cause problems in further analysis.

Since the confidence intervals for factor-specific incidence rates obtained from a 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 the incidence rates.

Over-dispersion

After fitting a generalized linear model to the data, to check the adequacy of the respective model, one usually computes a residual deviance for each cell. Thus, the deviance statistic for an observation reflects its contribution to the overall goodness of fit of the model. Plotting these residual deviances against corresponding quantiles for the normal distribution gives an indication of the adequacy of the fit of the model to the data. If the plot is approximately linear with unit slope, the fit is satisfactory.

Details are given in Chapter 7 of Venables and Ripley (2002).

A characteristic of the Poisson distribution is that its mean is equal to its variance. If the observed variance is greater than the mean the data are over-dispersed and the residual deviance plot will indicate that the model is not appropriate. Common reasons for over-dispersion are clustering of disease cases and the omission of relevant explanatory variables.

2.3 Negative binomial regression model

A problem with the Poisson regression model occurs when we encounter over-dispersion.

A method of dealing with such over-dispersion is to use the more general negative binomial time series model instead of the simple Poisson model (see, for example, Cameron and Trivedi 1998, page 71).

The negative binomial model where y is the number of trials until number of success occur is defined in terms of the density function of Y as

$$\text{Prob}(Y = y) = \frac{\Gamma(y + \theta)}{\Gamma(y + 1)\Gamma(\theta)} \left(\frac{\theta}{\theta + \lambda} \right)^\theta \left(\frac{\lambda}{\theta + \lambda} \right)^y. \quad (2.4)$$

The Poisson model arises in the limit as this dispersion parameter $\theta \rightarrow \infty$, so the over-dispersion parameter is actually the reciprocal of θ . The expected value of Y is λ and its variance is $\lambda + \lambda^2/\theta$.

2.4 Log-transformed linear regression model

The conventional model for handling data where the outcome is continuous is linear regression.

Let Y be a log-normally distributed variable with characteristic parameters mean μ and variance σ^2 . This implies that the probability density function of Y is the density function of the normal distribution, namely,

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(z - \mu)^2}{2\sigma^2}\right]. \quad (2.5)$$

In our studies, the incidence rates generally have positively skewed distributions so it is conventional to transform them by taking logarithms to obtain the outcome as

$$y_{ijt} = \ln \left(1000 \times \frac{n_{ijt}}{P_{ij}} \right). \quad (2.6)$$

Thus an additive linear model was fitted to the logarithms of the log-transformed incidence rates, namely,

$$y_{ijt} = \mu + \alpha_i + \beta_j + \gamma_t. \quad (2.7)$$

As in Equation (2.3), μ is a constant and α_i , β_j and γ_t are demographic group, location and period of time effects, respectively, with zero means.

Handling zeroes

If any count n_{ijt} is zero, Equation (2.6) needs to be modified to give a finite result, so that n_{ijt} is replaced by a positive value n_{ijt}^* .

Various methods may be considered for this data modification. Zero counts simply could be omitted, and the fitted model then used to impute counts for these cases before refitting the model (Ardkeaw and Tongkumchum 2009). This method has advantages in situations where under-reporting is known or suspected. Another method involves adding a constant c to all counts so that $n_{ijt}^* = n_{ijt} + c$. A third method involves replacing the zeroes by a suitably chosen constant d without changing any values of n_{ijt} greater than 0 (Kongchouy et al 2010).

Multiplicative models

Whereas age and/or gender effects are likely to remain fixed over the duration of a study, region effects are more volatile, particularly over the course of an epidemic. To allow for such interactions between spatial and period effects the additive model (2.7) needs to be extended, so we consider models of the form

$$y_{ijt} = \alpha_i + \sum_{k=1}^m \beta_j^{(k)} \gamma_t^{(k)} \quad (2.8)$$

This model contains m multiplicative space-time interaction terms. Since this is a non-linear model, its parameters cannot be fitted using linear regression, but Theil (1983) showed that the least squares estimates of the $\gamma_t^{(k)}$ parameters are the elements of the eigenvector of the matrix $Y_c^T Y_c$ corresponding to its k^{th} largest eigenvalue, where Y_c has elements $y_{ijt} - \bar{y}_i$ and Y^T denotes the transpose of Y . The corresponding least squares estimates of the $\beta_j^{(k)}$ parameters are then expressed in terms of the eigenvectors $\gamma_t^{(k)}$ as

$$\beta_j^{(k)} = \sum_{t=1}^T \gamma_t^{(k)} (y_{ijt} - \bar{y}_{ij}) . \quad (2.9)$$

However, if the $\gamma_t^{(k)}$ parameters are regarded as fixed, the model can still be fitted using linear regression, giving both estimates and standard errors for the remaining parameters. In practice, this assumption would be reasonable if the $\gamma_t^{(k)}$ were replaced by basis functions $g_t^{(k)}$ such as orthogonal polynomials or spline functions of degree k . The model then may be written

$$y_{ijt} = \alpha_i + \sum_{k=1}^m \beta_j^{(k)} g_t^{(k)} . \quad (2.10)$$

For the case $m = 1$, this is a simple generalization of the Lee-Carter model (McNeil and Tukey 1973, Lee and Carter 1992), in turn extended to m components by Booth et al (2002).

Generalized estimating equations (GEE) model

GEE is the extended generalized linear model to handle the correlated data (Zeger and Liang 1986, Yan and Fine 2004). It represents a class of model that is often utilized for data in which the responses are correlated.

Even though the GEE method gives estimates and standard errors of parameters adjusted for the assumed spatial correlations, the residuals themselves remain correlated. But since the correlation structure of the errors is either assumed or can be estimated, their correlation can be removed simply by using a linear filter (Ardkeaw and Tongkumchum 2009), and it may be advisable to remove this correlation before plotting them to assess the adequacy of the model.

2.5 Logistic Regression model

The logistic regression is used for model fitting in which the outcome variable is binary. In our first study investigating under-reporting, the unreported/reported status of a disease case is the outcome, so logistic regression is the appropriate method. In this case p_{ijt} denotes the proportion of unreported cases in district i , quarter j and year t . The logistic model takes the form

$$\ln\left(\frac{p_{ijt}}{1-p_{ijt}}\right) = \mu + \alpha_i + \beta_j + \gamma_t. \quad (2.11)$$

Equation 2.11 can be inverted to give an expression for the probability of the event as

$$p_{ijt} = \frac{1}{1 + \exp(-\mu - \alpha_i - \beta_j - \gamma_t)}. \quad (2.12)$$

The functional form of the right-hand side of Equation 2.12 ensures that its values are always between 0 and 1, as they should be, given that they are probabilities.

Chapter 3

Modeling Incidence Rates with Applications

3.1 Studies completed

We applied statistical methods for incidence rates to three studies. The first study examined patterns of diarrhea incidence in children less than 5 years of age in Thai provinces bordering Cambodia with the exception of Trad province. Zero or low count cases occurred in some districts where under-reporting was known or suspected. These counts were replaced by imputed values before fitting the model. The log-transformed linear regression model was then used to investigate the patterns of diarrhea incidence for district, quarter and year, and the GEE method was used to adjust for spatial correlation for residuals. The manuscript has been accepted for publication in *the Southeast Asian Journal of Tropical Medicine and Public Health* and appears in Volume 41 No.1 January 2010.

The second study aimed to model the trends in the annual tuberculosis incidence rates by gender in districts of Nepal from 2003 to 2008. This method investigates the regional and temporal pattern of this disease using the additive model given by Equation (2.3) with the negative binomial distribution to account for over-dispersion, but found that this model still has excessive deviance. Since the incidence rate for tuberculosis depends strongly on gender and in Nepal this gender effect varies over districts, the multiplicative model (2.8) was used with the additive effect for gender replaced by a factor combining gender and district, and we obtained a satisfactory fit with two multiplicative district-year components. The manuscript comprising the

second section of this chapter has been submitted to the *Asian Biomedicine (Research Review and news) Journal*.

Our third study involved an analysis of quarterly injury rates from July 2000 to June 2005 among boys and girls under 15 classified by 5-year age groups in 104 local government areas in NSW. The aim of this paper was to show that the simpler method based on a negative binomial generalized linear model is preferable to the complex and computationally expensive Bayesian method used by the NSW Department of Health in previous studies. This paper is currently being reviewed by the Department prior to submission to an appropriate international journal.

3.2 Preliminary analysis

Study 1: Diarrhea in Thai Provinces Bordering Cambodia: 1999-2004

This study focused on the monthly periods data of the number of under-reported cases for each district.

Figure 3.1 shows the number of monthly reported diarrhea cases in some selected districts. This graph illustrates the extent of probable under-reporting under our suspicion that the data are under reported. In top three panels, there is no strong evidence of under-reporting. However, in the other six districts there are noticeable gaps in the time series of reported cases which provide strong evidence of under-reporting.

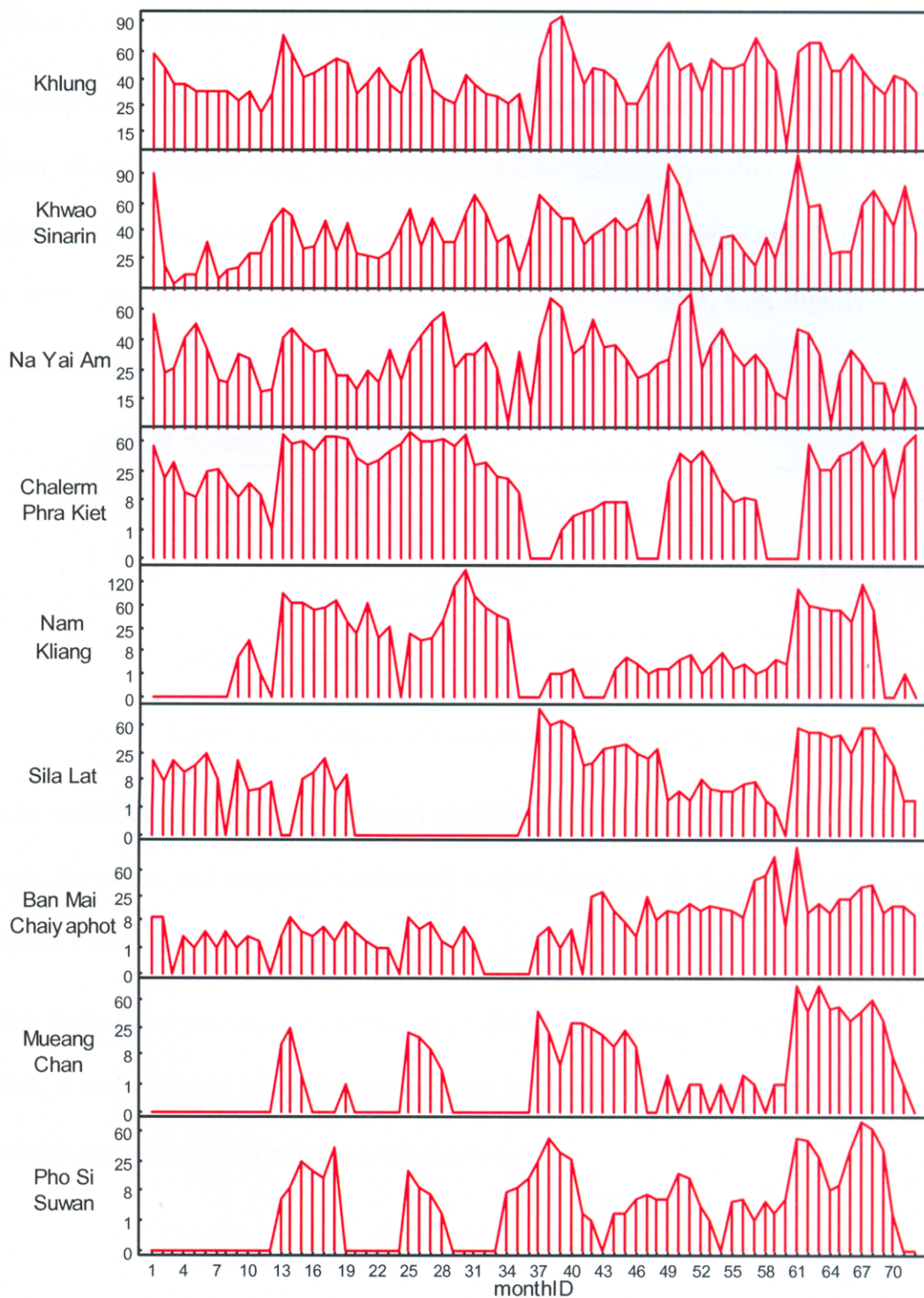


Figure 3.1: Numbers of monthly reported diarrhea cases in selected districts

Study 2: Tuberculosis (TB) in Nepal: 2003-2008

Figure 3.2 shows 95% confidence intervals for annual incidence rate per 1,000 by year. These are crude rates, unadjusted for gender and region within Nepal. The dotted horizontal lines represent the mean of annual TB incidence rate (1.14 per 1,000). The highest incidence rates occurred in 2005 and 2004, with slightly lower rates in 2006 and 2007, followed by a steep drop in 2008.

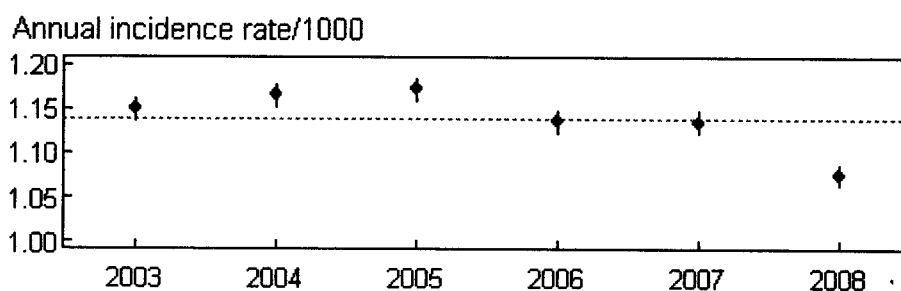


Figure 3.2: Tuberculosis incidence rates by year in Nepal

The preliminary analysis also showed a gender difference, with substantially higher rates for males, and regional differences, with higher rates in districts located in the lower altitudes.

The further analysis involved statistical modeling containing effects for gender, district and year, and it was found necessary to include interactions between these factors, as well as substantial over-dispersion.

Study 3: Injuries in NSW, Australia: July 2000- June 2005

Figure 3.3 presents 95% confidence intervals of annual injury-related hospitalization incidence rates per 100,000 for gender and age group (left panel) and six gender-age groups (right panel). These graphs show a strong interaction between gender and age group: rates for boys increase with age (particularly for 10-14 year-olds), but decrease for girls. An earlier study had fitted a complex Bayesian model using the BRugs package (Thomas 2004) to these data but had not taken this age-gender interaction into account. This model corrects for spatial correlation between pairs of districts with common boundaries. These districts comprised the 175 local government areas (LGAs) of NSW.

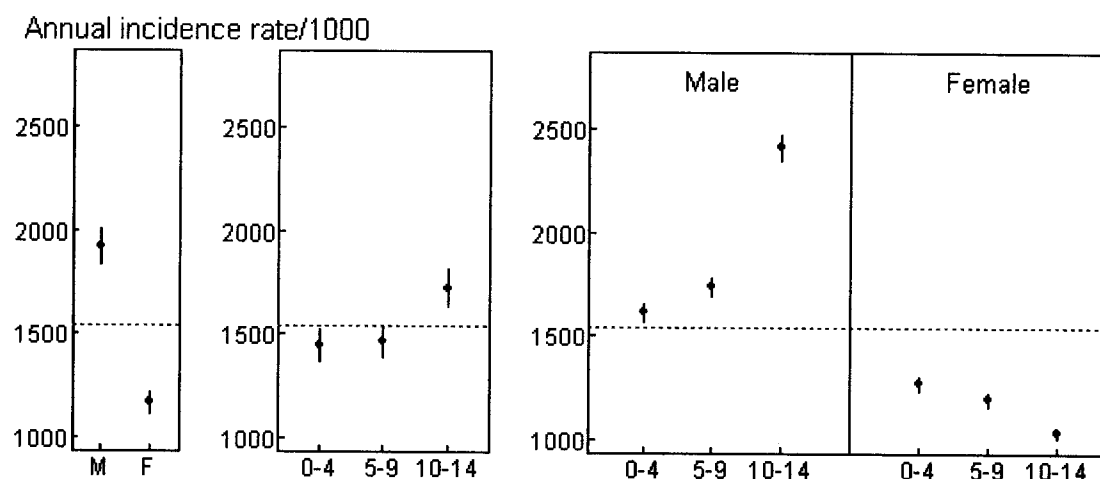


Figure 3.3: Injury incidence rates by gender, age-group, and age-group-gender

These data thus provided an opportunity to fit a simpler linear model to the log-transformed incidence rates with just two factors (age-gender group and district), and the GEE method to account for spatial correlation. Since the LGA populations range from a less than 1000 to more than 300,000, we aggregated LGAs with smaller populations into 104 super-LGAs, and assumed fixed interchangeable correlation matrices within each of the 17 larger regions.

3.3 Article 1

A STATISTICAL METHOD FOR ESTIMATING UNDER-REPORTED INCIDENCE RATES WITH APPLICATION TO CHILD DIARRHEA IN THAI PROVINCES BORDERING CAMBODIA

Sulawan Yotthanoo¹ and Chamnein Choonpradub²

¹Department of Statistics, School of Science and Technology, Naresuan University, Phayao; ²Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani, Thailand

Abstract. Diarrhea is a major health problem in Thailand, but reported data of disease incidence are known or suspected to be under-reported. This study aimed to develop a statistical model for estimating the annual incidence of hospital diarrhea cases among children under five years. Data regarding diarrhea patients 0-4 years old were collected for the National Notifiable Disease Surveillance (Report 506) about Thai provinces bordering Cambodia during 1999-2004 by the Ministry of Public Health. A log-linear regression model based on the prevailing seasonal-trend pattern was used for diarrhea incidence as a function of quarter, year and district, after imputing rates where under-reporting was evident, using populations obtained from the 2000 population census. The model also takes any spatial correlation between districts into account, using the generalized estimating equation (GEE) method. Diarrhea incidence had seasonal peaks in the first quarter (January to March) and the trend steadily increased from 1999 to 2004. Results from such studies can help health authorities develop prevention policies.

INTRODUCTION

Diarrhea is one of the world's top five infectious disease causes of death (Brownlie *et al*, 2006) and remains a major cause of morbidity and mortality among children in developing countries (Carlos and Saniel, 1990; Parashar *et al*, 2003). Children under five years of age have an average of 3.3 diarrhea episodes per year, and more than one-

third of all deaths in this age group are associated with diarrhea. Approximately 1.5 billion diarrhea episodes and 4 million deaths occur annually among children age less than five years (Vargas *et al*, 2004).

As in other developing countries, diarrhea in Thailand is a major health problem and accounts for approximately 50% of all hospital-reported infectious diseases (Thai Working Group on Burden of Disease and Injuries, 2002). The Bureau of Epidemiology (2002) reported that while diarrhea-related mortality declined from 1.11 per 100,000 in 1988 to 0.23 per 100,000 in 2002, morbidity increased from 1,488 cases per 100,000 in 1993 to 1,687 cases per 100,000 in 2002. In 2002, there were 1,055,393 cases of diarrhea in

Correspondence: Chamnein Choonpradub, Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani 94000, Thailand.

Tel: 66 (0) 894660803; Fax: 66 (0) 7331 2729

E-mail: cchamnein@bunga.pn.psu.ac.th

Thailand, of which one-third occurred among children under five years of age and 12% required hospitalization (Jiraphongsa *et al*, 2005).

Given that diarrhea morbidity remains high in Thailand, there is a need to improve treatment to prevent the disease, especially in provinces bordering Cambodia where cross-border migration may be a factor (Thimasarn *et al*, 1995; Konchom *et al*, 2003).

We investigated a statistical model based on linear regression for estimating the extent of under-reporting. We then classified patterns of child diarrhea in Thai provinces bordering Cambodia using the GEE model.

MATERIALS AND METHODS

Study area and data source

The border between Thailand and Cambodia is approximately 800 km long, stretching along the provinces of the lower north-east area of Thailand from a point known as "Chong Bog" in Ubon Ratchathani Province and ending in the Had Lek Subdistrict of Klong Yai in Trat Province (Fig 1). Due to its stronger economic growth, Thailand attracts many migrant workers from Cambodia. Cross-border migration has been connected with health problems, including infectious diseases (Thimasarn *et al*, 1995; Konchom *et al*, 2003).

Data regarding diarrhea cases from 1999 to 2004 in the border provinces of interest were taken from the National Notifiable Disease Surveillance Report (506), Bureau of Epidemiology, Ministry of Public Health. Each record contains the type of infectious disease, age, gender, subdistrict of residence, date of hospitalization, and disease severity of the patient. The resident population denominator used to compute the annual incidence was obtained from the Population and Housing Census of 2000, performed by the National Statistics Office of Thailand.

Data analysis

Although the registry included the village of residence and date of hospitalization of the patient, we used districts (statistical regions containing up to hundreds of villages ranging in population from 795 to 21,409). We did this to substantially reduce correlations between annual incidence outcomes in successive periods of time and in neighboring locations, while still enabling trends for place and time to be identified. Data from the Thai infectious disease registry are known, or suspected, to be seriously under-reported (Lumbiganon *et al*, 1990; Saengwonloey *et al*, 2003; Intusoma *et al*, 2008).

We first calculated disease incidence in cells defined by district (i) and month (j) of year (t) as the ratio of the number of reported cases (n_{ijt}) to the district population in 1,000s (P_i). For reasons evident from a detailed study of monthly disease counts (Table 2), any occurrence of zero cases in a cell was considered as a possible instance of under-reporting, and an additive linear model was fitted to the logarithms of the remaining incidence rates, namely,

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

In this model μ is a constant and α_i , β_j and γ_t are the effects of district $i=1,2,\dots,106$, month $j=1,2,\dots,12$ and year $t=1,2,\dots,6$, respectively, with zero means. After examining the plot of standardized residuals, this model was refitted after further omission of cells corresponding to residuals below a specified cut-off value. Having thus obtained an acceptable fit, the omitted occurrences were then imputed using the model. Next, we aggregated the monthly data into quarterly incidence rates and fitted a model similar to (1) with j now representing quarter instead of month, and the residuals from this model

were used to compute correlation coefficients between different districts. The averages of the correlation coefficients within each province were then used to fit a generalized estimating equation (GEE) model (Liang and Zeger, 1986; Yan and Fine, 2004) having a fixed block-diagonal correlation structure with the blocks corresponding to provinces. Residuals from this model were examined using normal quantile plots after filtering to remove their estimated correlation.

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

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

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 using sum contrasts (Venables and Ripley 2002) to compare the incidence rates for each level of a factor with the overall mean incidence rate.

Since the confidence intervals for factor-specific incidence rates and proportions obtained from this model (using the sum contrasts) 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 schematic maps of districts according to their estimated diarrhea annual incidence rates and under-reporting percentages.

We also estimated the extent of under-reporting data by district, quarter and year, by fitting a simple logistic regression model (Hosmer and Lemeshow, 2000; Kleinbaum and Klein, 2002) to the corresponding proportions imputed using the method described

above. These estimated proportions for levels of each factor after adjusting for the other factors in the model were again computed by requiring the weighted average of the adjusted proportions match the overall proportion, using a Newton-Raphson iteration procedure with Marquardt damping, and standard errors for differences between individual proportions and the overall mean.

All statistical analysis was carried out using the R program (R Development Core Team, 2007).

RESULTS

Preliminary analysis

Preliminary analysis indicated gender was not a major factor influencing diarrhea incidence (Ardkeaw and Tongkumchum, 2009), therefore it was not included in the model. Children less than five years old had the highest age-specific annual incidence rate (Bureau of Epidemiology, 2007), so this group was selected for study.

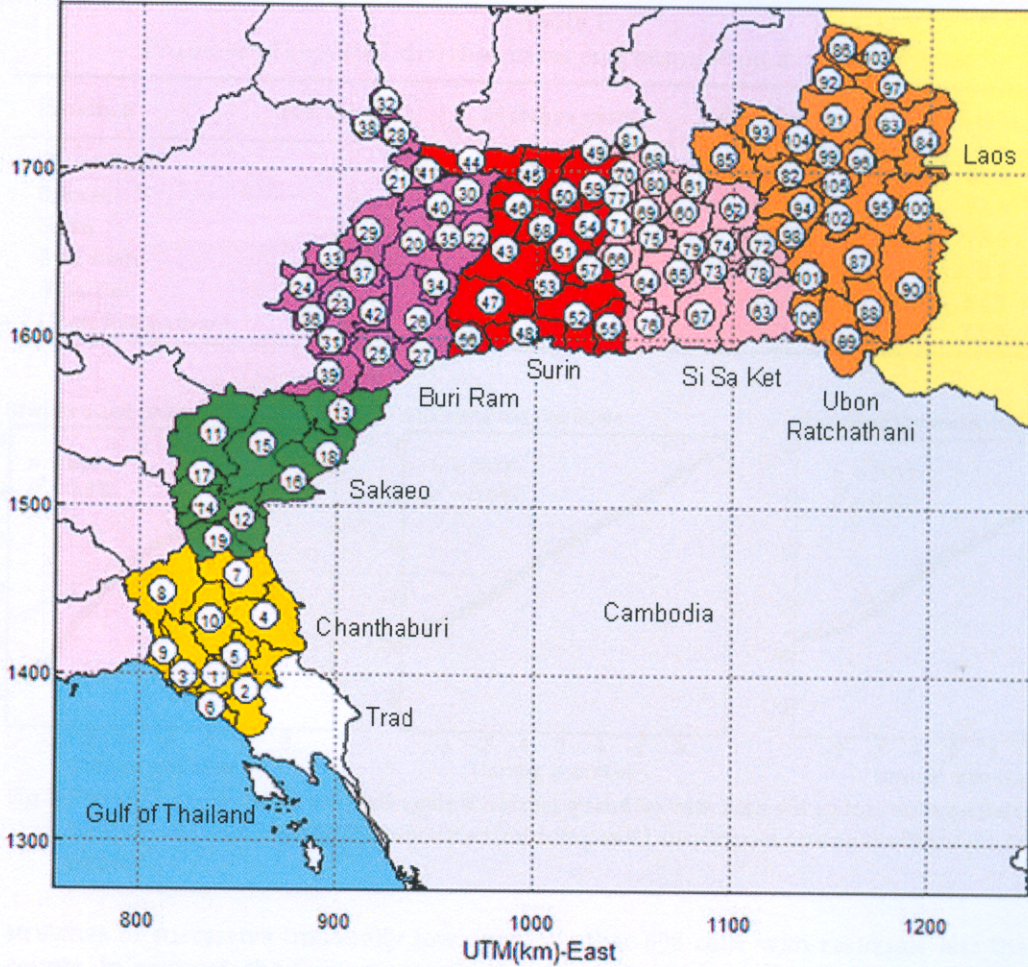
Data from Trat Province were not available for 2002 and so were excluded from the analysis.

During the study period from January 1999 to December 2004, 260,522 cases of diarrhea were reported from district hospitals in the Thai-Cambodia border provinces among children less than 5 years old. The number of cases reported in a month for each district varied from zero to 578 (average annual incidence rate 69.4 per 1,000). Among the six provinces, Surin (78.8 per 1,000) and Buri Ram (78.2 per 1,000) had relatively high rates (Table 1).

Under-reporting

Table 2 shows the number of monthly reported diarrhea cases in ten selected districts. Six of these districts Pho Si Suwan, Mueang Chan, Ban Mai Chaiyaphot, Sila Lat, Nam Kliang and Chalermphrakiet, had

UTM(km)-North



- | | | | | |
|-----------------------|-----------------------|--------------------|---------------------------|----------------------|
| 1 Mueang Chantchaburi | 23 Nang Rong | 45 Tha Tum | 67 Khun Han | 89 Nam Yuen |
| 2 Khlung | 24 Nong Ki | 46 Chom Phra | 68 Rasi Sabai | 90 Buntharik |
| 3 Tha Mai | 25 Lahan Sai | 47 Prasat | 69 Uthumphon Phisai | 91 Trakan Phut Phon |
| 4 Bong Nam Ron | 26 Prakhon Chai | 48 Kap Choeng | 70 Bung Bun | 92 Kut Khaopun |
| 5 Makham | 27 Eon Kruat | 49 Rattana Buri | 71 Huai Thap Than | 93 Mueng Sam Sip |
| 6 Laem Sing | 28 Phu Thai Song | 50 Sanom | 72 Non Khun | 94 Warin Chamrap |
| 7 Soydow | 29 Lam Plai Mat | 51 Sikhoraphum | 73 Si Patana | 95 Phibu Mangsahan |
| 8 Kaeng Hang Mueo | 30 Satuek | 52 Sangkha | 74 Nam Kling | 96 Tan Sum |
| 9 Na Yai Am | 31 Pakham | 53 Lam Duan | 75 Weng Hin | 97 Pho Sai |
| 10 Nao Kichakut | 32 Na Pho | 54 Samrong Thap | 76 Phu Sing | 98 Samrong |
| 11 Mueang Sa Kaeo | 33 Nong Hong | 55 Buachet | 77 Mueng Chan | 99 Don Mot Daeng |
| 12 Khlong Hat | 34 Phlapphlachai | 56 Phanom Dong Rak | 78 Benchalak | 100 Sitrindhom |
| 13 Ta Phraya | 35 Huai Rat | 57 Si Naxong | 79 Phayu | 101 Thung Si Udom |
| 14 Weng Nam Yen | 36 Non Suwen | 58 Kwao Sinazin | 80 Pho Si Suwen | 102 Na Yai |
| 15 Wattana Kakhon | 37 Channi | 59 Non Narai | 81 Sila Lat | 103 Na Tan |
| 16 Aranyaprathet | 38 Eon Mai Chaiyaphot | 60 Mueng Si Sa Ket | 82 Mueng Ubon Ratchathani | 104 Lao Sua Kok |
| 17 Khao Chahan | 39 Non Din Daeng | 61 Yong Chum Noi | 83 Si Mueng Mai | 105 Sawang Weerawong |
| 18 Khok Sung | 40 Eon Dan | 62 Kanthararom | 84 Khong Chiam | |
| 19 Weng Sombun | 41 Khaen Dong | 63 Kantharalak | 85 Khuang Nai | |
| 20 Mueang Buri Ram | 42 Chalem Phra Kiet | 64 Khukhan | 86 Khemarat | |
| 21 Khu Muang | 43 Mueang Surin | 65 Phrai Bueng | 87 Det Udom | |
| 22 Krasang | 44 Chumphon Buri | 66 Prang Ku | 88 Na Chaluai | |

Fig 1-Districts of Thai provinces bordering Cambodia (excluding Trad).

A STATISTICAL METHOD FOR UNDER-REPORTED CHILD DIARRHEA

Table 1
Number of reported diarrhea cases and population in each province.

Province	No. districts	Diarrhea cases	Population	Annual Incidence Rate
Chantaburi	10	12,552	33,108	63.2
Sakaeo	9	18,849	42,560	73.8
Surin	23	64,574	136,634	78.8
Buri Ram	17	58,940	125,642	78.2
Si Sa Ket	22	42,798	130,875	54.5
Ubon Ratchathani	25	62,809	153,410	68.2

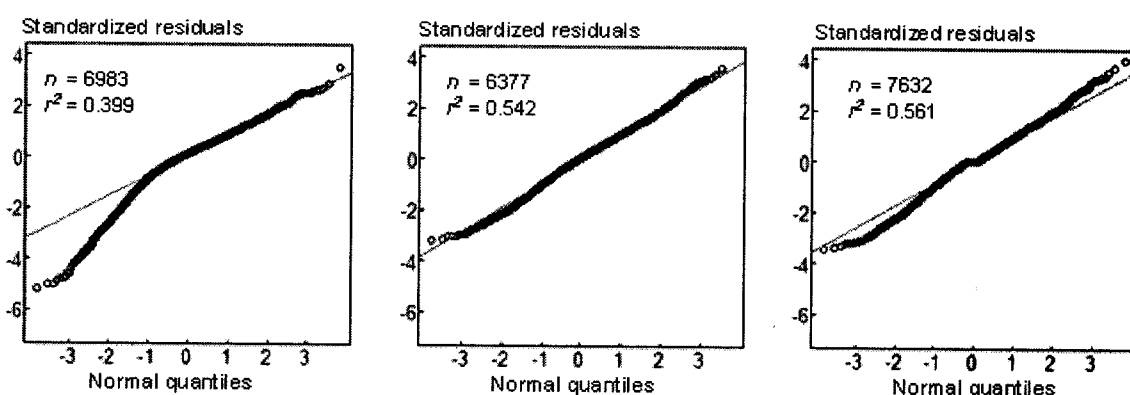


Fig 2—Plots of standardized residuals against normal quantiles with zero cell counts omitted (left panel), further cells with low residuals omitted (middle panel) and imputed using the fitted model (right panel).

stretches of successive unusually low case counts. In contrast, the four remaining selected districts (Na Yai Am, Kwao Si Narin, Khlung, and Chom Phra) showed no such evidence of under-reporting. Any outcome of zero or an extremely low number of reported cases was considered as a possible case of under-reporting.

The method involved first aggregating diarrhea cases by district counts for month and year from 260,522 individual cases into 7,632 records by cross-tabulating 106 districts over 72 months. The left panel of Fig 2 shows the plot of standardized residuals against normal quantiles after omitting the 649 cells with zero counts (8.5%) based on model (1). The residuals indicated a poor fit, which improved substantially when a fur-

ther 606 cells with residuals less than -1.4 were omitted (7.9%), as the middle panel shows. We thus used this latter model to impute cell counts for the omitted data, obtaining the plot shown in the right panel of Fig 2.

Table 3 lists the estimated proportions of under-reported cases. These proportions were highest in the October to December quarter and lowest in the April to June quarter. Similarly the estimated under-reporting rate was lowest in the year 1999 and gradually increased to 15.2% in 2001 and then decreased to 7.7% in 2004.

Based on the criterion we used, there were only four districts with no evidence of under-reporting (Mueang Sakaeo, Ban Kruat, Chom Phra and Phrasat). Chom Phra and

Table 2
Number of monthly reported diarrhea cases in ten selected districts.

Month ID	Pho Si Suwan	Mueang Chan	Ban Mai Chalyaphot	Sila Lat	Nam Klang	Chalerm Phra Kiet	Na Yai Am	Khwao Sinarin	Khlung	Chom Phra
1	0	0	10	18	0	53	57	88	59	193
2	0	0	10	7	0	19	24	22	48	141
3	0	0	0	19	0	34	26	15	37	88
4	0	0	3	12	0	11	42	18	37	44
5	0	0	1	16	0	9	50	18	32	59
6	0	0	4	24	0	25	35	33	32	53
7	0	0	1	8	0	26	21	16	33	43
8	0	0	4	0	0	16	20	20	33	47
9	0	0	1	18	5	9	33	21	28	45
10	0	0	3	4	15	16	30	27	33	58
11	0	0	2	5	1	10	17	27	22	73
12	0	0	0	7	0	1	18	45	31	121
13	5	13	3	0	88	74	41	55	75	101
14	9	25	10	0	64	57	48	51	58	112
15	25	2	4	8	66	62	39	30	41	92
16	18	0	3	12	53	48	34	31	44	45
17	13	0	6	21	55	68	35	47	49	69
18	37	0	2	4	72	69	23	29	54	115
19	0	1	7	10	35	63	23	45	51	110
20	0	0	4	0	21	40	18	27	31	60
21	0	0	2	0	65	31	25	26	38	53
22	0	0	1	0	17	37	20	25	48	62
23	0	0	1	0	28	48	35	29	36	41
24	0	0	0	0	0	59	21	42	31	79
25	17	20	10	0	21	77	34	56	53	91
26	9	17	5	0	15	61	43	31	63	74
27	6	10	7	0	16	60	52	48	34	72
28	2	3	2	0	37	64	59	34	29	49
29	0	0	1	0	108	54	26	33	27	57
30	0	0	6	0	160	71	32	53	42	84
31	0	0	2	0	81	32	32	67	37	111
32	0	0	0	0	56	33	39	53	31	118
33	0	0	0	0	45	22	26	33	30	85
34	7	0	0	0	37	20	9	37	27	36
35	9	0	0	0	0	11	34	19	31	39
36	14	0	0	1	0	0	13	37	11	39
37	25	42	3	91	0	0	42	67	55	381
38	51	20	6	60	1	0	70	58	87	173
39	34	4	1	66	1	1	63	49	94	187
40	26	30	5	54	2	3	33	48	61	91
41	2	30	0	16	0	4	38	32	37	113
42	1	24	26	17	0	5	53	38	47	153
43	0	19	31	30	0	7	36	42	46	178
44	2	12	14	31	2	7	37	48	39	178
45	2	22	7	33	5	7	30	41	26	94
46	5	12	3	24	3	0	22	45	27	118
47	6	0	24	20	1	0	24	68	38	136
48	5	0	9	30	2	0	28	28	55	104
49	5	2	13	2	2	18	30	100	68	187
50	16	0	12	4	4	44	65	77	46	217
51	14	1	19	2	6	34	74	46	51	171
52	3	1	14	8	1	48	26	27	32	103
53	1	0	17	5	3	32	38	17	54	105

A STATISTICAL METHOD FOR UNDER-REPORTED CHILD DIARRHEA

Table 2 (Continued).

Month ID	Pho Si Suwan	Mueang Chan	Ban Mai Chaiyaphot	Sila Lat	Nam Kliang	Chalerm Phra Kiet	Na Yai Am	Khwao Sinarin	Khlung	Chom Phra
54	0	1	16	4	7	13	47	36	48	91
55	4	0	14	4	2	7	34	37	48	117
56	5	2	10	6	3	9	27	29	51	123
57	1	1	46	7	1	8	32	22	73	56
58	4	0	52	2	2	0	26	36	54	76
59	2	1	87	1	4	0	17	25	45	99
60	5	1	6	0	3	0	15	48	11	106
61	51	85	111	54	98	0	47	111	60	124
62	48	43	12	48	60	58	45	57	69	116
63	28	81	19	50	57	28	33	60	69	170
64	8	45	12	41	53	27	9	27	45	101
65	10	47	23	47	51	41	24	29	46	115
66	35	31	22	25	32	47	35	29	59	182
67	73	41	35	55	113	62	28	60	45	274
68	64	60	39	54	53	30	20	72	36	219
69	36	31	12	26	0	51	20	56	31	131
70	2	7	17	16	0	8	11	43	42	103
71	0	1	17	2	1	54	22	76	40	125
72	0	0	11	2	0	72	13	39	33	129
Total	745	788	880	1,131	1,734	2,115	2,348	3,011	3,164	7,905
Population	12,306	9,654	14,226	10,926	24,336	20,526	13,128	16,506	21,480	32,658

Phrasat are located in the Surin Province. The highest estimates were found in Mueang Chan, Nam Kliang and Ban Mai Chaiyaphot Districts where the percentages exceeded 50%. Mueang Chan and Nam Kliang are located in the Si Sa Ket Province (see the left panel in Fig 5).

Diarrhea incidence

After aggregating the monthly cell counts (including those imputed from the under-reporting model for monthly data), we used the generalized estimating equation (GEE) model with a fixed correlation structure to account for spatial correlations 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 taken to be 0. Correlations between districts in pairs of dif-

Table 3
Estimates of under-reporting percentages
by quarter and year.

Factor	Percent
Quarter	
1 : Jan-Mar	8.1
2 : Apr-Jun	7.3
3 : Jul-Sep	9.6
4 : Oct-Dec	21.8
Mean	11.7
Year	
1999	6.4
2000	13.0
2001	15.2
2002	14.4
2003	8.9
2004	7.7
Mean	10.9

ferent provinces were similarly fixed at the means of the corresponding residual correlation, provided these means exceeded 0.1 in magnitude.

Table 4
Means and standard errors of residual correlations between districts within and between provinces.

Province	Chantaburi	Sakaeo	Surin	Buri Ram	Si Sa Ket	Ubon Ratchathani
Chantaburi	0.17 (0.04)					
Sakaeo	0.07 (0.03)	0.13 (0.05)				
Surin	0.03 (0.02)	0.05 (0.02)	0.03 (0.02)			
Buri Ram	-0.06 (0.02)	-0.05 (0.02)	0.03 (0.01)	0.17 (0.03)		
Si Sa Ket	-0.06 (0.02)	-0.02 (0.02)	0.00 (0.02)	0.09 (0.01)	0.12 (0.01)	
Ubon Ratchathani	0.02 (0.02)	0.09 (0.02)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.12 (0.02)

Table 4 shows the means and standard errors of the residual correlation coefficients between different districts in each province. These correlations were generally quite small, ranging from 0.03 in Surin to 0.17 in Chantaburi Province.

The results obtained by fitting the GEE model are shown in Fig 3. The left plot shows observed counts versus expected counts. The middle plot shows observed annual incidence per 1,000 versus the corresponding model-fitted values. Since both the cell counts and the corresponding incidence rates were strongly right-skewed, they were plotted on a cube root scale, which gave a squared correlation of 0.65 between the observed and fitted rates on this scale. The right plot shows the residuals (after filtering out the estimated spatial correlations) versus normal quantiles. Apart from a slight tilt, this plot shows little reason to doubt the normality assumption.

Fig 4 shows the fitted annual diarrhea incidence rates based on the GEE model. The graphs show 95% confidence intervals for annual diarrhea incidence/1,000 by quarter (left panel), year (middle panel) and district (right panel), each adjusting for the effects of the other two factors in the model. The dotted horizontal lines on each graph repre-

sent the overall mean annual incidence rate (19.4 per 1,000). The seasonal pattern of diarrhea incidence clearly indicates a peak in the January to March quarters with an adjusted annual incidence rate of 29.8 per 1,000 (95% CI 28.6 - 31.0). The trend steadily increased from 17.7 (95% CI 16.5 - 18.9) in 1999 to 20.0 (95% CI 18.9 - 21.0) in 2002, then dropping slightly to 18.6 (95% CI 17.8 - 19.4) in 2003 and increasing to 23.4 (95% CI 22.4 - 24.4) in 2004. The variation between districts was greater, ranging from 3.6 (95% CI 3.0 - 4.4) in Kantharalak District to 58.2 (95% CI 50.8 - 66.7) in Bung Bun District. The annual diarrhea incidence rate was generally higher than the mean in the districts of Buri Ram and Surin Provinces, generally lower than the mean in Chantaburi and Si Sa Ket Provinces, and typical of the whole in Sa Kao and Ubon Ratchathani Provinces.

The particular statistical and graphics methods used to produce Fig 5 enable under-reporting and incidence relativities between districts to be clearly illustrated. The right panel shows a thematic map of districts with diarrhea incidence coded according to whether the confidence interval exceeds, crosses, or is below the overall mean. Higher disease incidence occurred mainly in Buri Ram (14 of 23 districts) and Surin (8 of 17 districts) in the middle of the region.

A STATISTICAL METHOD FOR UNDER-REPORTED CHILD DIARRHEA

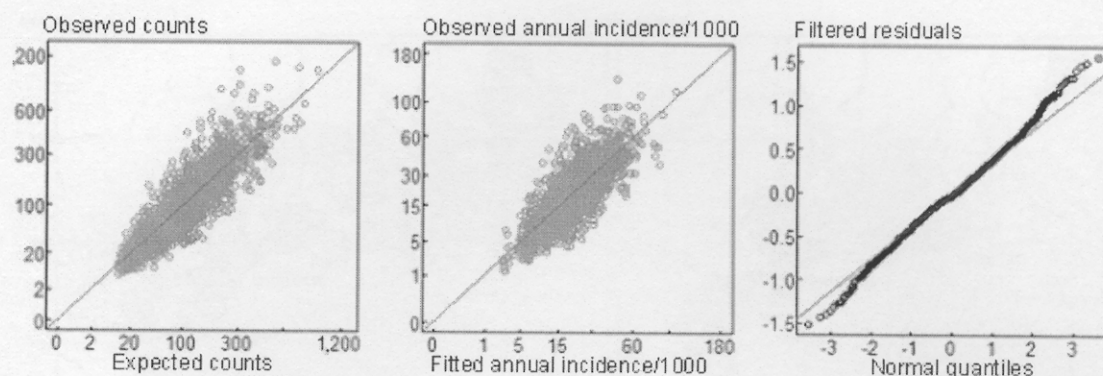
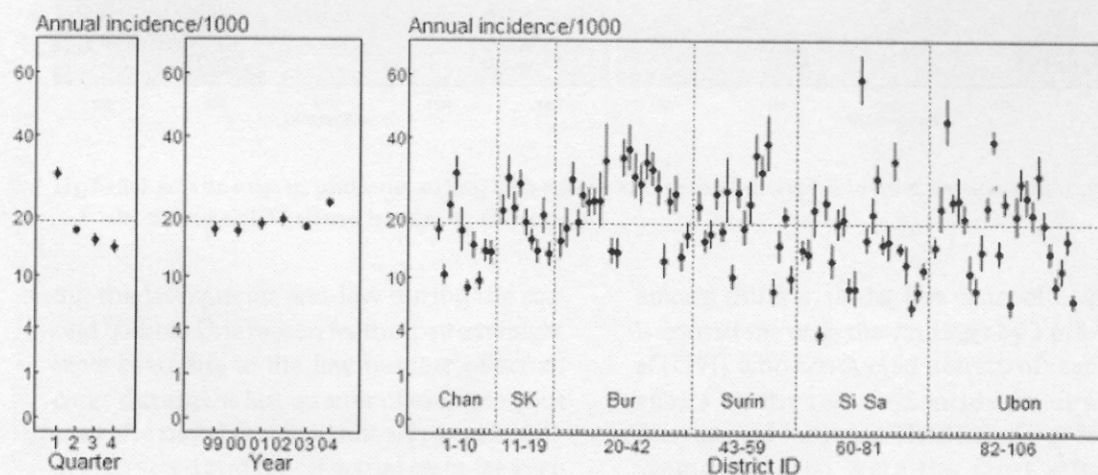


Fig 3—Plots of observed versus fitted counts and incidence rates (left panels) and residuals versus normal quantiles after fitting the GEE model.



Chan, Chantaburi; SK, Sakaeo; Bur, Buri Ram; Si Sa, Si Sa Ket; Ubon, Ubon Ratchathani

Fig 4—Confidence interval plots of annual incidence rates for each factor (quarter, year and district).

DISCUSSION

The results show the diarrhea incidence in the Thai provinces bordering Cambodia is a serious health problem (Staff of the Department of Planning and Health Information, 2008). The log-linear model was used to impute cell counts for the omitted data and the generalized estimating equation (GEE) model with a fixed correlation structure based on quarter, year and district which were used for analysis. The use of the GEE method for modeling spatial correla-

tion is discussed in detail in a recent review by Dormann *et al* (2007). Generalized linear models (GLMs) provide powerful statistical modeling (Aitkin *et al*, 1989) and the application of the GLMs to model epidemiological data was recommended by Flanders and Kleinbaum (1995). This method has also been applied to modeling diarrhea diseases by Kale and Hinde (2004), HIV/AIDS and other infectious disease mortality rates by Lim and Choonprabub (2007).

The estimated under-reporting of diarrhea case (Table 3) was relatively high dur-

SOUTHEAST ASIAN J TROP MED PUBLIC HEALTH

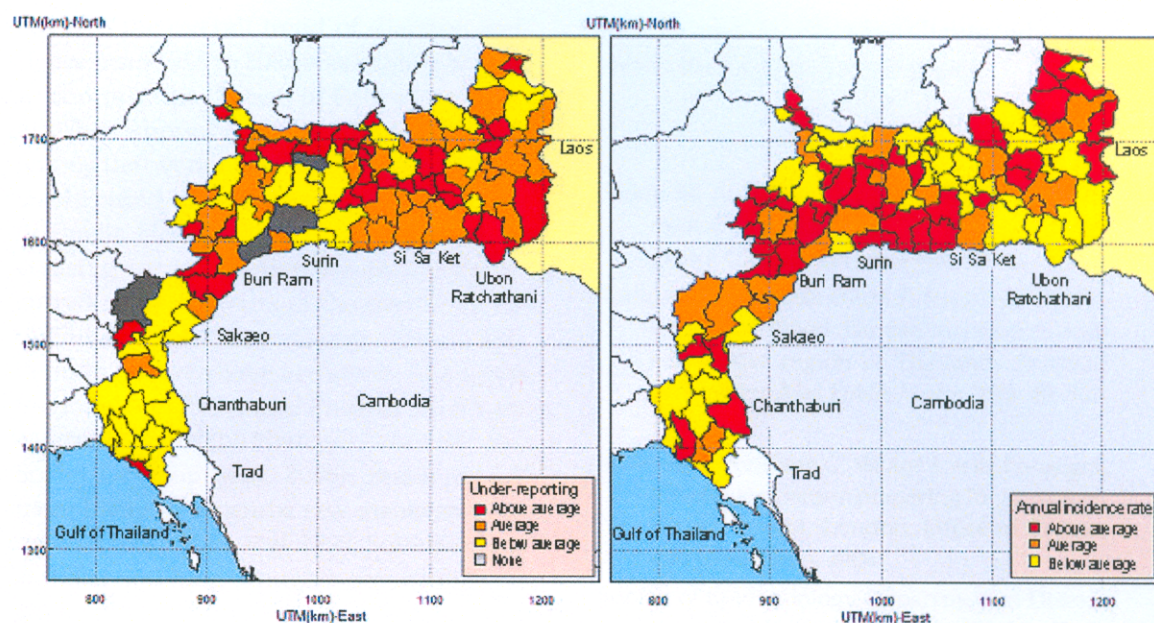


Fig 5—Schematic map of under-reporting (left panel) and annual diarrheal incidence rates (right panel) in districts of Thailand bordering Cambodia.

ing the last quarter and low during the second quarter. One reason for this pattern might have been due to the low number of actual cases during the last quarter of each year, but may also have been due to health worker failure to record and report actual cases for various reasons, including computer system problems. During the period 1999 to 2004, estimated under-reporting was lowest in 1999 and steadily increased in 2001 and then decreased by 2004. The highest estimates were found in the three districts Mueang Chan, Nam Kiang and Ban Mai Chaiyaphot, where our estimates exceed 50%. Evidence of under-reporting was also found in most districts of Si Sa Ket Province. Possible causes include overworked government personnel, inadequate coordination with private hospitals, and inadequate supervision at the central public health level (Saengwonloey *et al*, 2003). However, there were four districts with no evidence of under-reporting.

The highest diarrhea incidence occurred

among children under five years old, which is consistent with the findings by Pinfold *et al* (1991) who conducted a study of seasonal effects on the reported incidence of acute diarrhea in Northeast Thailand. They found young children were the most affected group, with a reported annual incidence of 2,952 per 100,000 for children less than five years old. Kosek *et al* (2003) found that diarrhea accounted for 21% of all deaths of children less than five years old. Early intervention for this age group could reduce the incidence.

According to this study, the diarrhea incidence during the study period was relatively high from January to March. This period mainly overlaps the winter and summer seasons and is associated with a high risk of diarrhea (Pinfold *et al*, 1991). Diarrhea disease was influenced by El Niño (Patz, 2005) in Peru, the number of diarrhea cases increased with an ambient temperature increase (Checkley, 2000). Moreover, we dis-

A STATISTICAL METHOD FOR UNDER-REPORTED CHILD DIARRHEA

covered the overall trend of diarrhea incidence from 1999 to 2004 was slightly higher. In contrast, the Bureau of Policy and Strategy has shown that the diarrhea trend among the overall Thai population remained stable based on statistics from the Ministry of Public Health (2001). The high, and increasing trends of diarrhea incidence occurred in the districts in the middle of the region. Possible reasons are as follows: residents in this area have a relatively low Gross Regional and Provincial Product (GRP) per capita (Office of the National Economic and Social Development, 2006); residents are more likely to consume raw meats and fermented foods (Lee *et al*, 1993; Somnasang *et al*, 1998) and long droughts occur in the region.

In conclusion, the data analysis model enables adjustments to be made to compensate for under-reporting and should thus provide more accurate estimates of disease incidence. The model may be useful for health planning in countries similar to Thailand where routine epidemiological reports of diarrhea and other disease cases are provided at the district level, because it provides a simple method based on readily available demographic data. The model can also be used to identify an unusually high annual incidence within the period of its occurrence, and thus enable health authorities to reduce the severity of ensuing epidemics by implementing preventative measures put in place for the demographic group at risk.

ACKNOWLEDGEMENTS

This study was funded by the Graduate School, Prince of Songkla University. We would like to thank the Ministry of Public Health for giving permission to use the data. We are grateful for Prof Don McNeil and Dr. Vorasith Sornsrivichai for their helpful advice and suggestions. We also thank Greig

Rundle and Dr Shanley Chong who corrected the English for our paper.

REFERENCES

- Aitkin M, Anderson D, Francis B, Hinde J. Statistical modelling in GLIM. Oxford, England: Clarendon Press, 1989.
- Ardkeaw J, Tongkumchum P. Statistical modeling of childhood diarrhea in the north-eastern border region of Thailand. *Southeast Asian J Trop Med Public Health* 2009; 40: 807-15.
- Brownlie J, Peckham C, Waage J, *et al*. Foresight. Infectious diseases: preparing for the future. Future threats. London: Office of Science and Innovation, 2006.
- Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health. Thailand: Annual epidemiology surveillance report 2001. Bangkok: War Veterans Organization of Thailand, 2002.
- Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health. Thailand: Annual epidemiology surveillance report 2006. Bangkok: War Veterans Organization of Thailand, 2007.
- Carlos CC, Santel CM. Etiology and epidemiology of diarrhea. *Phil J Microbiol Infect Dis* 1990; 19: 51-3.
- Checkley W. Effect of El Niño and ambient temperature on hospital admissions for diarrhoeal disease in Peruvian children. *Lancet* 2000; 355: 442-50.
- Dormann CF, McPherson JM, Araúejo MB, *et al*. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 2007; 30: 609-28.
- Flanders WD, Kleinbaum DG. Basic models for disease occurrence in epidemiology. *Int J Epidemiol* 1995; 24: 1-7.
- Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York: John Wiley and Sons, 2000.
- Intusoma U, Sornsrivichai V, Jiraphongsa C,

- Varavithaya W. Epidemiology, clinical presentations and burden of rotavirus diarrhea in children under five seen at Ramathibodi Hospital, Thailand. *J Med Assoc Thai* 2008; 91: 1350-5.
- Jiraphongsa C, Bresee JS, Pongsuwanna Y, et al. Epidemiology and burden of rotavirus diarrhea in Thailand: results of sentinel surveillance. *J Infect Dis* 2005; 192: S87-93.
- Kale PL, Hinde JP. Modeling diarrhea disease in children less than 5 years old. *Ann Epidemiol* 2004; 14: 371-77.
- Kleinbaum DG, Klein M. Logistic regression - A self learning text. 2nd ed. New York: Springer-Verlag, 2002.
- Konchom S, Singhasivanon P, Kaewkungwal J, et al. Trend of malaria incidence in highly endemic provinces along the Thai borders, 1991-2001. *Southeast Asian J Trop Med Public Health* 2003; 34: 486-94.
- Kosek M, Bern C, Guerrant RL. The magnitude of the global burden of diarrhoeal disease from studies published 1992-2000. *Bull World Health Organ* 2003; 81: 197-204.
- Lee HC, Steinkraus HK, Reilly PJ. Fish fermentation technology. United Nations University Press, 1993: 155-66.
- 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.
- Lumbiganon P, Panamonta M, Laopaiboon M, Pothinam S, Patithat N. Why are Thai official perinatal and infant mortality rates so low?. *Int J Epidemiol* 1990; 19: 997-1000.
- Ministry of Public Health. Public health statistic. Nonthaburi: Ministry of Public Health, 2001.
- Office of the National Economic and Social Development. Gross regional and provincial product. Bangkok: The Office of the National Economic and Social Development, 2006.
- Parashar UD, Bresee SJ, Glass RI. The global burden of diarrhoeal disease in children. *Bull World Health Organ* 2003; 81: 236.
- Patz JA, Lendrum DC, Holloway T, Foley JA. Impact of regional climate change on human health. *Nature* 2005; 438: 310-17.
- Pinfold JV, Horan NJ, Mara DD. Seasonal effects on the reported incidence of acute diarrhea disease in northeast Thailand. *Int J Epidemiol* 1991; 20: 777-86.
- R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2008. (Cited 2008 Jun 28). Available from: URL: <http://www.R-project.org>
- Saengwonloey O, Jiraphongsa C, Foy H. Thailand report: HIV/AIDS surveillance 1998. *J Acquir Immune Defic Syndr* 2003; 32: S63-7.
- Somnasang P, Moreno G, Chusil K. Indigenous knowledge of wild food hunting and gathering in north-east Thailand. *Food Nutri Bull* 1998; 19: 359-65.
- Staff of the Department of Planning and Health Information, Ministry of Health, the Reproductive Health Association of Cambodia, and PRB, No date. (Cited 2008 Dec 10). Available from: <http://www.prb.org/Articles/2002/ChildreninCambodiaFaceHighMortalityRate.aspx>
- Thai Working Group on burden of disease and injuries. Burden of disease and injuries in Thailand: priority setting for policy. Nonthaburi: Ministry of Public Health, 2002.
- Thimasarn K, Jatapadma S, Vijaykadga S, Strichaisinthop J, Wongsrichanalai C. Epidemiology of malaria in Thailand. *J Travel Med* 1995; 2: 59-65.
- Vargas M, Gascon J, Casals C, et al. Etiology of diarrhea in children less than five years of age in Ifakara, Tanzania. *Am J Trop Med Hyg* 2004; 70: 536-9.
- Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer-Verlag, 2002: 144-51.
- Yan J, Fine J. Estimating equations for association structures. *Stat Med* 2004; 23: 859-80.

3.4 Manuscript 2

Modeling Trends in Tuberculosis Incidence in Nepal

Sampurna Kakchapati^a, Sulawan Yotthanoo^a and Chamnein Choonpradup^{b,*}

^aPhD student, Program in Research Methodology.

^bDepartment of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Thailand.

*Author for correspondence; e-mail: cchamnein@bunga.pn.psu.ac.th

Abstract

Tuberculosis (TB) is a major cause of morbidity, mortality, and disability worldwide. It constitutes a large burden of infectious disease in Nepal. Every year more than 30,000 people develop active tuberculosis and 5,000-7,000 people have died due to TB. The magnitude of tuberculosis infection across the country is alarming and varies with location. The objective of the study was to model the trends in incidence of Tuberculosis from 2003 to 2008. A retrospective study was conducted in Nepal of tuberculosis incidence by gender and location over the six years period. Data were obtained for 198,734 tuberculosis cases from the South Asian Association for Regional Cooperation Tuberculosis and HIV/AIDS center (STAC). A negative binomial model using two trend eigenvectors as predictors was used to determine the trends of TB incidence in Nepal. The incidence rate of TB decreased from 1.32 to 1.24 per 1000 populations from 2003 to 2008. The result of fitting a negative binomial model was acceptable as indicated by the residual plots. The model extracted two trend eigenvectors that characterized (a) a decreasing trend of TB incidence, and (b) a tendency to increase during the first five years followed by a sharp drop in 2008. Tuberculosis is still a public health problem in Nepal. This study showed a steady decreasing trend in TB incidence but the numbers of cases are still very high. Gender differences exist in Tuberculosis incidence in Nepal. Higher rates were observed in Terai Region and urban areas. These findings highlight the need of tuberculosis control measures on a sustained and long term basis on high TB burden areas of Nepal.

Keywords: Tuberculosis, Negative binomial model, eigenvectors, count model

Introduction

Tuberculosis (TB) remains a major cause of infectious disease mortality, with an estimated 8.8 million new cases and 1.6 million deaths annually (WHO 2008, Werf and Borgdorff 2007). With 2.5% of all deaths worldwide attributed to TB, the disease is the seventh most common cause of death, after cardiovascular diseases, respiratory infections, chronic obstructive pulmonary disease, diarrheal diseases and HIV/AIDS (WHO 2004). TB remains a serious public health problem among particular patient populations, including children and elderly, and is prevalent in many urban areas (Dye et al 2005, Blumberg et al 2005). The global burden of TB is large, and is likely to remain high among public health problems in coming decades (Dye et al 1999, Lopez et al 2006).

Tuberculosis is a major public health problem in Nepal. Data from the National Tuberculosis Programme of Nepal shows that there are about 30,000 infectious cases and 5,000-7,000 deaths due to TB annually (STC 2007). It is the most common cause of death in the most economically productive age group comprising adults aged 15 to 49 years (Harries et al 1998). The reported incidence of all forms of tuberculosis amongst the general population was 176 per 100,000 in 2006 with mortality (including HIV) of 23 cases per 100,000 (WHO 2008). As in other countries, the tuberculosis epidemic in Nepal can be traced in part to poor working and living conditions. Many of these conditions persist to this day, and along with the Multi Drug resistance (MDR-TB) and HIV/AIDS epidemic, have fueled the current high levels of tuberculosis disease in the country.

The National Tuberculosis Center's annual reports provide evidence that the magnitude of tuberculosis infection across the country is high and varies with location. The reported caseload is extremely high in the Terai and the Hill parts of the country and is also high in many urban areas (STC 2007).

Several studies have found a gender difference in tuberculosis incidence (Uplekar et al 2001, Holmes et al 1998). In Nepal, the female/male ratio was found to be below parity among TB suspects undergoing sputum examination and for all types of TB case detection (Shrestha and Jha 2007).

Public health officials are often required to evaluate disease incidence in the country. They need to compare the standardized disease incidence rate within the area and time frame so that necessary actions can be taken. Statistical modeling may provide the quantitative framework for investigating key issues related to tuberculosis incidence, transmission dynamics and to predict the effects of different interventions. A mathematical model was used to describe the spatial and temporal variation in TB and HIV in India (Williams et al 2005). Another model was proposed to simulate TB dynamics and to evaluate the potential impact on active TB prevalence of several intervention strategies in highly endemic overcrowded prisons in Brazil (Legrand et al 2008). Poisson regression model was applied to determine the spatial and temporal variations in incidence of tuberculosis and identified an upward trend in the number of reported cases of tuberculosis in southern, eastern and middle Africa (Uthman et al 2005). Similar study on tuberculosis incidence in Portugal based on spatiotemporal clustering revealed that TB incidence showed clear spatial patterns, and high incidence rate space-time clusters were identified in three areas of the nation between 2000 and 2004 (Nunes 2007).

Investigating the regional and temporal pattern of disease can indicate areas with problems and possibly predict periods of likely disease epidemics. It can also help the concerned health authorities to plan an effective prevention program. The Poisson distribution and its extension to the negative binomial distribution to handle over dispersion is a standard approach to modeling event count data.

The aim of the study is to model the trends in incidence of tuberculosis in Nepal from 2003 to 2008.

Materials and methods

Study Area and Data Source

Nepal, officially the Federal Democratic Republic of Nepal, is a landlocked country in South Asia and is the world's youngest republic. It is bordered to the north by the People's Republic of China, and to the south, east, and west by the Republic of India. It has five development regions (eastern, central, western, mid western and far western), 14 zones, 75 districts, and has a current population growth of 2.2%

(Wikipedia 2009). On the basis of topography, it is divided into three distinct geographical regions; Mountain (7% of the population), Hill (43%) and Terai (50%), in decreasing altitude.

Table 1: Definitions and populations of super-districts

Code	Super-districts	Population	Code	Super-districts	Population
1	Darchula	138,455	33	Rupandehi	833,681
2	Baitadi	265,773	34	Chitwan	555,121
3	Dadeldhura	144,833	35	Makwanpur	457,349
4	Kanchanpur	451,790	36	Parsa	580,572
5	Bajhang+ Bajura	312,813	37	Bara	652,286
6	Doti	233,751	38	Rautahat	630,969
7	Achham	259,964	39	Rasuwa+ Sindu	398,709
8	Kailali	744,760	40	Dhading	388,103
9	Karnali (zone)	350,358	41	Nuwakot	330,100
10	Dailekh	256,064	42	Kathmandu	1,304,954
11	Jajarkot	153,719	43	Kavre	435,759
12	Surkhet	338,208	44	Bhaktapur	259,223
13	Bardiya	454,657	45	Lalitpur	393,228
14	Banke	457,903	46	Dolkha	233,748
15	Rukum	215,643	47	Ramechhap	242,525
16	Salyan	242,739	48	Sindhuli	322,698
17	Rolpa	236,419	49	Mahottari	637,294
18	Pyuthan	240,913	50	Dhanusha	779,388
19	Dang	540,577	51	Sarlahi	734,858
20	Mustang+ Myagdi	145,264	52	Solu+ Okhal	297,154
21	Baglung	303,556	53	Khotang	258,656
22	Parbat	179,200	54	Udaypur	339,163
23	Manang+ Lamjung	207,714	55	Siraha	661,030
24	Gorkha	325,824	56	Saptari	658,681
25	Kaski	444,787	57	Sankhu+ Tehra	308,761
26	Syangja	358,149	58	Bhojpur	227,545
27	Tanahu	362,300	59	Dhankuta	189,483
28	Gulmi	336,857	60	Sunsari	733,919
29	Arghakhanchi	237,762	61	Morang	978,441
30	Palpa	304,171	62	Taple+ Panchthar	379,783
31	Nawalparasi	665,258	63	Illam	329,176
32	Kapilvastu	567,152	64	Jhapa	795,779

Nepal is covered by the National Tuberculosis Control Program (NTC), which strictly follows the World Health Organization strategy (WHO 2008) defined as Directly Observed Treatment, Short Course, and regularly issues a progress report. The information used, regarding cases notified between 2003 to 2008 was provided by

STAC, the specific NTC information system managed by the NTC coordination team, working for the prevention and control of TB and HIV/AIDS in the Region (STC 2007). The reported cases for each year are available in computer files comprising characteristics of the disease, gender, address, and the severity of the illness. The independent variables are: (a) gender, (b) location and (c) calendar year (2003 to 2008).

To simplify the effect of location of residence when calculating incidence rates, one or more contiguous districts in each zone were grouped together to form 64 “super-districts” containing populations of above 100,000 on average, as shown in Table 1, where they are listed in order of geographical location from far western to eastern (keeping district within the same zone together) with their 2008 populations.

Statistical methods

Poisson regression is commonly used for modeling the number of cases of disease in a specific population within a certain time period. If λ_{ijt} denotes the mean incidence rate for gender i , geographical location j and year t , an additive model with this distribution is expressed as

$$\ln(\lambda_{ijt}) = \ln(P_{ij}) + \mu + \alpha_i + \beta_j + \gamma_t. \quad (1)$$

In this model, P_{ij} is the corresponding population at risk in 1000s and the terms α_i , β_j and γ_t represent gender, location and year effects that sum to zero so that μ is a constant encapsulating the overall incidence. The model fit is then assessed by the linearity in the plot of deviance residuals against normal quantiles, and also by examining plots of observed counts and appropriately scaled incidence rates against corresponding fitted values based on the model. The additive model including interaction between location and year takes the form

$$\ln(\lambda_{ijt}) = \ln(P_{ij}) + \alpha_i + \beta_{jt}. \quad (2)$$

To allow for possible interactions between gender, location and year effects, model (1) may be extended to

$$\ln(\lambda_{ijt}) = \ln(P_{ij}) + \alpha_{ij} + \beta_{jt}. \quad (3)$$

The interaction between the gender and year is omitted because it is reasonable to assume that the gender effect varies with location but not with year.

This model has interactions between two factors so if I, J and T are the numbers of respective levels of these factors, the number of parameters is $IJ+IT+JT-(I+J+T)-4$. For example with the two sexes, 64 locations and 6 years the number of parameters is equal to 448.

Since it is difficult to interpret a model with such a large number of parameters, we considered an alternative set of models that include interactions, these take the form, for $m < T$,

$$\ln(\lambda_{ijt}) = \ln(P_{ij}) + \alpha_{ij} + \sum_{k=1}^m \beta_j^{(k)} \gamma_t^{(k)}. \quad (4)$$

This model is non-linear so it cannot be fitted directly as a Poisson generalized linear model. However if y_{ijt} is the natural logarithm of the incidence rate in cell (i, j, t) , the model can be approximated by a non-linear model with additive errors ε_{ijt} , that is,

$$y_{ijt} = \alpha_{ij} + \sum_{k=1}^m \beta_j^{(k)} \gamma_t^{(k)} + \varepsilon_{ijt}. \quad (5)$$

To handle observations with zero cell counts, an appropriate adjustment is needed such as replacing any zero counts by a small positive constant. For specified values of i , Theil (1983) showed that the least squares estimates of the $\gamma_t^{(k)}$ parameters in model (5) are the elements of the eigenvector of the matrix $Y_c^T Y_c$ corresponding to its k^{th} largest eigenvalue, where Y_c has elements $y_{ijt} - \bar{y}_{ij}$ and Y^T denotes the transpose of Y . The corresponding least squares estimates of the $\beta_j^{(k)}$ parameters are then expressed in terms of the eigenvectors $\gamma_t^{(k)}$ as

$$\beta_j^{(k)} = \sum_{i=1}^2 \sum_{t=1}^T \gamma_t^{(k)} (y_{ijt} - \bar{y}_{ij}). \quad (6)$$

If the $\beta_j^{(k)}$ parameters are regarded as fixed, model (4) can be fitted using Poisson regression, giving both estimates and standard errors for the remaining parameters. In

practice, this assumption would be reasonable if the $\gamma_i^{(k)}$ were replaced by basis functions $g_i^{(k)}$ such as orthonormal polynomial or spline functions of degree k . The model then may be written

$$\ln(\lambda_{ij}) = \ln(P_{ij}) + \alpha_{ij} + \sum_{k=1}^m \beta_j^{(k)} g_i^{(k)}. \quad (7)$$

For $I=2$, $J=64$ and $T=6$, the number of parameters in this model varies from 192 when $m = 1$ to 448, the number of parameters in model (2), when $m = T-1$.

Model (7) also gives adjusted incidence rates for each factor of interest, obtained by replacing the parameters corresponding to the other factors by constants chosen to ensure that the total expected number of cases equals the observed number.

Sum contrasts (Venables and Ripley 2002, Tongkumchum and McNeil 2009) were used to obtain confidence intervals for comparing the adjusted incidence rates within each factor with the overall incidence rate. Since the confidence intervals for factor-specific incidence rates obtained from the 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 schematic maps of super-districts according to their estimated tuberculosis annual incidence rates.

Poisson models for disease counts are often over-dispersed due to clustering, in which case the negative binomial model is more appropriate (Venables and Ripley 2002).

The negative binomial model is an extension of the Poisson model for incidence rates that allows for the over dispersion that commonly occurs for disease counts.

The R program was used for all statistical analysis, graphs and maps (Venables et al 2008).

Results

Preliminary Analysis

During the study period 198,719 confirmed cases were notified from 2003 to 2008. Among all cases, 86,722 cases were new smear positive cases, 13,545 cases were relapse, 1,727 were failure, 1,914 were defaulters, 55,333 were new smear negative

and 39,458 were extra pulmonary cases. The number of cases varied from 0 to 1,801 per year, with 127,979 male cases and 70,740 female cases. The mean incidence rate of TB was 1.31 per 1,000 population. The incidence rates by year are shown in the Table 2.

Table 2: TB incidence rates by year

Years	No. of TB cases (N= 198719)	Population	Incidence
2003	31,637	23,961,451	1.32
2004	32,903	24,516,403	1.34
2005	34,077	25,266,209	1.34
2006	33,206	25,714,085	1.29
2007	33,450	26,284,014	1.27
2008	33,446	26,805,469	1.24

Figure 1 shows plots of the eigenvectors $\gamma_i^{(k)}$ for $k = 1, 2$. Since the first eigenvector shows a decreasing linear trend and the second eigenvector shows a linear trend for $t = 1$ to 5 followed by a sharp drop at $t = 6$ (a saw tooth with peak in 2007), it is reasonable to choose these functions as basis functions. Note that the first of these fixed functions has two parameters and the second has three parameters, and these five parameters can be determined by the requirement that they are normalized to have mean 0, sum of squares 1 and correlation 0. The functions are thus $g_t^{(1)} = 0.837 - 0.239 t$ and $g_t^{(2)} = -0.414 + 0.187 t$ for $0 < t < 6$ and $g_t^{(2)} = -0.736$ for $t = 6$. In Figure 1, the dotted lines denote these fixed functions.

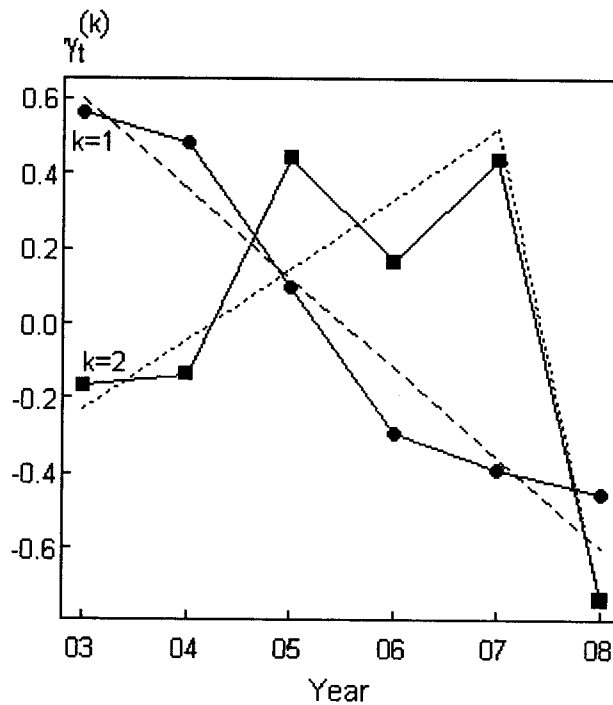


Figure 1: Plot of two eigenvectors and corresponding basis functions (dotted)

Statistical Analysis

Table 3 summarizes results obtained from fitting different Poisson and negative binomial models to the data. Models based on equation (7) with $m > 2$ have too many parameters to be easily interpreted. However, equations (2) and (3) have many more parameters than the two categories of equation (7), as can be seen in Table 3. To ensure that the negative binomial models are hierarchical, we used the value of θ estimated from model E (391.13 with standard error 60.0) for all models.

Table 3: Analysis of Deviance for Poisson and Negative Binomial models

Model	No. parameters	d.f.	Residual Deviance Poisson	Residual Deviance Negative Binomial
A: Equation (1)	70	698	3708.7	2183.4
B: Equation (2)	385	383	2331.6	1346.0
C: Equation (3)	448	320	717.2	484.4
D: Equation (7) (m=1)	192	576	1585.6	1022.6
E: Equation (7) (m=2)	256	512	1317.3	860.7

Figure 2 plots of the deviance residuals versus normal quantiles are compared, and it can be seen that the fit for the negative binomial model is more acceptable than that for the Poisson model. The largest two outliers identified in each graph were super-districts 27 and 46 during the years 2007 and 2004, respectively. The corresponding observed counts were 215 and 107 whereas the fitted values given by the negative binomial model were 124.4 and 62.8, respectively.

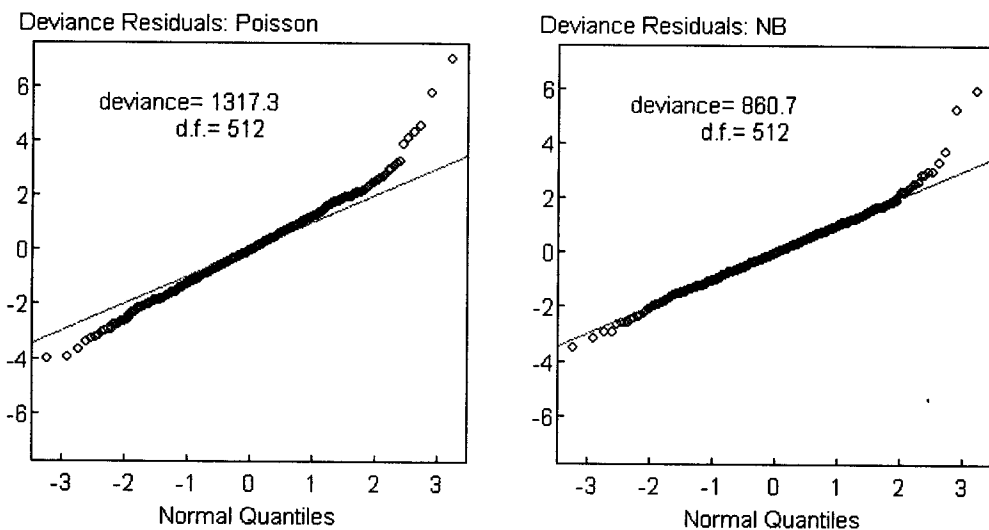


Figure 2: Diagnostic plots for Poisson and negative binomial models

Figure 3 shows plots of observed counts and observed annual incidence rates per 1000 population versus corresponding fitted values using the negative binomial model, indicating that the model fitted the data quite well.

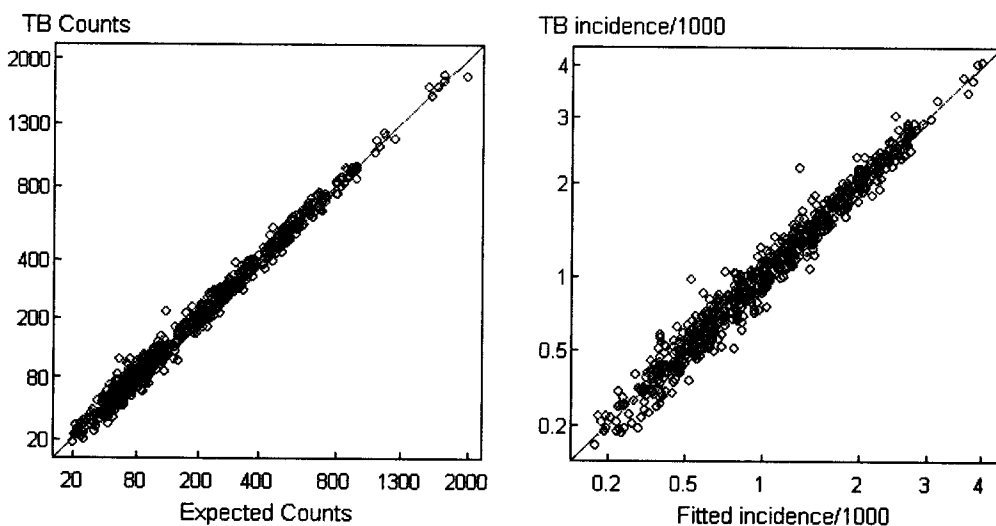


Figure 3: Plots of observed counts and observed incidence against fitted values

Figure 4 shows plots of the adjusted annual TB incidence rates/1000 for males and females and their confidence intervals for each super-district. The horizontal line corresponds to the overall incidence rates for males and females combined (1.31 per 1,000 population). The dark line represents the males and the light line represents females. The dotted horizontal dark line corresponds to the mean incidence rates for males (1.70 per 1000 population) and the dotted horizontal light line corresponds to the mean incidence rates for females (0.91 per 1000 population). In most of the super-districts, there is high incidence of TB in males.

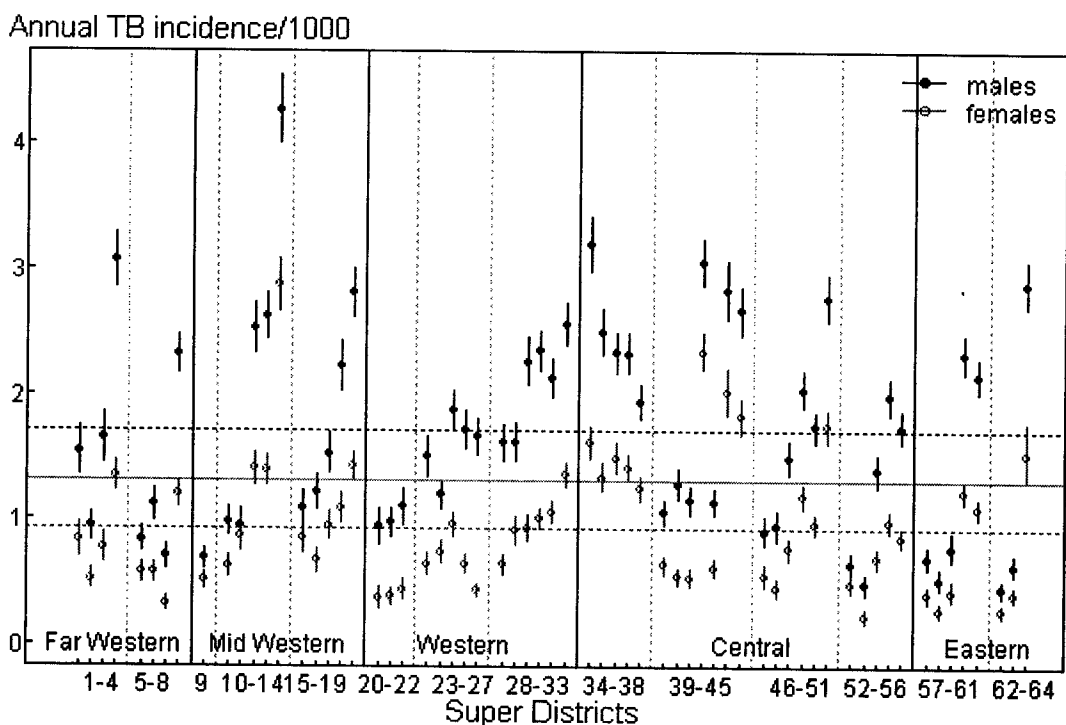


Figure 4: Annual TB incidence/1000 for males and females

Figure 5 shows a schematic map of the male and female adjusted incidence rates for super-districts, using the confidence intervals plotted in Figure 4 to classify these regions as above the mean (darkest shade), below the mean (lightest shade) or not evidently different from the mean (intermediate shade). Higher TB incidence rates occurred for males and females in most of the super-districts of the Terai region and some super-districts of the Hill region. The plot also shows similar patterns for males and females, suggesting that it might be reasonable to fit a simpler model that does not contain an interaction between gender and super-district. However, when the simpler model was fitted, the residual deviance increased to 1719.5 with 575 degrees

of freedom, and the plot of deviance residuals also indicated a poor fit, so the model containing the interaction was preferred.

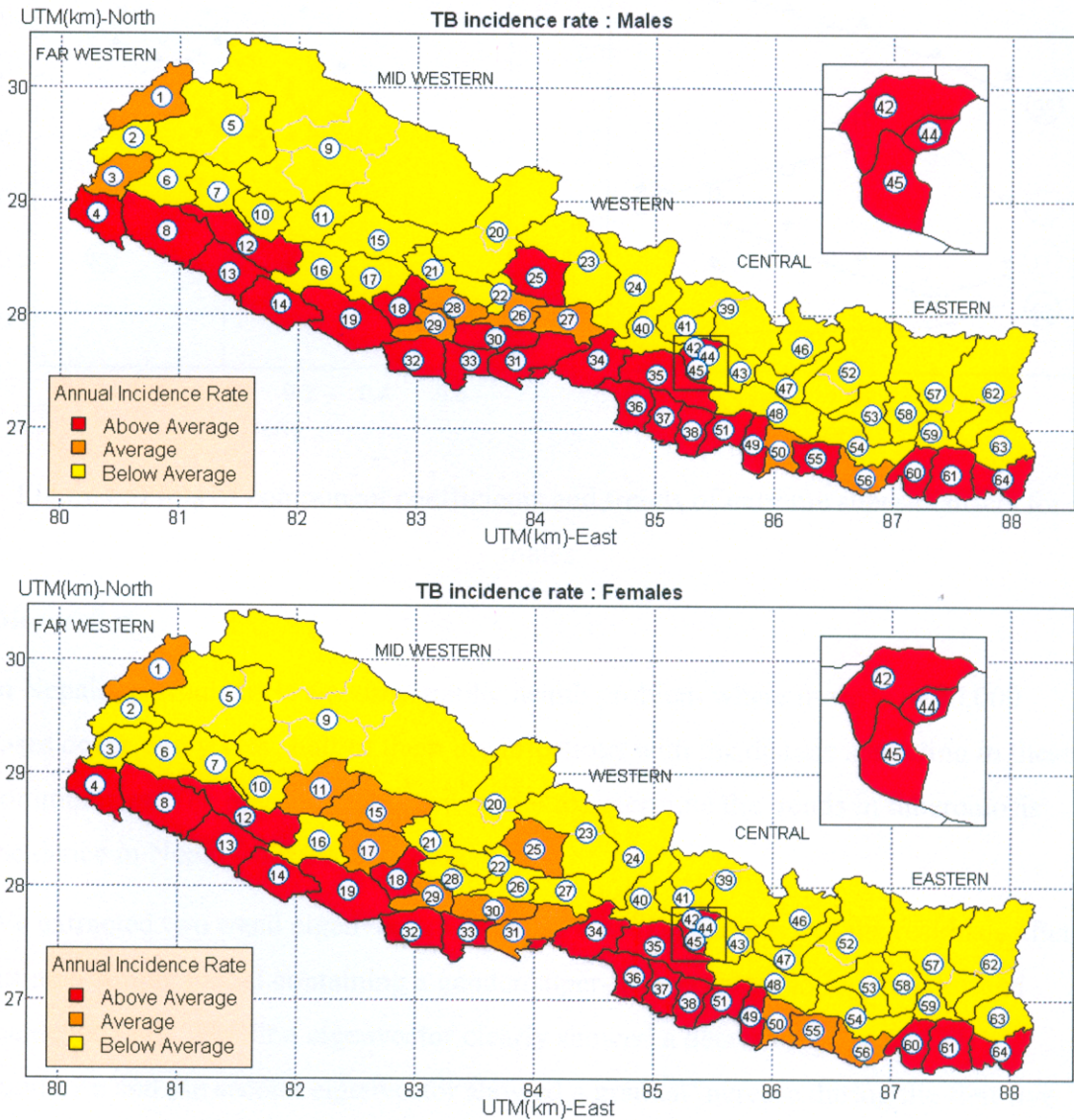


Figure 5: Schematic map of annual tuberculosis incidence rates for males and females

Figure 6 shows a plot of the estimated trend coefficients with circles indicating confidence regions (left panel) and annual incidence rates for four selected districts for males corresponding to extreme coefficients (right panel). The lines are the trends of incidence rates given by fixed functions corresponding to the dotted lines in Figure 1. Super-district 35 had a pure downward trend, super-district 17 had a trend similar to the saw tooth shape in figure 1 and super-districts 11 and 53 had a hybrid trends composed of weighted linear combinations of the two trend functions.

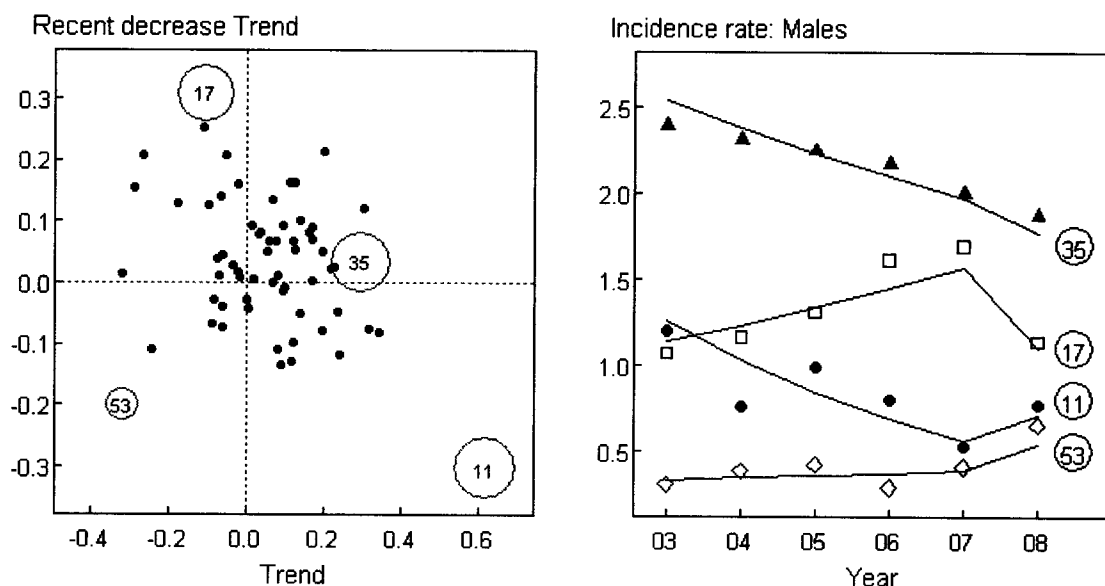


Figure 6: Plots of component coefficients and trends of extreme super-districts for males

Discussion

In Nepal, tuberculosis is a serious public health problem where more than 30,000 cases occur every year; half of them are infectious with the disease spreading in these communities. The present study used model to determine the trends in tuberculosis incidence in Nepal from 2003 to 2008.

We extracted two trend eigenvectors from the covariance matrix of the residuals after fitting a simple model containing a gender/super-district factor to log transformed incidence rates. The first eigenvector clearly showed a decreasing trend of TB incidence and the second eigenvector showed a gradual increase during the first five years followed by a sharp drop in 2008. We then replaced these data generated eigenvectors by fixed orthonormal functions comprising a straight line and bent line and then fitted generalized linear Poisson and negative binomial models containing the gender and super-district factor and the two fixed functions with super-district specific parameters. The negative binomial model fitted reasonably well as indicated by plots of deviance residuals. We also plotted confidence intervals for the adjusted incidence rates for gender and super-district.

The overall annual incidence rate of TB was found to be 1.31 per 1,000 population. The findings showed that gender differences existed in the incidence of TB; the male to female incidence ratio was 1.86. This is reasonably consistent with TB incidence and gender patterns found in recent studies in Nepal (Shrestha and Jha 2007).

Epidemiological findings demonstrate that in most settings, tuberculosis incidence rates are higher for males, at all ages except in childhood, when they are higher for females. Studies have reported that sex differentials in prevalence rates begin to appear between 10 and 16 years of age, and remain higher for males than females thereafter (WHO 2003). A possible reason for the higher prevalence in post-adolescent males is that biological factors associated with being female (such as hormonal factors) may protect post-adolescent females from TB infection (Dolin 1998).

The decreases in trends of TB incidence as fitted by fixed functions over the six year periods were consistent with the WHO report and STC annual report on tuberculosis. The decrease in TB incidence may be attributed to successful TB control programs in the country with the expansion of DOTS, case finding and treatment success in the recent years in Nepal (STC 2007). Similarly a recent first national tuberculin survey carried out among school children in Nepal shows the ARTI (Annual Risk of Tuberculosis Infection) in Nepal to be lower than previous estimates, indicating a decrease in transmission of tuberculosis (Shrestha et al 2008).

There were pronounced spatial variations in TB incidence for males and females with higher rates occurring in the Terai region, followed by the Hill and Mountain regions. Thus it can be concluded that tuberculosis is more prevalent in the Terai region. Studies from the UK and Spain have shown seasonal variations in tuberculosis rates and higher notification rates over summer and in hotter regions (Douglas et al 1996 and Douglas et al 2000). This increase has been attributed to impaired host defence mechanisms (Davies 1997). But the notification rate in the Terai region can be attributed not only to medical factors, but also social and environmental factors. The Terai region is characterized by high temperatures, low socio-economic status, malnutrition, high levels of poverty and social deprivation, all contributing to TB infection. However the lower incidence rates of tuberculosis in mountain areas are

consistent with studies from Kenya and Mexico, which reported that the tuberculosis incidence decreases strongly with increasing altitude (Mansoor et al 1999, Vargas et al 2004). The cause of the close inverse relationship of altitude and TB incidence might be related to the well known changes in alveolar oxygen pressure at different altitudes (Vargas et al 2004, West et al 1990). In this study, the steady decline of the barometric pressure as altitude decreases leads to lower alveolar oxygen pressure, which in turn inhibits the development of tuberculosis lesions.

TB incidences were also found to be higher in urban areas. The high number of cases in cities may be due to increasing poverty, migration, and homelessness in cities that seems to be linked with the reemergence of TB (Carolyn 1996). Associations among tuberculosis, urbanization, and poverty have been noted in studies from countries as diverse as India and the Philippines (Rangan et al 2003, Tupasi et al 2000). It is clear that growing numbers of poor, malnourished people living in unhygienic, overcrowded conditions can facilitate the transmission of TB in Nepal.

Our study had some limitations. We analyzed a short period of time (from 2003 to 2008). Additional analyses are needed to evaluate the trends of tuberculosis using data for a longer study period, or more detailed incidence data (monthly, quarterly). Second, we could not incorporate age, which is considered as the one of the risk factors for tuberculosis, due to unavailability of age-specific incidence data.

In conclusion, this study presents insights into the incidence of TB by gender, year and location. These findings require further investigation, but highlight the importance of selectively monitoring geographic locations and planning future intervention strategies.

Acknowledgements

Our study was partially funded by the Graduate School, Prince of Songkla University, Thailand. We would like to express our gratitude to the SAARC Tuberculosis and HIV/AIDS Centre (STAC), Bhaktapur, for permission to use their data. We are indebted to Prof. Don McNeil for supervising our research.

References

- Blumberg, H.M., Leonard, M.K.J. and Jasmer, R.M. 2005. Update on the treatment of tuberculosis and latent tuberculosis infection. *JAMA*, 293: 2776-84.
- Carolyn, S. 1996. Healthy Cities or Unhealthy Island? The Health and Social Implications of Urban Inequality. *Environ Urban*, 8: 9-30.
- Davies, P.D.O. 1997. Seasonality of tuberculosis. *Thorax*, 52: 398.
- Dolin, P. 1998. Tuberculosis epidemiology from a gender perspective. In: Diwan VK, Thorson A, Winkvist A (Eds). *Gender and Tuberculosis. Göteborg: Nordic School of Public Health*, 29-40.
- Douglas, A.S., Strachan, D.P. and Maxwell, J.D. 1996. Seasonality of tuberculosis: the reverse of other respiratory diseases in the UK. *Thorax*, 51: 944-6.
- Dye, C., Scheele, S., Dolin, P., Pathania, V. and Raviglione, M.C. 1999. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. *JAMA*, 282: 677-86.
- Dye, C., Watt, C.J., Bleed, D.M., Hosseini, S.M. and Raviglione, M.C. 2005. Evolution of tuberculosis control and prospects for reducing tuberculosis incidence, prevalence, and deaths globally. *JAMA*, 293: 2767-75.
- Harries, A., Maher, D. and Uplekar, M. 1998. *National Tuberculosis Programme of Nepal: A Clinical Manual*, National Tuberculosis Centre. Thimi, Bhakatpur.
- Holmes, C.B., Hausler, H. and Nunn, P. 1998. A review of sex differences in the epidemiology of Tuberculosis. *Int J Tuberc Lung Dis*, 2: 96-104.
- Legrand, J., Sanchez, A., Le, P.F., Camacho, L. and Larouze, B. 2003. Modeling the Impact of Tuberculosis Control Strategies in Highly Endemic Overcrowded Prison. *PLoS ONE*, 3: e2100.
- Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T. and Murray, C.J. 2006. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*, 367: 1747-57

- Mansoer, J.R., Kibuga, D.K. and Borgdorff, M.W. 1999. Altitude: a determinant for tuberculosis in Kenya? *Int J Tuberc Lung Dis*, 3: 156–61.
- Nunes, C. 2007. Tuberculosis incidence in Portugal: spatiotemporal clustering. *Int J Health Geogr*, 6: 30.
- Rangan, S., Gupte, H., and Bandiwadekar, A. 2003. Tackling tuberculosis in urban areas: experiences from Mumbai city. *Health Administrator*, 15: 72-9.
- Rios, M., Garcia, J.M., Sanchez, J.A. and Perez, D. 2000. A statistical analysis of the seasonality in pulmonary tuberculosis. *Eur J Epidemiol*, 16: 483–8.
- SAARC Tuberculosis and HIV/AIDS Centre (STC). 2007. *Tuberculosis Control in the SAARC Region - An update 2007*. STC Thimi, Bhaktapur, 36-40.
- Shrestha, K.B., Malla, P., Jha, K.K., Shakya, T.M., Akhtar, M. and Gunneberg, C. 2008. First national tuberculin survey in Nepal. *Int J Tuberc Lung Dis*, 12: 909–15.
- Shrestha, L. and Jha, K.K. 2007. Gender disparity among TB suspects & new TB patients-a record-based retrospective study in SAARC member states. *SAARC J. Tuberc, Lung Dis HIV/AIDS*, 4: 8-18.
- Theil, H. 1983. *Linear algebra and matrix methods in econometrics*. North-Holland Publishing Company, Amsterdam, Netherlands.
- Tongkumchum, P. and McNeil, D. 2009. Confidence intervals using contrasts for regression model. *Songklanakarin J Sci Technol*, 31(2): 151-6.
- Tupasi, T.E., Radhakrishna, S., Quelapio, M.I.D., Villa, M.L.A., Pascual, M.L.G., and Rivera A.B. 2000. Tuberculosis in the Urban Poor Settlements in the Philippines. *Int J Tuberc Lung Dis*, 4: 4-11.
- Uplekar, M.W., Rangan, S., Weiss, M.G., Ogden, J., Borgdorff, M.W. and Hudelson, P. 2001. Attention to gender issues in tuberculosis control. *Int J Tuberc Lung Dis*, 5: 220-224.
- Uthman, O.A. 2008. Spatial and Temporal Variations in Incidence of Tuberculosis in Africa, 1991 to 2005. *World Health Popul*, 10: 5-15.

- Vargas, M.H., Furuya, M.E., and Perez-Guzman, C. 2004. Effect of altitude on the frequency of pulmonary tuberculosis. *Int J Tuberc Lung Dis*, 8: 1321–4.
- Venables, W.N. and Ripley, B.D. 2002. *Modern Applied Statistics with S*. New York, Springer-Verlag.
- Venables, W.N., Smith, D.M. and the R Development Core Team. 2008. *An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics Version 2.6.2 (2008-2-08)*. Available at: <http://cran.r-project.org/doc/manuals/R-intro.pdf> (accessed February 2008).
- Werf, M.J. and Borgdorff, M.W. 2007. Targets for tuberculosis control: how confident can we be about the data? *Bull World Health Organ*, 85: 370-6.
- West, J.B. 1990. *Respiratory physiology. The essentials. 4th ed.* Baltimore, MD: Williams and Wilkins.
- Wikipedia. 2009. *The free encyclopedia 2009*. Available at: http://en.wikipedia.org/wiki/List_of_countries_by_population (accessed March 2009).
- Williams, B.G., Granich, R., Chauhan, L.S., Dharmshaktu, N.S. and Dye, C. 2005. The impact of HIV AIDS on the control of tuberculosis in India. *Proc Natl Acad Sci U S A*, 102: 9619-24.
- World Health Organization. 2003. *Gender and Tuberculosis Global Tuberculosis Control. WHO Report 2003*. Available at: www.who.int/entity/gender/documents/en/TB.factsheet.pdf (accessed January 2009).
- World Health Organization. 2008. *Global Tuberculosis Control: surveillance, planning, financing*. WHO report 2008 WHO/HTM/TB/2008.376 Geneva, Switzerland: World Health Organization.
- World Health Organization. 2004. *The top 10 causes of death. Geneva, Switzerland: World Health Organization*. Available at <http://www.who.int/mediacentre/factsheets/fs310/en/index.html> (accessed November 2008).

3.5 Manuscript 3

Spatial analysis in hospitalised injury morbidity in NSW, Australia

Shanley S.S.Chong¹, Sulawan Yotthanoo² and Rebecca Mitchell¹

¹Injury Risk Management Research Centre, UNSW, Australia

²School of Science and Technology, Naresuan University Phayao, Thailand

Abstract

Objective: To demonstrate the utility of a simple additive statistical model in accounting for the spatial, demographic and temporal patterns of unintentional injuries among children, as a first step in explaining child injury risk inequity.

Setting: Data were obtained on all hospital separations of NSW residents aged below 15 years who were unintentionally injured from 1 July 2000 to 30 June 2005.

Analysis: A generalized linear model was used to estimate injury incidence rates across local government areas in New South Wales, Australia.

Results: The incidence rates for males in all age groups were higher than those for females, with age-specific rates increasing for males and decreasing for females. Higher incidence rates were observed for children residing in rural and remote areas, and in the eastern suburbs of Sydney, and incidence rates were generally highest during summer and lowest during winter.

Conclusions: A simple generalized linear model containing age/gender group, location of residence and quarterly period as additive factors can provide a useful method for comparing injury risk, and has practical advantages over other methods that have been used in the literature. Such analysis provides useful information for further studies using socioeconomic indicators and leading to a better understanding of injury inequity, and better planning of injury prevention strategies within communities.

Keywords: Injury, count model, spatial analysis, inequity

Introduction

Locations with different environmental settings have different injury rates (WHO 2009). Exploring spatial patterns in injury risk may help to understand these differences. Spatial analysis studies the distribution and risk factors of disease and health-related events in relation to the geographical and environmental factors that determine the distribution (Waller and Gotway 2004). The increasing amount of geo-coded health and population data, combined with appropriate statistical methods, computing technology, and user-friendly geographic information systems, has facilitated investigations of spatial variation of injury risk (Braddock et al 1994, Poulos et al 2008, Bell et al 2008).

Many spatial analysis studies use Bayesian statistical methods (Vacchino 1999, Clements et al 2006, Zhu et al 2006, Law and Haining 2004). Lawson et al (2003) have provided a comprehensive description of these methods. However, despite their popularity among many statisticians, Bayesian methods are complex, computationally intensive, require special-purpose software, and do not provide conventional estimates of population parameters and their standard errors.

The aim of this paper is to illustrate the application of a conventional generalized linear model with additive determinants and to compare the results obtained with those based on a corresponding Bayesian model.

Methods

Data were obtained on all hospital separations of New South Wales (NSW) residents aged below 15 years who were unintentionally injured from 1 July 2000 to 30 June 2005. These data include information on inpatient separations of NSW residents from public and private hospitals, private day procedures and public psychiatric hospitals. They include data on episodes of care in hospital, which end with the discharge, transfer, or death of the patient, or when the service category for the admitted patient changes. The hospitalisation data were coded using ICD-10-AM (National Centre for Classification in Health 1998, National Centre for Classification in Health 2000, 2002, 2004). Injury-related hospitalised injuries were selected if they met the following criteria:

- the hospitalisation was for a patient resident in NSW;
- injury or poisoning was the principal diagnosis (ICD-10-AM range S00-T98);
- the external cause code was in the ICD-10-AM range V00-X59.

Hospitalizations relating to transfers and statistical discharges were excluded in order to partly eliminate ‘multiple counts’. These exclusions refer to transfers between hospitals or changes in the service category, such as a change from acute to rehabilitation for a patient during one episode of care in a single facility (Population Health Division 2004).

As geographic-based estimates may be unstable where there were small numbers of injuries due to a small population, local government areas (LGAs) with population less than 20,000 were combined with adjacent LGAs, and designated as super LGAs, as shown in the Appendix. In preliminary data analysis, injury rates were computed with respect to 17 statistical divisions of NSW created from the 11 statistical districts used by the Australian Bureau of Statistics and dividing the Sydney Metropolitan District into seven divisions defined by their geographic location (see Appendix).

Statistical analysis

A generalized linear model with a negative binomial distribution allowing for extra-Poisson dispersion (Venables and Ripley 2002, Chapter 7) was used to model the number of injury outcomes in the population. The data were classified by gender, age group, location of residence (104 super LGAs) and quarterly time periods, thus giving $2 \times 3 \times 104 \times 20 = 12,120$ data cells. If λ_{ijt} denotes the number of injuries for age group-gender combination i , geographic location j and quarter t , an additive model with this distribution is expressed as

$$\ln(\lambda_{ijt}) = \mu + \ln(P_{ijt}) + \alpha_i + \beta_j + \delta_t .$$

In this model, P_{ijt} is the corresponding population at risk in 100,000s and α_i , β_j and δ_t represent the gender-age, location (super LGA) and quarter effects. To check the fit of the model, deviance residuals were plotted against normal quantiles, and observed counts and appropriately scaled incidence rates were plotted against corresponding fitted values based on the model.

Sum contrasts (Venables and Ripley 2002, Chapter 6) were used to obtain confidence intervals for comparing the adjusted incidence rates for each factor with the overall incidence rate. Since these confidence intervals divide naturally into three groups according to their location entirely above the mean, around the mean, or entirely below the mean, this trichotomy was used to create thematic maps of super LGAs according to their estimated injury annual incidence rates.

In the corresponding Bayesian method the Besag-York-Mollie model was used (Lawson et al 2003, pages 123-124). This model is widely used in spatial statistics, taking into account both spatially correlated variation in rates in adjacent LGAs and uncorrelated variability in rates across areas. It gives the posterior distribution of the expected relative risk. The expected number of cases in each LGA was calculated by multiplying the overall age-gender-specific rate for NSW with the age-gender-specific population of the LGA, in order to adjust for the changes in age distribution across LGAs. The age-gender-specific hospital separation ratios were calculated for each LGA by comparing the number of injury cases in each LGA with the expected number. This expected number of cases was determined by multiplying the age-gender-specific rate for NSW by the age-gender-specific population of the LGA. For each LGA, the mean of the posterior distribution was taken as the best estimate of the smoothed hospital separation ratio. The 95% credible interval for the relative risk was used to compare the risk for each LGA with the NSW state average. The models were checked using the Gelman-Rubin diagnostic, and spatial autocorrelations in the residuals were examined using Moran's I statistic (Lawson et al 2003).

The R program (Venables and Smith 2004) was used for all conventional statistical analysis, graphs and maps, and the BRugs package (Thomas 2004) was used for the Bayesian analysis.

Results

There were 91,573 NSW residents aged below 15 who were admitted to hospital for an unintentional injury during the period 1 July 2000 to 30 June 2005. Of these, 63.6% were male. The rates were clearly lower for females, particularly for those aged 10-14 years. The rates were also higher in the western regions of NSW (i.e. Murrumbidgee and North-West districts) and lower in the inner-city and central suburbs of metropolitan Sydney. The quarterly rates varied from a minimum of 1022.3 per 100,000 in July-September 2002 to a maximum of 1508.2 in January-March 2001 (Table 1).

Age-gender group	Count	Rate	Quarter	Count	Rate
Male 0-4	16826	1518.5	Jul-Sep 2000	4402	1313.4
	17470	1519.9	Oct-Dec 2000	4833	1441.9
	23943	2048.7			
Female 0-4	12355	1178.7	Jan-Mar 2001	4745	1415.7
	11730	1074.6	Apr-Jun 2001	4704	1403.5
	9249	833.1	Jul-Sep 2001	4293	1279.1
			Oct-Dec 2001	4885	1455.5
Statistical division	Count	Rate			
01:Richmond	3,369	1365.4	Jan-Mar 2002	5062	1508.2
02:Mid-North	4,607	1556.1	Apr-Jun 2002	4511	1344.0
03:Hunter	7,069	1172.8	Jul-Sep 2002	3415	1022.3
04:Illawarra	5,766	1379.0	Oct-Dec 2002	4818	1442.2
05:South East	2,484	1218.1			
06:Murray	1,871	1534.2	Jan-Mar 2003	4983	1491.6
07:North West	2,990	1804.4	Apr-Jun 2003	4599	1376.7
08:Northern	3,005	1531.1	Jul-Sep 2003	4306	1295.8
09:Central West	3,351	1713.1	Oct-Dec 2003	4690	1411.3
10:Murrumbidgee	3,266	1887.9			
11:Sydney Outer	7,843	1499.7	Jan-Mar 2004	4917	1479.6
12:Sydney: Inner	2,283	1124.3	Apr-Jun 2004	4900	1474.5
13:Sydney North	7,803	1305.9	Jul-Sep 2004	4037	1216.0
14:Sydney West	16,126	1345.3	Oct-Dec 2004	4412	1328.9
15:Sydney South	11,433	1350.6			
16:Sydney	4,410	967.3	Jan-Mar 2005	4555	1372.0
17:Sydney	3,897	1525.7	Apr-Jun 2005	4506	1357.2

Table 1: Hospitalisation injury incidence rates per 100,000 in NSW by age-gender group, quarter and statistical division, 2000-01 to 2004-05

Figure 1 shows a plot of observed cell counts versus corresponding fitted values based on the negative binomial model, together with the plot of deviance residuals versus normal quantiles. The negative binomial model fits the data quite well as evident by the linear in the residuals plot. The comparison of the crude injury hospitalisation rates (Table 1) and the adjusted injury hospitalisation rates (Figure 1) suggests little evidence of confounding in these data.

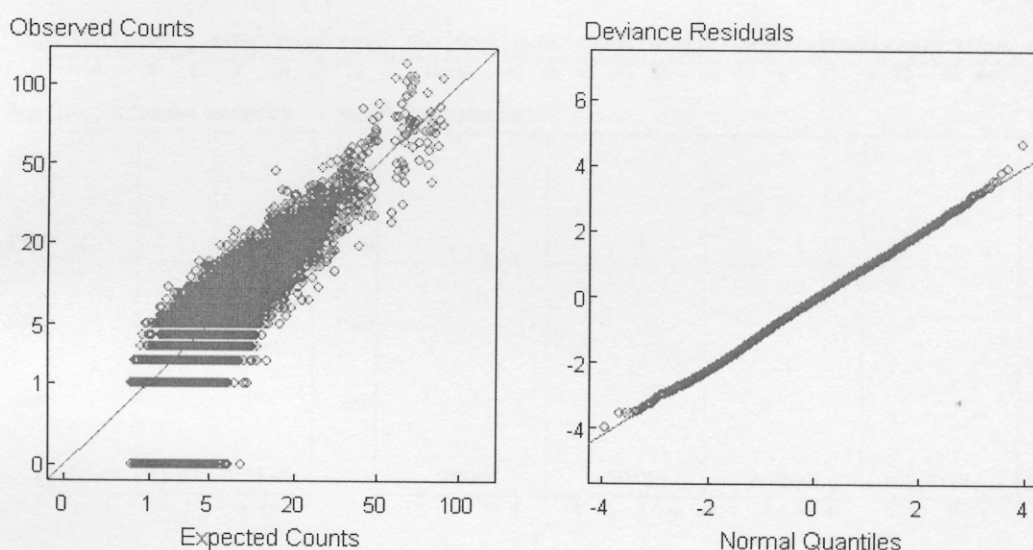


Figure 1: Plots of observed counts against fitted values (left) and deviance residuals versus normal quantiles (right).

Figure 2 shows 95% confidence intervals of annual injury-related hospitalisation incidence rate per 100,000 by super LGA (top panel), and sex and age group (bottom left panel), and year (right bottom panel), each adjusted for the effects of the other terms in the model. The vertical dotted lines separate the 17 divisions based on the statistical districts with the metropolitan district divided into the seven geographical regions. The incidence rates for males for all age groups were higher than those for females, with rates increasing by age group for males, and decreasing by age group for females. Also incidence rates were lower during July to September and higher during January to March.

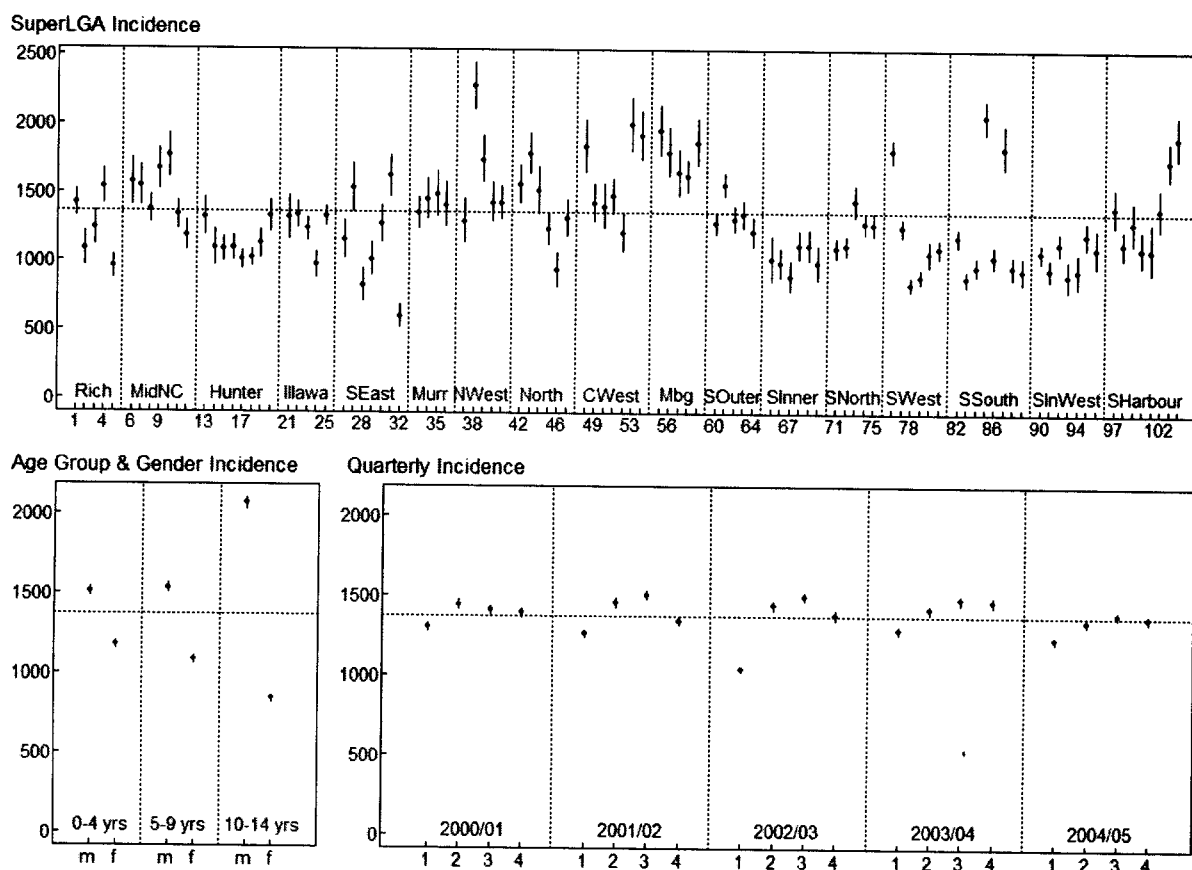


Figure 2: Injury-related hospitalisation incidence rate/100,000 by super LGA (top panel), and sex and age group (bottom left panel), and year (right bottom panel), each adjusted for the effects of the other terms in the model.

The upper panel of Figure 3 shows a thematic map of the adjusted annual injury-related hospitalisation rates in super-LGAs after fitting the generalized liner model, using the confidence intervals to classify them as above the mean (darkest shade), below the mean (light shade) or not evidently different from the mean (moderate shade). The lower panel of Figure 3 shows a similar map of the relative risk of injury in LGAs after using the Bayesian model. In both maps, injuries are higher in the remote and rural areas mainly in the western regions and also in some areas close to the coast of NSW.

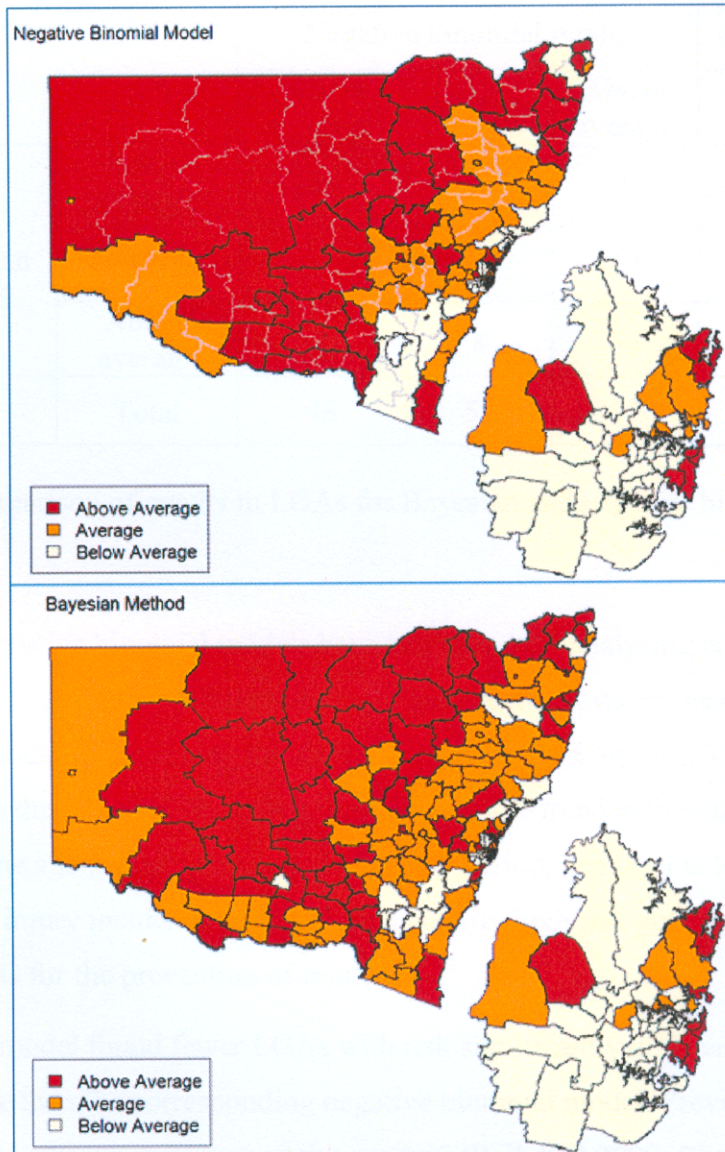


Figure 3: Thematic maps of adjusted annual injury-related hospitalisation rates in NSW with the insert of Sydney metropolitan areas. The upper panel shows rates for super-LGAs based on confidence intervals plotted in the upper panel of Figure 2, and the lower panel shows rates for LGAs based on an alternative Bayesian method.

Table 2 shows a cross tabulation of the LGAs based on the 95% credible intervals given by the Bayesian method and the corresponding 95% confidence intervals given by the negative binomial model. The two methods give similar results. However, there were 19 LGAs with above average injury rates using the negative binomial model, whereas using the Bayesian method only 8 LGAs were identified as having an above average injury risk.

		Negative binomial model			Total
		Below average	Average	Above average	
Bayesian model	Below average	41	1	1	43
	Average	4	43	19	66
	Above average	1	8	57	66
Total		46	52	77	175

Table 2: Comparison of results in LGAs for Bayesian and negative binomial models

Discussion

Poisson and negative binomial models have been used for analysing count data, though less commonly for spatial modeling. This paper reports on the effectiveness of using such a model in describing the spatial pattern of child injuries. This model is able to analyse the effect of age-group, gender and time trend at the same time. After adjusting for the age-group, gender, and quarterly period, geographic location was used to model injury incidence rates and thus identify high risk geographical locations as priority areas for the prevention of injuries.

The Bayesian model found fewer LGAs with risk significantly different from the overall average than the corresponding negative binomial model. Previous studies on spatial statistics of injuries have used this method (Bell et al 2009, Chong and Mitchell 2008, Poulos et al 2008). However, the Bayesian method is computationally intensive and time consuming (Carroll et al 2006).

The negative binomial model performs as well as the Bayesian model. Using a simple count model to distinguish areas with increased risk makes complicated data simpler to understand by stakeholders planning injury prevention program for different areas. In addition, it enables estimated rates to be calculated without lengthy or technically complicated procedures.

A limitation of this study is that the location of the incident is based on the location of residence of the hospitalised individual and it is possible that the injury may not have occurred near or in the home environment.

This study clearly shows that majority of the metropolitan areas have lower injury risk. However, most rural areas situated in the western part of NSW have higher injury risk.

Thus the patterns found in this study suggest socioeconomic status may help to explain the differences in injury risk and model will be fitted to adjust for this status to account for spatial variation we observe in this study.

Acknowledgements

The authors wish to thank the Centre for Epidemiology and Research at the NSW Health Department for providing access to the Health Outcomes and Information Statistical Toolkit (HOIST) analyzed in this study.

References

- Bell, N., Schuurman, N. and Hamed, S. M. (2008). Are injuries spatially related? Join count spatial autocorrelation for small-area injury analysis. *Injury Prevention*, 346-353.
- Braddock, M., Lapidus, G., Cromley, E., Cromley, R., Burke, G., and Banco, L. (1994). Using a Geographic Information System to Understand Child Pedestrian Injury. *American Journal of Public Health*, 1158-1161.
- Carroll, R. J., Ruppert, D. and Stefanski, L. A. (2006). *Measurement error in nonlinear models. A Modern Perspective*. Chapman and Hall/CRC.
- Clements, A. C., Lwambo, N. J., Blair, L., Nyandindi, U., Kaatano, G., Kinung'hi S., Webster, J. P., Fenwick, A. and Brooker, S. (2006). Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. *Tropical Medicine and International Health*, 490-503.
- Law, J. and Haining, R. (2004). A Bayesian approach to modelling binary data: the case of high-intensity crime areas. *Geographical Analysis*. 197-216.
- Lawson, A. B., Browne, W. J. and Vidal Rodeiro, C. L. (2003). *Disease Mapping with WinBUGS and MLwiN*. Wiley.

- Potter-Forbes, M. and Aisbett, C. (2003). *Injury costs: A valuation of the burden of injury in NSW 1998-1999*. Sydney: NSW Injury Risk Management Research Centre, University of NSW.
- Poulos, R., Hayen, A., Chong, S. and Finch, C. (2008). Geographic mapping as a tool for identifying communities at high risk of fire and burn injuries in children. *Burns*, 35: 417-424.
- Schmertmann, M. and Finch, C. (2004). *NSW Injury Profile: A Review of Injury Deaths During 1998-2002*.
- Thomas, A. (2004). *BRugs User Manual*, Version 1.0. University of Helsinki: Department of Mathematics and Statistics.
- Venables, W. N. and Ripley B. D. (2002). *Modern Applied Statistics with S*. New York, Springer-Verlag. Chapter 6.
- Venables, W. M. and Smith, D. M. (2008). The R development Core Team. *An Introduction to R: Notes on R: A Programming Environment for Data Analysis and Graphics Version 2.6.2 (2008-2-08)*. Available from: <http://cran.r-project.org/doc/manuals/R-intro.pdf>.
- Waller, L. A. and Gotway, C. A. (2004). *Applied spatial statistics for public health data*. Wiley-Interscience.
- WHO. (2009). *Accidents and injuries*. Available from: <http://www.who.int/ceh/risks/cehinjuries2/en/index.html>
- Zheng, B. (2005) Summarizing the Goodness of fit on Generalized Linear Models for Longitudinal Data. *Statistics in Medicine*, 19: 1265-1275.
- Zhu, Li., Gorman, D. M. and Horel, S. (2006). Hierarchical Bayesian spatial models for alcohol availability, drug “hot spots” and violent crime. *International Journal of Health Geographics*, 5: 54.
- Vacchino, M. N. (1999). Poisson regression in mapping cancer mortality. *Environmental research*, 1-17.

Appendix: Geographical Divisions

LGA	Div	SID	LGA	Div	SID	LGA	Div	SID	LGA	Div	SID
Ballina	1	1	Tallaganda	5	32	Narrabri	8	45	Hawkesbury	11	62
Byron	1	2	Yass	5	32	Quirindi	8	45	Wollondilly	11	63
Kyogle	1	3	Boorowa	5	33	Barraba	8	46	Wyong	11	64
Richmond Valley	1	3	Crookwell	5	33	Bingara	8	46	Ashfield	12	65
Lismore	1	4	Harden	5	33	Manilla	8	46	Drummoyne	12	66
Tweed	1	5	Young	5	33	Nundle	8	46	Leichhardt	12	67
Coffs Harbour	2	6	Albury	6	34	Parry	8	46	Marrickville	12	68
Grafton	2	7	Corowa	6	35	Uralla	8	46	South Sydney	12	69
Pristine Waters	2	7	Culcairn	6	35	Walcha	8	46	Sydney	12	70
Greater Taree	2	8	Holbrook	6	35	Tamworth	8	47	Baulkham Hills	13	71
Hastings	2	9	Hume	6	35	GlenInnes	8	47	Hornsby	13	72
Kempsey	2	10	Tumbarumba	6	35	Guyra	8	47	Ku-ring-gai	13	73
Copmanhurst	2	11	Berrigan	6	36	Severn	8	47	Pittwater	13	74
Maclean	2	11	Conargo	6	36	Tenterfield	8	47	Warringah	13	75
Bellingen	2	12	Deniliquin	6	36	Bathurst	9	48	Blacktown	14	76
Nambucca	2	12	Jerilderie	6	36	Blayney	9	49	Camden	14	77
Cessnock	3	13	Urana	6	36	Evans	9	49	Campbelltown	14	78
Great Lakes	3	14	Balranald	6	37	Oberon	9	49	Fairfield	14	79
Lake Macquarie	3	15	Murray	6	37	Rylstone	9	49	Liverpool	14	80
Maitland	3	16	Wakool	6	37	Cabonne	9	50	Penrith	14	81
Newcastle	3	17	Wentworth	6	37	Cowra	9	50	Bankstown	15	82
Port Stephens	3	18	Windouran	6	37	Forbes	9	51	BotanyBay	15	83
Dungog	3	19	BrokenHill	7	38	Bland	9	51	Canterbury	15	84
Gloucester	3	19	Dubbo	7	39	Weddin	9	51	Hurstville	15	85
Merriwa	3	19	Mudgee	7	40	Greater Lithgow	9	52	Kogarah	15	86
Murrurundi	3	19	Wellington	7	40	Orange	9	53	Randwick	15	87
Scone	3	19	Coolah	7	41	Lachlan	9	54	Rockdale	15	88
Muswellbrook	3	20	Coonabarabran	7	41	Parkes	9	54	Sutherland	15	89
Singleton	3	20	Gilgandra	7	41	Griffith	10	55	Auburn	16	90
Kiama	4	21	Narromine	7	41	Carrathool	10	56	Burwood	16	91
Shellharbour	4	22	Bogan	7	42	Hay	10	56	Concord	16	92
Shoalhaven	4	23	Bourke	7	42	Leeton	10	56	Holroyd	16	93
Wingecarribee	4	24	Brewarrina	7	42	Murrumbidgee	10	56	Parramatta	16	94
Wollongong	4	25	Central Darling	7	42	Coolamon	10	57	Ryde	16	95
Bega Valley	5	26	Cobar	7	42	June	10	57	Strathfield	16	96
Bombala	5	27	Coonamble	7	42	Narrandera	10	57	Hunters Hill	17	97
Snowy River	5	27	Walgett	7	42	Temora	10	57	Lane Cove	17	98
Yarrowumla	5	27	Warren	7	42	Cootamundra	10	58	Manly	17	99
Cooma-Monaro	5	28	Unincorporated NSW	7	42	Gundagai	10	58	Mosman	17	100
Eurobodalla	5	29	Armidale Dumaresq	8	43	Tumut	10	58	North Sydney	17	101
Goulburn	5	30	Inverell	8	44	Lockhart	10	59	Waverley	17	102
Queanbeyan	5	31	Moree Plains	8	44	Wagga Wagga	10	59	Willoughby	17	103
Gunning	5	32	Yallaroi	8	44	Blue Mountains	11	60	Woollahra	17	104
Mulwaree	5	32	Gunnedah	8	45	Gosford	11	61			

Chapter 4

Summary and Conclusions

This thesis has focused on using statistical methods to model incidence rates with application to diarrhea morbidity among young children in Thai Provinces bordering Cambodia, tuberculosis incidence rates for all persons classified by gender, location and years in Nepal, and injuries to young persons classified by gender, age group and local government area in NSW (Australia). This chapter summarizes the results and general conclusions, and suggests directions for further research by statisticians.

4.1 Summary of Study Results

Child diarrhea in Thai provinces bordering Cambodia

The original objective for this study was to investigate spatio-temporal patterns of hospital-reported diarrhea incidence for young children in districts of Thai provinces bordering Cambodia. However, the preliminary investigation revealed extensive under-reporting in the surveillance database. Because the data set that we used reported for other common diseases and the reported data for the same months and districts as for the diarrhea data were also zeroes, the evidence for under-reporting was even stronger than the preliminary analysis indicated. It should also be noted that infectious diseases in Thailand are substantially under-reported for two reasons. First, many people with the disease do not seek hospital treatment, and even when they do, hospitals do not always provide complete reports to the surveillance system lacking the information on patient enrolment and data collection in terms of validity and representative (Leelarasamee et al 2004). Thus the study actually had two objectives.

First, it was necessary to develop a method for imputing the under-reported data, and this became the primary objective of the published paper. The second objective was to subsequently use the imputed data to investigate the diarrhea incidence patterns.

For the first objective, a log-transformed linear regression model was used to estimate the extent of under-reporting of hospital-diagnosed cases of diarrhea. Age was not a factor in the model, and the data were aggregated as combinations of districts and months. Zero cases occurred in some districts where under-reporting was known or suspected. We omitted the zero cases before fitting an additive linear model to the log-transformed monthly incidence rates, and then deleted the cells corresponding to residuals below a specified cut-off value (-1.4). This value was chosen to satisfy the normality assumption in the residuals plot. The fitted model was then used to impute the omitted occurrences. Table 4.1 shows the number of cases thus imputed, by district and year. Districts with no under-reporting are in bold.

Table 4.1: Summary of under-reported cases categorized by district and year

Chanthaburi Province

District	Year						Average
	1999	2000	2001	2002	2003	2004	
1 Mueang Chantchaburi	13	0	0	0	0	0	2.1
2 Khlung	0	0	19	0	18	0	6.1
3 Tha Mai	0	26	0	0	0	0	4.4
4 Pong Nam Ron	0	0	0	0	0	43	7.1
5 Makham	6	0	13	0	0	0	3.3
6 Laem Sing	6	5	23	20	0	17	11.7
7 Soydow	0	0	0	24	0	12	5.9
8 Kaeng Hang Maeo	18	0	0	0	0	10	4.6
9 Na Yai Am	0	0	5	0	0	16	3.5
10 Nao Kichakut	6	0	9	0	0	20	5.8
Average	4.9	3.2	6.9	4.3	1.8	11.7	5.4

Table 4.1: (Continued)

Sa Kaeo Province

District	Year						Average
	1999	2000	2001	2002	2003	2004	
11 Mueang Sa Kaeo	0	0	0	0	0	0	0.0
12 Khlong Hat	0	0	0	21	0	0	3.5
13 Ta Phraya	0	0	0	187	156	151	82.3
14 Wang Nam Yen	13	0	0	16	0	0	4.8
15 Wattana Kakhon	0	0	0	0	16	0	2.7
16 Aranyaprathet	0	0	21	0	17	0	6.3
17 Khao Chakan	0	0	109	250	45	29	72.2
18 Khok Sung	0	3	0	49	3	11	10.9
19 Wang Sombun	8	15	0	28	35	22	18.1
Average	2.3	2.0	14.4	61.2	30.2	23.7	22.3

Buri Ram Province

District	1999	2000	2001	2002	2003	2004	Average
20 Mueang Buri Ram	0	0	462	414	47	0	153.8
21 Khu Muang	76	177	41	265	186	0	124.1
22 Krasang	0	0	0	0	39	0	6.4
23 Nang Rong	161	0	0	0	0	111	45.4
24 Nong Ki	0	0	29	184	0	0	35.5
25 Lahan Sai	0	60	28	298	0	0	64.4
26 Prakhon Chai	0	0	0	250	0	0	41.6
27 Ban Kruat	0	0	0	0	0	0	0.0
28 Phu Thai Song	0	0	10	72	0	7	14.9
29 Lam Plai Mat	0	0	0	181	0	0	30.1
30 Satuek	0	0	80	855	273	0	201.4
31 Pakham	0	0	0	171	79	103	58.9
32 Na Pho	0	101	0	0	19	103	37.2
33 Nong Hong	0	0	25	179	0	104	51.3
34 Phlapphlachai	0	0	0	243	0	36	46.5
35 Huai Rat	0	0	16	0	0	267	47.1
36 Non Suwan	0	0	102	0	24	151	46.1
37 Chamni	0	0	0	0	0	16	2.7
38 Ban Mai Chaiyaphot	65	100	107	99	0	0	61.9
39 Non Din Daeng	13	14	13	138	0	37	35.8
40 Ban Dan	0	27	0	64	0	0	15.1
41 Khaen Dong	0	7	22	78	0	0	17.6
42 Chalem Phra Kiet	49	0	34	273	80	77	85.3
Average	15.8	21.1	42.1	163.6	32.5	44.0	53.2

Table 4.1: (Continued)

Surin Province

District	Year						Average
	1999	2000	2001	2002	2003	2004	
43 Mueang Surin	0	0	290	0	0	0	48.3
44 Chumphon Buri	20	53	42	43	0	24	30.3
45 Tha Tum	42	231	436	0	0	237	157.8
46 Chom Phra	0	0	0	0	0	0	0.0
47 Prasat	0	0	0	0	0	0	0.0
48 Kap Choeng	1	0	125	0	66	0	32.1
49 RattanaBuri	0	0	0	186	298	0	80.7
50 Sanom	0	0	50	58	41	15	27.2
51 Sikhoraphum	0	0	116	0	113	0	38.2
52 Sangkha	0	0	0	0	90	0	15.1
53 Lam Duan	52	0	0	0	0	55	17.9
54 Samrong Thap	137	97	413	0	0	0	107.9
55 Buachet	28	50	381	32	0	0	81.7
56 Phanom Dong Rak	0	5	6	6	37	27	13.5
57 Si Narong	0	48	121	0	142	41	58.8
58 Kwao Sinarin	27	0	10	0	9	27	12.3
59 Non Narai	6	7	0	14	66	20	18.9
Average	18.4	28.9	117.2	20.0	50.8	26.3	43.6

Si Sa Ket Province

District	1999	2000	2001	2002	2003	2004	Average
60 Mueng Si Sa Ket	0	0	33	0	0	84	19.5
61 Yong Chum Noi	40	0	42	9	17	54	26.9
62 Kanthararom	0	239	657	164	84	58	200.4
63 Kantharalak	145	54	50	0	12	35	49.3
64 Khukhan	0	0	0	286	516	185	164.4
65 Phrai Bueng	22	0	0	75	53	53	33.8
66 Prang Ku	258	0	0	302	24	0	97.4
67 Khun Han	0	120	41	43	181	97	80.2
68 Rasi Salai	39	11	10	0	0	13	12.2
69 Uthumphon Phisai	12	16	42	164	13	61	51.4
70 Bung Bun	30	25	36	0	37	25	25.6
71 Huai Thap Than	0	83	205	0	42	38	61.3
72 Non Khun	57	145	216	149	32	38	106.2
73 Si Ratana	0	232	0	0	0	46	46.4
74 Nam Kling	174	13	70	224	196	67	124.0
75 Wang Hin	36	14	197	214	143	0	100.8
76 Phu Sing	26	160	99	40	37	0	60.4
77 Mueng Chan	76	54	51	19	78	19	49.5
78 Benchalak	16	16	52	43	120	10	42.6
79 Phayu	35	24	31	30	5	5	21.6
80 Pho Si Suwan	53	27	25	21	15	18	26.6
81 Sila Lat	17	40	72	4	44	10	31.3
Average	47.1	57.9	87.7	81.1	75.0	41.7	65.1

Table 4.1: (Continued)

Ubun Ratchathani Province

District	Year						Average
	1999	2000	2001	2002	2003	2004	
82 Mueang Ubun Ratchathani	0	764	0	0	0	0	127.3
83 Si Mueng Mai	25	19	0	57	0	0	16.7
84 Khong Chiam	77	54	0	0	0	114	40.8
85 Khuang Nai	130	0	0	0	0	238	61.2
86 Khemarat	184	360	0	139	0	0	113.8
87 Det Udom	61	453	338	0	0	75	154.5
88 Na Chaluai	0	12	0	11	109	93	37.5
89 Nam Yuen	0	76	72	81	0	133	60.3
90 Buntharik	135	306	142	25	23	74	117.5
91 Trakan Phut Phon	0	0	0	0	0	632	105.3
92 Kut Khaopun	41	0	0	0	0	0	6.8
93 Mueng SamSip	0	0	14	45	19	0	12.9
94 Warin Chamrap	0	0	120	0	0	0	20.1
95 Phibu Mangsahan	15	59	0	0	24	83	30.3
96 Tan Sum	52	30	127	54	0	18	47.0
97 Pho Sai	14	45	0	0	0	0	9.9
98 Samrong	77	86	105	0	0	33	50.0
99 Don Mot Daeng	17	20	0	85	48	81	41.7
100 Sirindhorn	0	254	124	0	0	0	63.0
101 Thung Si Udom	0	0	0	33	18	20	11.7
102 Na Yai	6	65	17	0	11	13	18.7
103 Na Tan	21	37	14	25	30	33	26.7
104 Lao Sua Kok	0	6	0	4	22	0	5.2
105 Sawang Weerawong	0	30	0	29	8	169	39.2
106 Nam Khun	11	33	11	10	4	16	14.2
Average	110.2	181.1	118.7	100.0	89.2	147.3	49.3

Next, we fitted a model to the under-reported rates using logistic regression. In this model the outcome was taken as the binary variable indicating whether or not under-reporting had occurred in cells comprising districts for each of the 24 quarterly periods over the 6 years. The factors in this model were taken as the 106 districts, four seasons (January-March, April-June, July-September, and October-December) and the six years. The four districts showing no evidence of under-reporting were omitted from the model.

Figure 4.1 shows the results after fitting this model as 95% confidence intervals for the percentages of under-reported data by season (left panel), year (middle panel) and district (right panel), with the first level for each factor used as the referent group. The dotted horizontal lines represent the overall percentage of under-reported cells (10.8%). The proportions of under-reported cells were highest in the fourth quarter (October-November). The percentages increased sharply after 1999, remained stable for the next three years, and then dropped substantially again in 2003-2004. The percentage of under-reported cases was higher than average in districts of SiSaket province, and lower than average in districts of Chanthaburi province

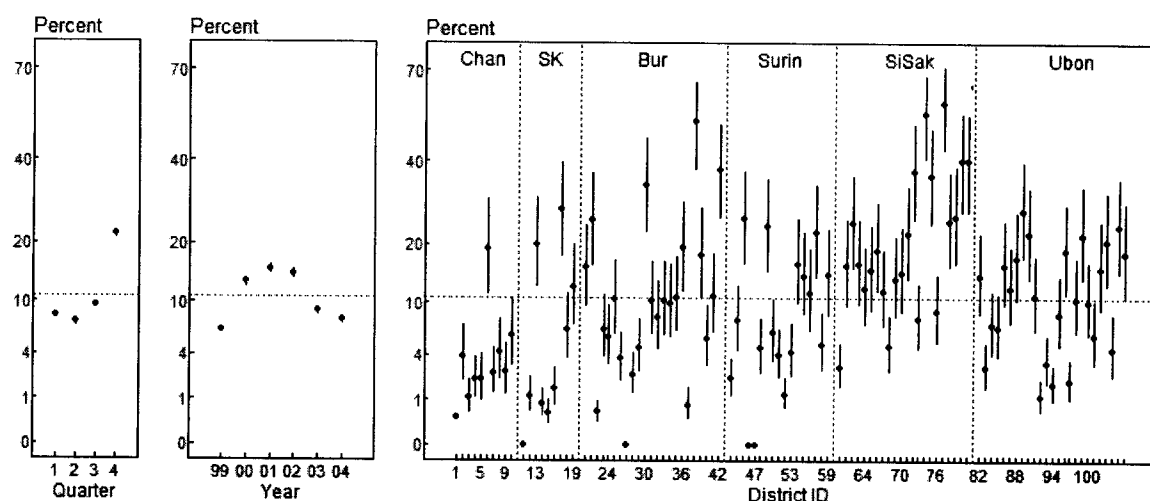


Figure 4.1: Results of fitting logistic regression model

For the second objective, after fitting the log-transformed linear regression model to the quarterly incidence rates, we examined the spatial correlation structure of the residual errors between pairs of districts. As expected, these correlations were quite low and mostly not statistically significant for districts in different provinces, but statistically significant correlations were found between districts in the same province. We used the GEE method to adjust for these within-province correlations, assuming a fixed correlation structure with the same correlation between all districts in the same

province, but allowing these common correlations (which ranged from 0.03 for Surin province to 0.12 for Si Sa Ket and Ubon Ratchathani, 0.13 for Sakaeo and 0.17 for Chantaburi and Buri Ram) to vary with province. To reduce correlations occurring in disease counts in successive months, we aggregated the data into quarterly incidence rates before fitting the model to the log-transformed rates.

The results (Figure 4 of the article 1) show that the diarrhea incidence rates were higher in the first quarter (January to March), with no trend except for slightly higher rates in 2004. There were pronounced differences between districts, with higher rates in Buri Ram, Surin and Si Sa Ket provinces.

Tuberculosis in Nepal

The tuberculosis data in Nepal were aggregated by year and the 75 districts were combined to 64 super-districts to achieve a more equal balance of populations in different regions. There was no evidence of under-reporting. Different models based on both log-transformed linear model and Poisson and negative binomial generalized linear models were considered, taking into account the need to provide a satisfactory fit to the data without an excessive number of parameters in the model.

This criterion proved difficult to meet with purely additive linear models. In the end we selected a nonlinear model with two multiplicative components similar to an extension of the Lee-Carter model used for mortality forecasting (Lee and Carter 1992, Booth et al 2002). The log-transformed model with assumed Gaussian errors fitted poorly compared with an equivalent negative binomial generalized linear model for which the fit was satisfactory.

The multiplicative components could be estimated as eigenvectors of a covariance matrix, and when these were smoothed using linear spline basis functions and used as fixed predictors it was possible to fit the negative binomial model straightforwardly.

The first basis function comprised a linear decreasing trend of TB incidence and the second function comprised a linear increasing trend during the first five years followed by a sharp drop in 2008. The decrease in trends of TB over this period may be attributed to successful TB control, case finding and treatment success in the recent years in Nepal.

Injuries in NSW

For this study comprising children's injury incidence rates in local government areas of NSW, generalized Poisson and negative binomial linear models with additive determinants were compared with the Bayesian model that the statisticians in the Department of Health had recommended. We first fitted a linear model to the log-transformed incidence rates with an appropriate modification to handle the zero counts. This method also incorporated GEEs with fixed correlation structure within the 17 larger divisions similar to that used in the first study. This model fitted reasonably well, but the GEE adjustment proved to be largely unnecessary because the spatial correlations were small, probably due to the fact that the district effects accommodated these correlations. So we then fitted a standard negative binomial generalized linear model and found that this fitted the data even better than the log-transformed linear model. The results obtained from this model were found to be very similar that those given by the Bayesian model.

4.2 Conclusions

The conclusions from our three investigations may be summarized as follows.

While much more study is needed with more extensive data sets, it would appear that appropriate methods for analyzing incidence rates already exist, and the new methods that continue to be developed in the statistical literature are largely unnecessary for most studies involving incidence rate outcomes. In particular, we found no reason to use generalized linear models other than negative binomial models, or generalized linear models with GEEs, zero-inflated models, generalized additive models, mixed models, Bayesian models, Markov chain Monte Carlo models, or any of the other complex models that have been developed. These models are often preferred for good reasons in the statistical literature, and have been found useful for many applications, but they have costs including computational complexity, lack of available software, and difficulty in explaining them to scientists with limited understanding of statistical theory.

Despite arguments that Gaussian models for transformed outcomes are inappropriate for analyzing incidence rates (see, for example Crawley 2005, page 125), we found that this model, with appropriate modification for handling zeros, fits incidence rates well in a wide variety of situations. Given that the Gaussian model is well understood and has been generalized to handle more complex situations including correlated errors with respect to both space and time, errors with non-stationary variances, weighted errors, multivariate outcomes, as well as non-linear models, we argue that it should not be too readily abandoned.

Therefore we suggest that future studies with incidence rate outcomes should seriously consider using an appropriate model with Gaussian errors as a basis for data analysis.

Transformed linear models could possibly be improved further by incorporating weights into the linear regression model, with the weights increasing with the number of disease counts in a cell (see, for example, Faraway 2006). Our reason for suggesting this is that un-weighted linear regression just deals with the incidence rates and does not take into account the number of cases, but it could be argued that the number of cases (not just the incidence rate) should determine the standard error of an estimated incidence rate.

References

- Ardkeaw, J. and Tongkumchum, P. 2009. Statistical Modelling of Childhood Diarrhea in the North-eastern border region of Thailand. *Southeast Asian J Trop Med Public Health*, 40(3): 807-815.
- Besag, J., York, J. and Mollie, A. 1991. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 3(1):1-20.
- Bishop, Y. M., Fienberg S. E. and Holland, P. W. 1975. *Discrete Multivariate Analysis*. Cambridge, Massachusetts: MIT Press.
- Booth H., Maindonald, J. and Smith, L. 2002. Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56(3): 325-336.
- Cameron, A. and Trivedi, P. 1998. *Regression analysis of count data*. Cambridge: University Press.
- Cheung, Y. B. 2002. Zero-inflated models for regression analysis of count data: a study of growth and development. *Statistics in medicine*, 21: 1461–1469.
- Crawley, M. 2005. *Statistics an introduction using R*, (1st ed.), John Wiley: New York.
- Davis, R., Dunsmuir, W. and Streett, S. 2003. Observation-driven models for Poisson counts. *Biometrika*, 90(4): 777-790.
- Demidenko, E. 2007. Poisson Regression for Clustered Data. *International Statistical Review*, 75(1): 96–113.
- Dormann, C. 2007. Effects of incorporating spatial autocorrelation into the analysis of species distribution data, *Global Ecology and Biogeography*, 16: 129-138.

- Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B. Schurr, F. M. and Wilson, R. 2007. Methods to account for spatial autocorrelation in the analysis of distributional species data: a review. *Ecography*, 30: 609–628.
- Evans, S. and Li, L. 2005. A comparison of goodness of fit tests for the logistic GEE model, *Statistics in Medicine*, 24: 1245–1261.
- Faraway, J. J. 2006. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (1st ed.). London: Chapman and Hall.
- Fienberg, S. E. 1980. *The Analysis of Cross-classified Categorical Data* (2nd ed.). Massachusetts: MIT Press.
- Greenland, S. 1989. Modeling and variable selection in epidemiologic analysis. *American Journal of Public Health*, 79(3): 340-349.
- Hastie, T. and Tibshirani R. 1990. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, 46: 1005-1016.
- Jansakul, N. and Hinde, J. P. 2004. Linear mean-variance negative binomial models for analysis of orange tissue-culture data. *Sonklanakarinn Journal of Science and Technology*, 26(5): 683-696.
- Kaewsompak, S., Boonpradit, S., Choonpradub, C. and Chaisuksant, Y. 2005. Mapping acute febrile illness incidence in Yala province. *Songklanagarind Journal of Medicine*, 23(6): 455-462.

- Kongchouy, N., Choonpradub, C. and Kuning, M. 2010. Methods for modeling incidence rates with application to pneumonia among children in SuratThani, Thailand. *Chiang Mai Journal of Science*, 37(1): 29-38.
- Lambert, D. and Roeder, K. 1995. Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Association*, 90: 1225-1236.
- Lawless, J. L. 1987. Negative binomial and mix Poisson regression. *The Canadian Journal of Statistics*, 15(3): 209-225.
- Lawson, A. B., Browne, W. J. and Vidal Rodeiro, C. L. 2003. *Disease Mapping with WinBUGS and MLwiN*. New York: John Wiley and Sons.
- Lee, H. S. 1996. Analysis of overdispersed paired count data. *The Canadian Journal of Statistics*. 24(3): 319-326.
- Lee, R. D. and Carter, L. R. 1992. Modeling and forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87(419): 659-671.
- Leelarasamee, A., Chupaprawan, C., Chenchittikul, M. and Udompanthurat, S. 2004. Etiologies of acute undifferentiated febrile illness in Thailand. *Journal of the medical association of Thailand*, 87(5): 464-472.
- Lewsey, J. D. and Thomson, W. M. 2004. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. *Community Dentistry and Oral Epidemiology*, 32(3): 183-189.
- Lim, A. and Choonpradub, C. 2007. A Statistical Method for Forecasting Demographic Time Series Counts, with application to HIV/AIDS and other Infectious Disease Mortality in Southern Thailand. *Southeast Asian Journal of Tropical Medicine and Public Health*, 38(6): 1029-1040.

- Maul, A., El-Shaarawi, A. H. and Ferard, J. F. 1991. Application of negative binomial regression models to the analysis of quantal bioassays data. *Environmetrics*, 2(3): 253-261.
- McCullagh, P. and Nelder, J. A. 1989. *Generalized Linear Models* (2nd ed.). Chapman and Hall, London.
- McNeil, D. R. and Tukey, T. W. 1973. Higher-order diagnosis of two-way tables, illustrated on two sets of demographic empirical distributions. *Biometrics*, 31: 487-510.
- Paul, S. and Saha, K. K. 2007. The generalized linear model and extensions: a review and some biological and environmental applications. *Environmetrics*, 18: 421-443.
- Poston, D. L. and McKibben, S. L. 2003. Using zero-inflated count regression models to estimate the fertility of U.S. women. *Journal of Modern Applied Statistical Methods*, 47(2): 371-379.
- R Development Core Team. 2008. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. [Cited 2008 Jun 28]. Available from: URL: <http://www.R-project.org>
- Ridout, M., Hinde, J. And Demétrio, C.G.B. 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, 57(1): 219-223.
- Sriwattanapongse, W. and Kuning, M. 2009. Modeling malaria incidence in North-Western Thailand. *Chiang Mai Journal of Science*, 36(3): 403-410.

- Sriwattanapongse, W., Kunning, M. and Jansakul, N. 2008. Malaria in North - Western Thailand. *Sonklanakarin Journal of Science and Technology*, 30(2): 207-214.
- Swennen, C., Sampantarak, U. and Ruttanadakul, N. 2009. TBT-pollution in the Gulf of Thailand: a re-inspection of imposex incidence after 10 years. *Marine Pollution Bulletin*, 58(4): 526-532.
- Theil, H.1983. *Linear algebra and matrix methods in econometrics*. North-Holland Publishing Company, Amsterdam, Netherlands.
- Thomas, A. 2004. BRugs User Manual, Version 1.0. University of Helsinki: Department of Mathematics and Statistics.
- Thurston, S. W., Wand M. P., and Wiencke, J. K. 2000. Negative Binomial Additive Models. *Biometrics*, 56: 139-144.
- Tiensonwan, M., Lertprapai, S., Sirichaisinthop, J. and Lawmepol, A. 2000. Application of log-linear models to malaria patients in Thailand. *Statistics in Medicine*, 19(14): 1931-1945.
- Tiensonwan, M., Yimprayoon, P. and Lenbury, Y. 2005. Application of log-linear models to cancer patients: a case study of data from the national cancer institute. *Southeast Asian Journal of Tropical Medicine and Public Health*, 36(5): 1283-1296.
- Tongkumchum, P. and McNeil, D. 2009. Confidence intervals using contrasts for regression model. *Sonklanakarin Journal of Science and Technology*, 31(2): 151-156.
- Ugarte, M. D., Ibanez, B. and Militino, A. F. 2004. Testing for Poisson Zero Inflation in Disease Mapping. *Biometrical Journal*, 46(5): 526-539.

Venables, W. N. and Ripley, B. D. 2002. *Modern Applied Statistics with S* (4th ed.).

New York: Springer-Verlag.

Warton, D. 2004. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, 16: 275-289.

Yan, J. and Fine, J. 2004. Estimating Equations for Association Structures. *Statistics in Medicine*, 23(6): 859–880.

Zeger, S. L. and Liang, K. Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42: 121-130.

Vitae

Name: Miss Sulawan Yotthanoo

Student ID: 4920330002

Educational Attainment:

Degree	Name of institution	Year of Graduation
B.Sc. (Statistics)	Chiang Mai University	1999
M.Sc. (Applied Statistics)	Chiang Mai University	2001

Work-Position and Address:

Lecturer in Department of Statistics, School of Science and Technology, Naresuan University Phayao, Thailand.

List of Proceedings and Publications:

Proceedings:

Yotthanoo, S. and Choonpradub, C. Diarrhea prevalence pattern in Thai provinces bordering Cambodia. The International Conference on Health and the Changing World on November 10-13 2008. Bangkok, Thailand.

Publications:

Yotthanoo, S. and Choonpradub, C. 2010. A statistical method for estimating under-reported incidence rate with application to child diarrhea in Thai provinces bordering Cambodia. *The Southeast Asian Journal of Tropical Medicine and Public Health*, 41(1): 203-214.

Kakchapati, S., Yotthanoo, S. and Choonpradub, C. 2010. Modeling tuberculosis incidences in Nepal. *Asian Biomedicine (Research, Reviews and News)*, 4(2): April issue.