2018

Spatial analysis of tuberculosis amongst teenagers in Kenya

Elvis Muriithi
Strathmore Institute of Mathematical Sciences (SIMs)
Strathmore University

Spatial Analysis of Tuberculosis amongst Teenagers in Kenya

Elvis Muriithi

A Research Project submitted to the Strathmore Institute of Mathematics (SIMS) In Partial Fulfillment of the Requirement for the Degree of Masters of Science in Statistical Sciences at the Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University Nairobi, Kenya.

June,2018

Abstract

Tuberculosis (TB) is the ninth leading cause of death worldwide and the leading cause from a single infectious agent, ranking above HIV/AIDS. Despite being preventable and curable, tuberculosis (TB) is the leading cause of death from infectious disease globally, with nearly 10 million people developing TB and 1.5 million people dying from TB in 2014. Among teenagers, it is a top fifteen cause of mortality. Much emphasis in terms of TB surveillance is given to adults especially people living with HIV.

This study provides a spatial model to determine the dynamics of TB epidemic among the notified teenage TB cases in Kenya over three years (2013, 2014 and 2015). Conditional autoregressive model emerges as the best model compared to generalized mixed model.

Our results show that TB incidence remained constant over time and there is no much variation by sex. We found significance pattern of TB epidemic by counties.

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name of Candidate: Elvis Muriithi

Signature: .....

Date: .....

.

Approval

Approval The thesis of Elvis Muriithi was reviewed and approved by the following:

Dr. Humphrey Njogu,

Senior Lecturer, Faculty of Information Technology,

Strathmore University

Dr. Joseph Orero,

Dean, Faculty of Information Technology,

Strathmore University

Professor Ruth Kiraka,

Dean, School of Graduate Studies,

Strathmore University

Dedication

This thesis is dedicated to:

My family and friends for all the support they gave me and being patient with me as I worked on this project.

Special dedication goes to my grandfather who ensured and encouraged me that enriching oneself with knowledge give you power to become the greatest among many.He also taught me that the best kind of knowledge to have is that which is learned for its own sake. I would never fail to dedicate this work to my grandmother who constantly kept me in check by enquiring how far I was with the project.

It is also dedicated to my lovely mother, who taught me that even the hardest task can be accomplished if it is done one step at a time.

To my uncles and aunts especially uncle Njagi who always tells me that hard work pays. His words has kept me going through all odds and true to his words God has immensely blessed me in so many ways.

Acknowledgements

First and foremost, I must acknowledge my limitless thanks to the Almighty God for all His blessings. I am cognizant of the fact that this work would have never become true without His guidance. I owe a deep debt of gratitude to our university for giving us an opportunity to complete this work.

I would also like to thank the Institute of Mathematical sciences whose steadfast support of this project was greatly needed and deeply appreciated.

Prof. Samuel Mwalili has been the ideal thesis supervisor. His sage advice, insightful criticisms, and patient encouragement aided the writing of this thesis in innumerable ways.

I also would like to express my wholehearted thanks to my family for their generous support they provided me throughout my entire life and particularly through the process of pursuing the master degree. Because of their unconditional love and prayers, I have the chance to complete this thesis

I want to acknowledge the National TB program through the leadership of Dr. Kamene for allowing me to use their data in this project.

I am also grateful to my fellow students who kept me in check from the start and for their never ending encouragement and support throughout the learning period.

Lastly, deepest thanks go to all people who took part in making this thesis real.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Tuberculosis (TB) is ranked among the leading cause of mortality and morbidity globally. Kenya is among the 20 high burden countries that contribute 80% of the Global TB burden and is ranked number 15 worldwide while in Africa, it is five. Among teenagers, it is a top fifteen cause of mortality. Much emphasis in terms of TB surveillance is given to adults especially people living with HIV. However teenagers with TB are given low priority in most national health programs falling into the neglected group in this epidemic. This research project presents a GIS approach to analyze the spatial dynamics of TB epidemic among the notified teenage TB cases en Kenya over three years (2013, 2014 and 2015). There is no study that has been done to explore these patterns of teenage tuberculosis in Kenya.

1.2 Problem Statement and Justification

Teenage Tuberculosis is rated among the top fifteen causes of mortality today worldwide. The importance of teenage TB needs to be emphasized due to the fact that this is a very mobile group in that if infected with the disease they are more likely to transmit to others. Also, TB if not treated can be a pool for future TB epidemics.

Although affirmative action to TB reduction has been portrayed, very little has been done in regard to understanding the burden of teenage TB as well as identification of hot spots. This is because the epidemic (amongst teenagers) has been given a low priority and

neglected with the perception that they rarely transmit the disease and due to the high correlation of HIV and TB therefore more emphasis is given to adults.

The Prevalence Survey conducted in 2015-16 estimated that 558 per 100,000 population had TB against the WHO estimates of 233 per 100,000 population meaning nearly 40% of the TB cases are missed in Kenya. The survey showed that 83% of TB cases were HIV negative. This suggests that interventions to control TB among People Living with HIV have been successful and a large burden of TB now exists among people not infected with HIV. Therefore there could be a possibility that these missed cases are among the teenagers.

Detection of spatial clustering will be useful in identifying higher risk areas, where surveillance and control need to be targeted.

This study examines the distribution of teenage TB at the counties and country at large. The findings of this study will provide vital information to the National TB program and can be used by policy makers in coming up with control measures to avert this future epidemic.

1.3 Objectives

1.3.1 Main Objective

To determine the spatial distribution patterns of notified teenage TB cases in Kenya over three years (2013-2015).

1.3.2 Specific Objectives

i. To determine the spatial trend across the three years

ii. To classify the counties according to teenage TB incidence

iii. To obtain the teenage TB hot spot map for Kenya by sex.

Chapter 2

Literature Review

2.1 History of Tuberculosis

It has been hypothesized that the genus Bacterium originated more than 150 million years ago (Haman, 1984), with archaeological evidence in Egypt, documented more than 5000 years ago, with typical skeletal abnormalities of tuberculosis, including characteristic Potts deformities (Zimmerman, 1979) (Cave, 1939, and in America (Daniel T. , 2000). Hippocrates understood that TB attacked mainly people between the age of eighteen and thirty-five (Hippocrates, 1982). Mycobacterium tuberculosis was originally isolated by Robert Koch in 1882. Chemotherapeutic agents were discovered which revolutionized TB management between 1943 and 1957 (A. Schwartz, 1944) (Daniel T. M., 2006).

2.2 Incidence of TB in Kenya

The last TB prevalence survey was conducted before independence in 1958-59 and therefore the true burden of Tuberculosis in Kenya has been pegged on estimates. The current estimated TB incidence in Kenya stands at 109/100,000 for those with TB and HIV, and 268/100,000 for the TB cases with and without HIV with a case detection rate of 75%. Mortality rate from TB with HIV is 21/100,000, while TB without HIV is 20/100,000 (WHO, Global Tuberculosis Report, 2015).

Due to social - economic factors and it nature of transmission, TB burden in Kenya is higher than previously thought. The findings of the survey done in 2016 Justified the above

claim and showed that 558 per 100,000 population had TB against the WHO estimates of 233 per 100,000. In 2015, 82,000 people were diagnosed with TB meaning 40% of the cases go undiagnosed. From the same survey, 138,105 (NTLD, Kenya Tuberculosis Prevalence Survey, 2016). people fall sick of TB in Kenya every year. This pool of missed cases continues to fuel the spread of TB, considering that one undiagnosed and untreated individual can infect 10-15 people.

Worldwide, an estimated two billion people are infected with Mycobacterium tuberculosis (CDC, 2006) and approximately two million die from TB annually (Dermot Maher, Global epidemiology of tuberculosis, 2005).

Notified TB cases provides a good indicator of TB incidence in countries that have good surveillance systems which capture TB cases at the lowest level of health care to the highest giving insight of testing trends in the country

TB affects all age categories. It mostly affects the productive age group which is 15-44 years (NTLD, 2016), and this is a fact since the time of Hippocrates (Hippocrates, 1982). Therefore TB is categorized among the top leading cause of deaths in Kenya.

2.3 Natural history of TB

In studies of the natural history of the disease among sputum smear-positive and HIV-negative cases of pulmonary TB, around 70% died within 10 years; among culture-positive (but smear-negative) cases, 20% died within 10 years. The duration of tuberculosis from onset to cure or death is approximately 3 years and appears to be similar for smear-positive and smear-negative tuberculosis (Edine W. Tiemersma, 2011).

When one gets TB Infection,it refers to the presence of TB bacilli in the body without any clinical signs and symptoms of the disease. People who get TB infection are at greater risk of developing TB disease at the initial stages of the infection which stretches up to two years.

When an infected person coughs, droplet nuclei containing tubercle bacilli are inhaled, then they enter the lungs and travel to the alveoli. In the alveoli, they multiply, with a small number of tubercle bacilli entering the bloodstream, thus spreading throughout the body triggering the immune system produces macroscopic that surround the bacilli containing and keeping them in control.

In case the immune system fails to keep the bacilli under control, they multiply rapidly moving from infection to TB disease.

If TB disease is left untreated, 50-60% of the cases die, 20-25% are spontaneously cured and 20-25% of the cases remain as chronic coughers within 5 years.

2.4 Challenges of TB

Drug resistant TB is a form of TB infection caused by bacteria that are resistant to treatment with at least two of the most powerful first-line anti-TB medications drugs (isoniazid and rifampin). Drug resistant TB has proven to be a major threat in the campaign to end TB in Kenya. Prolonged treatment of drug resistant TB lump together with poor treatment outcomes not forgetting the high cost of treating one case are among the key issues in the fight against TB in resource limited countries like Kenya(NTLD-P, 2013).

HIV has also contributed to the resurgence of TB in Africa but with the findings of the prevalence survey, the co-infection rate of TB-HIV stands at 17% meaning 83% of the TB cases in Kenya are HIV negative .

2.5 Spatial Analysis

GIS and Spatial Scan Statistic have been used to describe the spatial distribution of diseases. Detection of clusters has been useful is the regular surveillance of diseases, by identifying factors behind the spread of the diseases. Combination of GIS and spatial analysis has been used to analyze the spatial distribution of diseases in endemic areas, thereby identifying clusters (Hoeven TA, 2008).

Chapter 3

Methodology

3.1 Study area, Population and Data

The study was carried out at different counties in Kenya with the aim of giving a report of the spatial distribution patterns of notified teenage TB cases in Kenya. This factored in the ages between 13-19 years of cases reported in years 2013, 2014 and 2015. The data was obtained from the National TB program.

3.2 Visualization

The first step was to visualize the spatial characteristics of the data. This allowed for an appreciation of any patterns that might be present, identification of obvious errors, and the generation of hypotheses about factors that might influence the observed pattern. This was also important for communicating the findings using maps of a disease distribution.

Data aggregation for the three year was done which involved getting the total cases notified(total number of teenagers infected with tuberculosis) into a single value. This was then assigned a spatial location which in our study was a county.

Disease points were then expressed as a function of the population size to provide estimate of prevalence rate per unit area. Choropleth lap which shows information by 'filing' (coloring) each component area with color, providing an indication of the magnitude of the variable of interest which was then used as a means for visualizing this data.

3.3 Spatial model

The study was divided into $n = 47$ counties. County and area are used interchangeably to imply the same thing. Let θ_i denote the unknown relative risk for the ith county $(i = 1, ..., n)$. Furthermore, Y_i and E_i are the observed and expected cases of the disease in the ith area, respectively. The simplest measure to assess the risk of a disease is through the standardized mortality ratio (SMR), calculated for a given area i by $SMR_i = Y_i/E_i$, $i = 1, ..., n$. Values greater than one imply a risk of disease higher than expected, while below one indicate a risk lower than expected for the respective area. However, a low value of E_i can happen if the population of a certain place is too low or if the disease under study is rare, which implies a non-plausible high risk for the region in question.

To overcome this problem, Bayesian hierarchical spatial models can be adopted. Using a Bayesian approach, we can obtain the joint posterior distribution for process and parameters given data. Such models allow the use of covariates that can provide information on the risk of mortality, as well as a set of random effects that capture the dependence between neighbouring regions. This is often facilitated using Markov chain Monte Carlo (MCMC). A general formulation of this class of models is given by

$$\begin{aligned} Y_i \mid E_i, \theta_i &\sim \text{Poisson}(E_i\theta_i) \\ \log(\theta_i) &= \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \phi_i,\ i = 1, 2, ..., n \end{aligned} \tag{3.1}$$

where θ_i denotes the unknown relative risk for the ith area which is modeled by an intercept term α common to all regions, a set of p covariates $x_{i1}, x_{i2}, ..., x_{ip}$ and a random effect ϕ_i. The regression parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ and the intercept term α are assigned weakly informative Gaussian prior distributions.

In Poisson distribution, $Var(Y_i) = E(Y_i)$, although in practical cases $Var(Y_i) > E(Y_i)$ occur. This scenario is known as *overdispersion*. The random effect are included to model any overdispersion and/or spatial correlation present in the data. Overdispersion and spatial correlation may occur due to the presence of unmeasured risk factors, which thus cannot be included as covariates in the model.

For areal data, the spatial dependency among the areal units is taken into account through

the neighborhood structure for the areas (regular or irregular) of interest. Once such neighbourhood structure is defined, autoregressive models (analogous to discrete temporal series models) are adopted to model the spatial random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$. One of the most popular is the class of conditionally autoregressive (CAR) models which are a type of Markov random field model.

For geostatistical data, distance-based exponential or Matérn covariance functions are adopted. Other choices may be considered and easily implemented through the hierarchical modeling framework (Cressie and Wickle, 2011; Banerjee *et al*., 2015).

Similarly, for modeling spatio-temporal structure in the data, temporally correlated error terms may be considered. A common approach to model temporal correlation is through autoregressive models (e.g., see Musenge *et al*; 2013). However, more complex methods may be used such as those for dynamical spatio-temporal models (Cressie and Wikle, 2011).

3.3.1 Conditional Autoregressive(CAR) Models

Due to its flexibility, multivariate normal distribution is usually used to represent the joint distribution of the random effects. For a vector of correlated random effects $\boldsymbol{\phi}$, it may be assumed that:

$$\boldsymbol{\phi} \sim \text{MNormal}\,(\mu, \sum) \tag{3.2}$$

where μ is the mean vector and $\sum$ is the covariance matrix which describes the spatial dependence of the data.

Let n be the total number of areas in a study. In general, the covariance matrix $\sum$ is given by:

$$\sum = \sigma^2\, \boldsymbol{C}; \quad \boldsymbol{C} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{M} \tag{3.3}$$

where σ^2 is a general variance parameter, $\boldsymbol{I}$ is a $n \times n$ identity matrix, ρ is a parameter that measures the spatial dependence, $\boldsymbol{W} = \left[w_{ij}\right]_{i,j=1,..,n}$ denote the so-called spatial proximity matrix with $w_{ii} = 0$ and $w_{ij} = 1$ if the ith and the jth areas are neighbours (denoted $j \sim i$), and 0 otherwise, and and $\boldsymbol{M} = m_{ii}$ is a diagonal matrix. Two areas are said neigh-

bours if they share common border, although other ways of defining neighbourhoods can be adopted (Cressie, 1993).

In order to ensure symmetry of $\sum$, we should specify $m_{jj}w_{ij} = m_{ii}w_{ji}$ such that the matrix $\boldsymbol{C} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}\boldsymbol{M}$ is symmetric and positive-definite. Thus, the matrix of variances is positive-definite for $\rho \in (\rho_{\min}, \rho_{\max})$, where $1/\rho_{\min}, 1/\rho_{\max}$ are the smallest and largest eigenvalue of the matrix $\boldsymbol{W}$, respectively. In practice, we expect a positive spatial dependence. Thus, it is possible to limit the range of ρ at $(0, \rho_{\max})$. A value of $\rho = 0$ implies spatial independence.

For a CAR model, the full conditional distributions of ϕ_i given all the remaining components $\boldsymbol{\phi}_{-i} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_n)$ can be generally defined as:

$$\phi_i \mid \boldsymbol{\phi}_{-i} \sim \text{Normal}\left(\mu_i + \rho \sum_{j \sim i}^{n} w_{ij}(\phi_j - \mu_j), \sigma^2 m_{ii}\right) \tag{3.4}$$

A number of different conditional autoregressive prior models have been proposed in a disease mapping context. We describe here four that are most commonly used.

3.3.2 Intrinsic CAR model

The intrinsic conditional autoregressive model (named iCAR hereafter) was proposed by Besag *et al.*(1991). it is one of the simplest CAR models for the distribution of the random effects in hierarchical spatial modeling. This model is obtained from Equation (3.4) by letting $w_{ij} = 1/n_i$ if areas i and j are adjacent and 0 otherwise, $m_{ii} = 1/n_i$, n_i the number of neighbours to an area i and $\rho = \rho_{\max}$, which under these specifications of w_{ij} and m_{ii} implies $\rho = 1$. Here the covariance matrix $\boldsymbol{C}$ is not positive-definite and hence there is no proper joint distribution for $\boldsymbol{\phi}$, as it would be possible to add any constant to each ϕ_i without changing the distribution. This issue can be rectified by fixing a constraint such as $\sum_{i=1}^{n} \phi_i = 0$ which can be implemented numerically at each iteration of an MCMC algorithm used for model fitting.

Imposing such constraints, the random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)$ has a multivariate normal distribution with mean 0 and a singular variance covariance matrix $\sigma^2 \boldsymbol{D}^{-1}$ (where $\boldsymbol{D}^{-1}$ is

the generalized inverse of $\boldsymbol{D}$) with ijth element of matrix $\boldsymbol{D}$ is defined as

$$d_{ij} = \begin{cases} n_i, \text{ if } j = i \\ -1, \text{ if } j \sim i \\ 0, \text{Otherwise} \end{cases}$$

If we let $\mu_i = 0$ for each i, then the conditional distribution of ϕ_i is now expressed as:

$$\phi_i \mid \boldsymbol{\phi}_{-i} \sim \text{Normal}\left(\frac{1}{n_i}\sum_{j\sim i}^{n}\phi_j, \frac{\sigma^2}{n_i}\right) \tag{3.5}$$

The conditional expectation of random effect ϕ_i is the average of the effects of its neighbours. Its conditional variance depends on its number of neighbours n_i, so that if an area has many neighbours then its variance will be smaller.

The iCAR model has some disadvantages. The strength of spatial dependence between random effects is always considered maximum$(\rho = 1)$. Furthermore, the variance parameter σ^2 is used to capture both overdispersion and spatial dependence(Leroux *et al*., 1999).

3.3.3 Convolution model

To model the random effects, Besag *et al*.,(1991) also proposed another popular model known as the convolution model (named BYM hereafter) in which the random effect ϕ_i is decomposed into two parts; a spatially unstructured component to account for pure overdispersion and a spatially structured component to account for spatial dependence:

$$\phi_i = v_i + u_i \tag{3.6}$$

where $v_1, ..., v_n$ are the unstructured random effects, independently and identically distributed with a normal distribution. The random effects u_i have a spatially structured prior distribution with iCAR distribution.

BYM model presents some identifiability problems of the spatial and non-spatial effects. That is, each data point is represented by two random effects but only their sum $v_i + u_i$ is identifiable. In addition, Eberly and Carlin (2000) noted some convergence problems in

estimation, which can be affected by the choice of prior distributions for the parameters and even the chain chosen for the initial values.

3.4 Bayesian approach

Lets assume we have a statistical model $p(\mathbf{y} \mid `)$ for the vector of observed data y depending on a vector of unknown parameters `. In a Bayesian context the parameter ` is assumed to be random with prior distribution $\pi(`)$.Using Bayes theorem the posterior distribution of ` is given by:

$$p(\mathbf{y} \mid `) = \begin{array}{c} \frac{p(\mathbf{y}|`).\pi(`)}{\int p(\mathbf{y}|`).\pi(`)d`} \\ \infty p(\mathbf{y} \mid `).\pi(`) \end{array} \tag{3.7}$$

In general, the posterior distribution for complex statistical models is a high dimensional, not analytically and numerically tractable function. Markov chain Monte Carlo (MCMC) provides a method to generate approximate samples from the posterior distribution which can be used to approximate quantities like the posterior mean, mode etc. by its empirical counterpart.

In this section Bayesian approach is briefly summarized. For more information see Gilks *et al*. (1996) and Gelman *et al*. (2004). A good overview about Bayesian methods is also given by Dellaportas and Roberts (2003).

The two basic algorithms used in MCMC are the Gibbs sampler first introduced by Geman and Geman (1984) and discussed by Gelfand and Smith (1990) and the Metropolis Hastings sampler developed by Metropolis *et al*. (1953) and Hastings (1970). These algorithms are described in the following sections.

3.4.1 Gibbs Sampler

Suppose we want to sample from a posterior distribution $p(\mathbf{y} \mid `)$ where the parameter ` is divided into d components $` = (\theta_1, ..., \theta_d)'$ . The Gibbs sampler is based on the full conditional distributions denoted by $p(\theta_i \mid `_{-i}) := p(\theta_i \mid `_{-i}, \mathbf{y})$, where $`_i = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d)'$ denotes the vector of ` without the ith component. The algorithm of the Gibbs sampler pro-

ceeds as follows: nnnnn

1. choose a starting value $\grave{}^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})'$ and set $t = 1$

2. sample $\grave{}^{(t)} = (\theta_1^{(t)}, ..., \theta_d^{(t)})'$by

$$\theta_1^{(t)} \sim p(\theta_1 \mid \theta_2^{(t-1)}, ..., \theta_d^{(t-1)})$$
$$\theta_2^{(t)} \sim p(\theta_2 \mid \theta_1^{(t)}, \theta_3^{(t-1)}, ..., \theta_d^{(t-1)})$$
$$\vdots$$
$$\theta_d^{(t)} \sim p(\theta_d \mid \theta_1^{(t)}, , ..., \theta_{d-1}^{(t-1)})$$

3. increase t by 1 and return to step 2.

The Gibbs Sampler has gained considerable popularity, particularly in applications in medicine, where hierarchical Bayesian models are commonly applied (see, e.g., Gilks *et al*., 1993). This popularity is also due to the availability of software which allows its application in a variety of problems (e.g.,WinBUGS).

3.4.2 Metropolis Hastings (MH) Sampler

Metropolis-Hastings (MH) sampler is considered as the atom of the MCMC techniques, introducing the basic notions and different properties. From section 3.4.1, when we want to samples from the full conditional distributions we consider the Gibbs sampler . There are case where samples cannot be obtained directly and sampling becomes tedious This happens where the full conditionals do not belong to any standard distribution there we the Metropolis Hastings (MH) algorithm is considered.

Instead of sampling directly from the full conditional of θ, a candidate value $\widetilde{\theta}$ from an arbitrary proposal distribution $q(\widetilde{\theta},.)$ is drawn and accepted with a certain probability. Both samplers(Gibbs sampler and the MH sampler, updates the parameters component by component. However, for notational simplicity the MH algorithm is described for one single component in this section. The algorithm works as follows:

1. choose a starting value $\theta^{(0)}$ and set $t = 1$

2. – propose a candidate value $\widetilde{\theta} \sim q(\widetilde{\theta})$

   – accept $\widetilde{\theta}$ with probability $min\left\{1, \frac{p(\widetilde{\theta}|\mathbf{y})}{p(\theta^{(t-1)}|\mathbf{y})} \frac{q(\theta^{(t-1)},\widetilde{\theta})}{q(\widetilde{\theta},\theta^{(t-1)})}\right\}$ and set $\theta^t = \widetilde{\theta}$, otherwise set $\theta^t = \theta^{(t-1)}$

3. increase t by 1 and return to step 2.

3.4.3 MH versus Gibbs Algorithms

There are advantages and disadvantages to MH and Gibbs methods. The Gibbs Sampler provides a single new value for each ` at each iteration, but requires the evaluation of a conditional distribution. On the other hand the MH step does not require evaluation of a conditional distribution but does not guarantee the acceptance of a new value. In addition, block updates of parameters are available in MH, but not usually in Gibbs steps (unless joint conditional distributions are available). If conditional distributions are difficult to obtain or computationally expensive, then MH can be used and is usually available.

In summary, the Gibbs Sampler may provide faster convergence of the chain if the computation of the conditional distributions at each iteration are not time consuming. The MH step will usually be faster at each iteration, but will not necessarily guarantee exploration. In straightforward hierarchical models where conditional distributions are easily obtained and simulated from, then the Gibbs Sampler is likely to be favoured. Thus, in this thesis Gibbs Sampler is used.

3.5 Model Comparison and Goodness of Fit (GOF) Measures

After fitting a model, The next an important step is model choice and comparison. Deviance information criterion (DIC) does this in a bayesian way where it factors in the fit of the data to the model and the corresponding complexity of the model.

3.5.1 Deviance Information Criterion (DIC)

Assuming you have a probability model $p(\mathbf{y} \mid \grave{})$. The Bayesian deviance $D(\grave{})$, which is used as a measure for goodness of fit, is defined as:

$$D(\grave{}) = -2\log p(\mathbf{y} \mid \grave{}) + 2\log f(\mathbf{y})$$

where $f(\mathbf{y})$ is some fully specified standardizing term.

To measure the model complexity, you need to introduce the effective number of parameters pD defined by:

$$pD = E[D(\grave{} \mid \mathbf{y})] - D(E[\grave{} \mid \mathbf{y}])$$

$=$ posterior mean of the deviance $-$ deviance of the posterior means.

The sum of the posterior mean of the deviance and the effective number of parameters is what define the deviance information criterion (DIC) as shown below:

$$DIC = E[D(\grave{} \mid \mathbf{y})] + pD$$
$$= 2E[D(\grave{} \mid \mathbf{y})] - D(E[\grave{} \mid \mathbf{y}])$$

The above criterion was suggested by Spiegelhalter *et al*. (2002). The model with the smallest DIC is to be preferred according to this criterion . pD and DIC are easily computed using the available MCMC output by taking the posterior mean of the deviance to obtain $E[D(\grave{} \mid \mathbf{y})]$ and the plug-in estimate of the deviance $D(E[\grave{} \mid \mathbf{y}])$ using the posterior means $E[\grave{} \mid \mathbf{y}]$ of the parameter $\grave{}$. An information theoretic discussion of the DIC as criterion for posterior predictive model comparison is given in van der Linde (2005).

Chapter 4

Results and Findings

4.1 Teenage TB Cases Notified

A total of 20,754 TB cases were notified between 2013 and 2015. 54% of these were male are the remaining 46% were females. Across the three years, more males than females were notified.

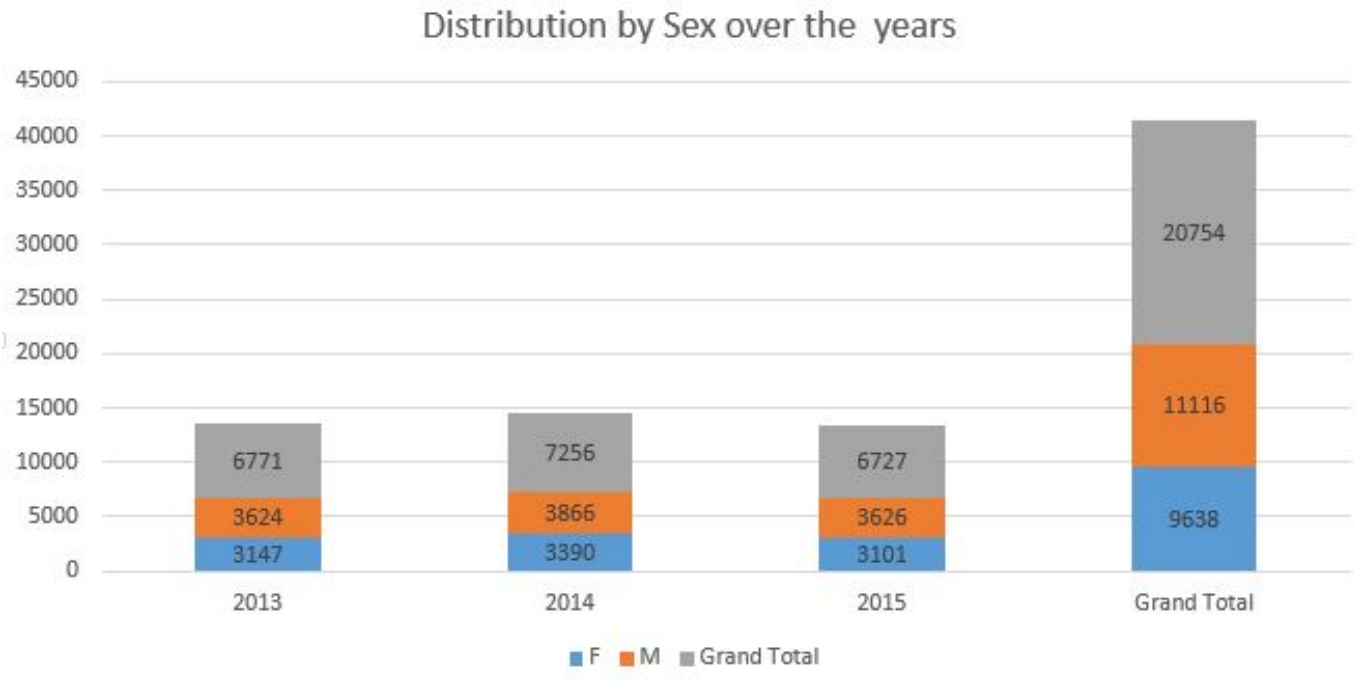

Figure 4.1: Distribution by Sex over the years

Table 4.1 shows the distribution of case per county and sex across the three years.

Table 4.1: Cases Notified by County

County	Sex	2013	2014	2015	Grand Total
Baringo	F	32	46	40	118
	M	37	42	41	120
Baringo Total		69	88	81	238
Bomet	F	93	113	109	315
	M	94	176	152	422
Bomet Total		187	289	261	737
Bungoma	F	91	86	70	247
	M	76	71	74	221
Bungoma Total		167	157	144	468
Busia	F	47	55	27	129
	M	53	55	37	145
Busia Total		100	110	64	274
Elgeyo Marakwet	F	23	25	29	77
	M	31	29	40	100
Elgeyo Marakwet Total		54	54	69	177
Embu	F	34	39	38	111
	M	47	37	41	125
Embu Total		81	76	79	236
Garissa	F	44	53	33	130
	M	92	81	70	243
Garissa Total		136	134	103	373
Homa Bay	F	106	118	72	296

Table 4.1: Cases Notified by County

County	Sex	2013	2014	2015	Grand Total
	M	89	95	76	260
Homa Bay Total		195	213	148	556
Isiolo	F	24	16	18	58
	M	41	27	34	102
Isiolo Total		65	43	52	160
Kajiado	F	61	81	65	207
	M	56	81	66	203
Kajiado Total		117	162	131	410
Kakamega	F	88	83	97	268
	M	71	79	75	225
Kakamega Total		159	162	172	493
Kericho	F	109	89	149	347
	M	137	202	177	516
Kericho Total		246	291	326	863
Kiambu	F	133	138	115	386
	M	109	142	117	368
Kiambu Total		242	280	232	754
Kilifi	F	54	51	52	157
	M	72	72	67	211
Kilifi Total		126	123	119	368
Kirinyaga	F	42	42	49	133
	M	38	39	36	113
Kirinyaga Total		80	81	85	246
Kisii	F	96	92	69	257
	M	83	84	67	234
Kisii Total		179	176	136	491
Kisumu	F	127	119	101	347

Table 4.1: Cases Notified by County

County	Sex	2013	2014	2015	Grand Total
	M	95	96	85	276
Kisumu Total		222	215	186	623
Kitui	F	50	63	75	188
	M	83	106	74	263
Kitui Total		133	169	149	451
Kwale	F	33	30	29	92
	M	42	32	38	112
Kwale Total		75	62	67	204
Laikipia	F	25	29	20	74
	M	36	28	32	96
Laikipia Total		61	57	52	170
Lamu	F	12	8	10	30
	M	15	10	11	36
Lamu Total		27	18	21	66
Machakos	F	74	75	63	212
	M	82	85	55	222
Machakos Total		156	160	118	434
Makueni	F	58	41	48	147
	M	53	54	52	159
Makueni Total		111	95	100	306
Mandera	F	27	32	26	85
	M	51	48	47	146
Mandera Total		78	80	73	231
Marsabit	F	44	44	49	137
	M	39	49	56	144
Marsabit Total		83	93	105	281
Meru	F	121	149	151	421

Table 4.1: Cases Notified by County

County	Sex	2013	2014	2015	Grand Total
	M	215	266	231	712
Meru Total		336	415	382	1133
Migori	F	73	103	83	259
	M	82	80	79	241
Migori Total		155	183	162	500
Mombasa	F	157	126	99	382
	M	152	171	142	465
Mombasa Total		309	297	241	847
Murang'a	F	55	64	46	165
	M	71	66	70	207
Murang'a Total		126	130	116	372
Nairobi	F	423	408	364	1195
	M	432	388	396	1216
Nairobi Total		855	796	760	2411
Nakuru	F	126	184	129	439
	M	173	192	179	544
Nakuru Total		299	376	308	983
Nandi	F	22	38	37	97
	M	42	36	28	106
Nandi Total		64	74	65	203
Narok	F	53	67	94	214
	M	100	114	144	358
Narok Total		153	181	238	572
Nyamira	F	45	36	41	122
	M	41	37	36	114
Nyamira Total		86	73	77	236
Nyandarua	F	29	31	41	101

Table 4.1: Cases Notified by County

County	Sex	2013	2014	2015	Grand Total
	M	24	34	35	93
Nyandarua Total		53	65	76	194
Nyeri	F	51	50	34	135
	M	49	26	49	124
Nyeri Total		100	76	83	259
Pokot	F	60	72	58	190
	M	78	97	107	282
Pokot Total		138	169	165	472
Samburu	F	18	34	42	94
	M	49	57	41	147
Samburu Total		67	91	83	241
Siaya	F	60	65	57	182
	M	38	57	48	143
Siaya Total		98	122	105	325
Taita Taveta	F	13	17	15	45
	M	19	19	9	47
Taita Taveta Total		32	36	24	92
Tana River	F	15	14	20	49
	M	22	24	24	70
Tana River Total		37	38	44	119
Tharaka Nithi	F	48	72	59	179
	M	67	72	56	195
Tharaka Nithi Total		115	144	115	374
Trans Nzoia	F	53	58	51	162
	M	53	67	60	180
Trans Nzoia Total		106	125	111	342
Turkana	F	66	106	111	283

Table 4.1: Cases Notified by County

County	Sex	2013	2014	2015	Grand Total
	M	129	89	134	352
Turkana Total		195	195	245	635
Uasin Gishu	F	60	79	72	211
	M	76	72	87	235
Uasin Gishu Total		136	151	159	446
Vihiga	F	51	30	19	100
	M	44	33	19	96
Vihiga Total		95	63	38	196
Wajir	F	21	19	25	65
	M	46	49	32	127
Wajir Total		67	68	57	192
Grand Total		6771	7256	6727	20754

4.2 Results: Application to TB Data

In this section the model is applied to the TB data. The model was fitted using the WinBUGS program. As shown in Table 4.2 Conditional Autoregressive (CAR) model was the best, with the lowest DIC value of 2398. Thus CAR model was parsimonious as compared to generalized linear random effects model.

Table 4.2: Deviance Information Criterion (DIC) for models comparison

Model	Deviance D	Effective parameters pD	DIC $= D + pD$
Random effects model	2358.340	48.202	2406.542
CAR model	2349.670	48.409	2398.080

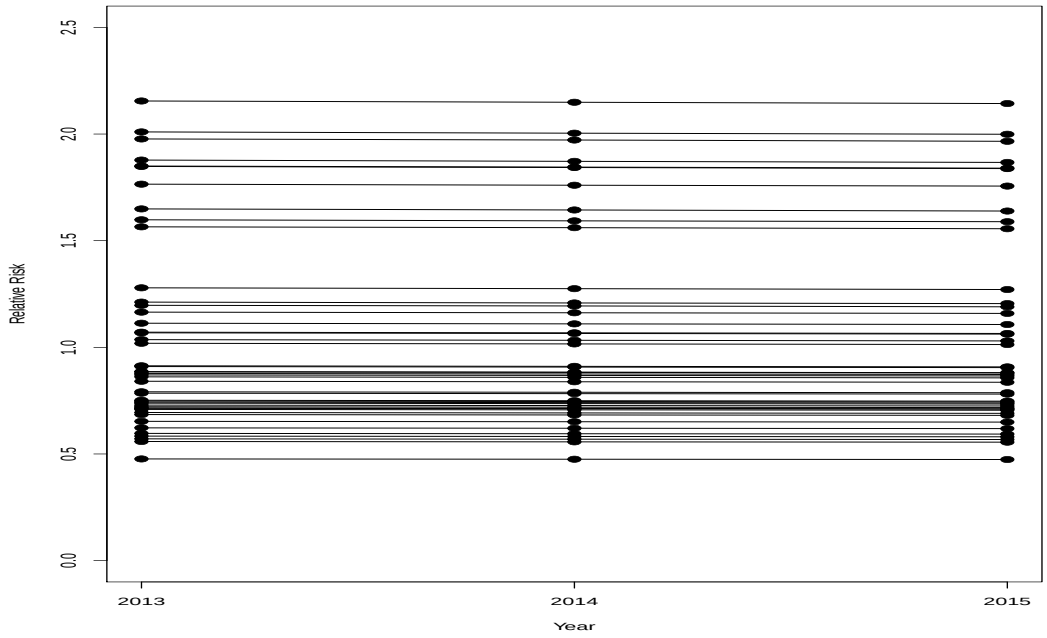

Figure 4.2: TB relative risk over time

As can be seen in Figure 4.2, TB incidence remained constant between 2013 to 2015, meaning there is no change in TB incidence over the three years. However, the figure shows clustering of counties into low, medium and high TB incidence. The highest TB incidence rate was reported in more arid counties such as Garissa, Isiolo, Lamu, Turkana and West Pokot as well as in most populated counties like Nairobi, Nakuru, Meru and Mombasa. This is displayed in Table 4.3.

Table 4.3: Classification of the Counties' TB incidence based on CAR model results

Low TB incidence $RR < 1.0$	Medium TB incidence D $1.0 \leq RR < 1.5$	High TB incidence counties $RR \geq 1.5$
Baringo	Homa Bay	Bomet
Bungoma	Kajiado	Garissa
Busia	Kisumu	Isiolo
Elgeyo Marakwet	Lamu	Kericho
Embu	Migori	Marsabit
Kakamega	Nairobi	Meru
Kiambu	Nakuru	Mombasa
Kilifi	Narok	Samburu
Kirinyaga	Turkana	Tharaka Nithi
Kisii		West Pokot
Kitui		
Kwale		
Laikipia		
Machakos		
Makueni		
Mandera		
Murang'a		
Nandi		
Nyamira		
Nyandarua		
Nyeri		
Siaya		
Taita Taveta		
Tana River		
Trans Nzoia		
Uasin Gishu		
Vihiga		
Wajir		

4.3 Discussion

Most counties with high relative risk are hard to reach. People living in these counties are pastoralists meaning they move from one place to another and this increases chances of transmitting the disease. High numbers of incidence could also be attributed to poor or inaccessible health care services. Therefore there is need to carry out assessment to distinguish these claim.

Counties like Tharaka Nithi and Kericho had more male case than female meaning the males are at a higher risk of contracting TB. This could be attributed by poor health seeking behaviors among men or due to occupational factors.

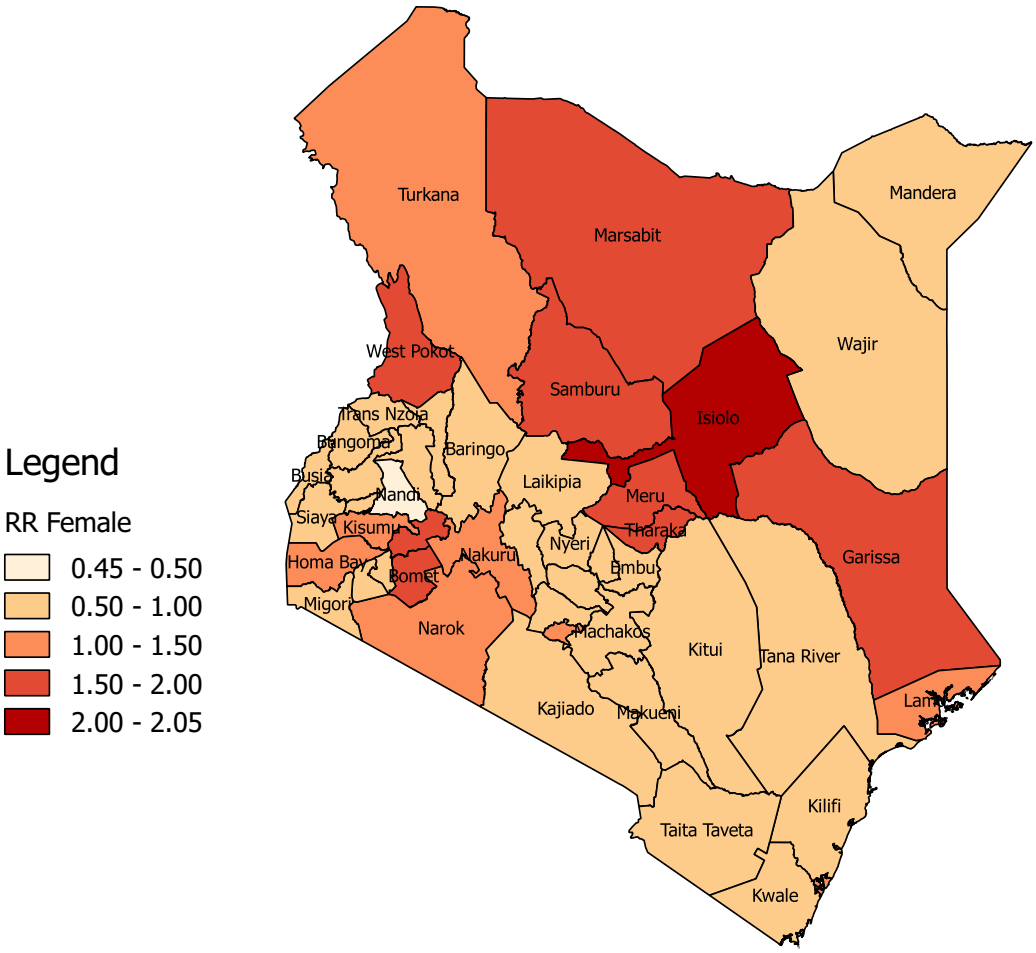

Figure 4.3: TB relative risk for female

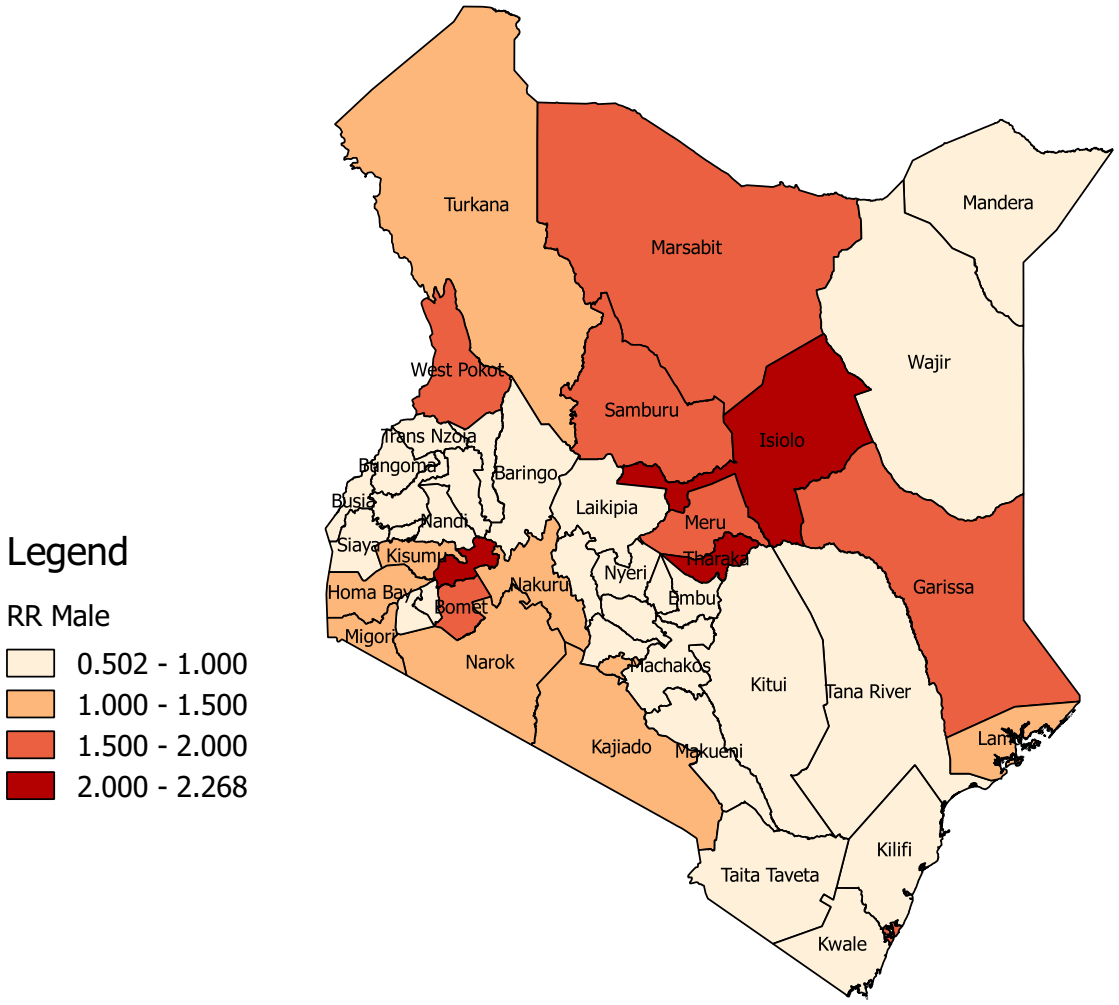

Figure 4.4: TB relative risk for male

Chapter 5

Conclusions and Recommendations for Further Research

5.1 Conclusions

TB incidents are not going down. There is need to intensify case finding especially in school where these teenagers spend most of their time. This will help identify TB case, manage them in order to prevent further transmission of the disease.

The key step to addressing the problem faced from tuberculosis to is the fact that there is no published estimates of tuberculosis incidence among adolescents and young adults in Kenya. This study is therefore important in providing the initial finding of adolescent TB rate in Kenya.

Increasing attention should be given to the challenges of management and prevention of tuberculosis in children and adolescents. This efforts should be geared at both county and national levels in order to address the wide gaps for implementation of fight against adolescent Tuberculosis.

5.2 Recommendations for Further Research

The work focused mainly at the county level analysis. However, it is likely adolescent TB in some counties may be concentrated or clustered at sub-county levels. A low-level granular analysis will be required. We also recommend exploration of various spatial models and use of semi-parametric approach.

REFERENCES

1. Besag J, York J and Molli´ e A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann Inst Stat Math.*, 43, 1–59.

2. NTLD-P: Kenya Tuberculosis Prevalence Survey, 2016 [http:/www.nltp.co.ke/tb-prevalence-survey-2015-2016/], Accessed 02 February 2018.

3. Best NS, Richardson S, Thompson A. (2005). A comparison of Bayesian spatial models for disease mapping. *Stat Methods Med Res*;14:35–59.

4. Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P., editors (2010). Handbook of Spatial Statistics. *Chapman & Hall.*

5. Gelman A., Carlin JB., Sternn HS. and Rubin DB. (2004). Bayesian Data Analysis; *2nd edn. Boca Raton: Chapman and Hall.*

6. Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transaction of Pattern Analysis and Machine Intelligence 6, 721–741.

7. Gilks, W., S. Richardson, and D. Spiegelhalter (1996). Markov chain Monte Carlo in Practice. *Boca Raton: Chapman & Hall/CRC.*

8. Gilks, W. R., D. G. Clayton, D. J. Spiegelhalter, N. G. Best, A. J. McNeil, L. D. Sharples, and A. J. Kirby (1993). Modelling complexity: Applications of Gibbs Sampling in medicine. *Journal of the Royal Statistical Society* B 55, 39–52.

9. Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

10. Dirk U. Pfeiffer et al Mark Stevenson, K. S. (2008). Spatial Analysis in Epidemiology . United States: Oxford University Press Inc., New York.

11. Society, A. T. (2012). Respiratory and Critical Care Medicine. Treatment of Tuberculosis

12. USAID. (2009). INFECTIOUS DISEASES. Retrieved January 16, 2016, from USAID: USAID Health: Infectious diseases, Tuberculosis, Countries, Kenya

13. NTLD-P. Tuberculosis Control in Kenya. Retrieved January 17, 2017, from www.nltp.co.ke

14. NTLD-P. (2013). Guidelines for Management of Tuberculosis and Leprosy in Kenya .Retrieved February 17, 2017, from www.nltp.co.ke

15. Kathryn J. Snow, Charalambos Sismanidis, Justin Denholm, Susan M. Sawyer, Stephen M. Graham (2017). The incidence of tuberculosis among adolescents and young adults: a global estimate.
European Respiratory Journal. 2018 51: 1702352; DOI: 10.1183/13993003.02352-2017

16. Wikipedia.www.wikipedia.org. [Online] Wikipedia, January 17, 2012. [Cited: January 19, 2012.] http://en.wikipedia.org/wiki/Tuberculosis.

17. Schiffman, George. Tuberculosis. http://www.medicinenet.com. [Online] July 22, 2011. [Cited: February 10, 2012.]
http://www.medicinenet.com/tuberculosis/page6.htm.

18. Phaisarn Jeefoo, Nitin Kumar Tripathi and Marc Souris (2010) Spatio-Temporal Diffusion Pattern and Hotspot Detection of Dengue in Chachoengsao Province, Thailand. Thailand: Environmental Research and Public Health.

19. Wikipedia. Spatial analysis. http://en.wikipedia.org. [Online] January 27, 2012. [Cited: February 10, 2012.] http://en.wikipedia.org/wiki/Spatial_analysis.

20. John D. H. Porter, (2002) Geographical information systems (GIS) and the tuberculosis DOTS strategy. http://onlinelibrary.wiley.com. [Online] A European journal TM & IH, January 5, 2002. [Cited: February 14, 2017.]
https://onlinelibrary.wiley.com/doi/full/10.1046/j.1365-3156.1999.00475.x

21. PK Moonan, (2004) Using GIS technology to identify areas of tuberculosis transmission and incidence. http://www.biomedcentral.com. [Online] International Journals of Health Geographics, October 13, 2004. [Cited: February 14, 2017.] http://www.biomedcentral.com/1476072X/3/23

22. Antonio L Rodrigues-Jr, A. R.-N. (2006). Spatial distribution of M. tuberculosis/HIV coinfection in São Paulo State, Brazil,1991-2001

23. Wikipedia. (2012, February 9). Moran's I. Retrieved April 12, 2012, from en.wikipedia.org: http://en.wikipedia.org/wiki/Moran's_I

24. CDC. (2006, March 24). Emergence of Mycobacterium tuberculosis with Extensive Resistance to Second-Line Drugs — Worldwide, 2000–2004. Morbidity and Mortality Weekly Report, 55(11), pp. 301-308. Retrieved from http://www.cdc.gov/mmwr/pdf/wk/mm5511.pdf

25. Hayman, J. (1984). Mycobacterium ulcerans: an infection from Jurassic time? The Lancet, 1015-1016

26. Zimmerman, M. (1979). Pulmonary and osseous tuberculosis in an Egyptian mummy. Bulletin of the New York Academy of Medicine, pp. 604-608.

27. Cave, A. (1939, July). The evidence for the incidence of tuberculosis in ancient Egypt. British Journal of Tuberculosis, 33(3), 142-152.

28. Daniel, T. (2000). The origins and precolonial epidemiology of tuberculosis in the Americas: can we figure them out? International Journal of Tuberculosis and Lung Disease, 4, 395-400.

29. Hippocrates. (1982). The aphorisms of Hippocrates with a Translation into Latin, and English (Special ed edition ed.). Birmingham: Classics of Medicine Library.

30. A. Schatz, E. B. (1944). Streptomycin, a substance exhibiting antibiotic activity against Gram-positive and Gram-negative bacteria. Proceedings of the Society for Experimental Biology and Medicine, 55, 66-69.

31. Edine W. Tiemersma, M. J. (2011). Natural History of Tuberculosis: Duration and Fatality of Untreated Pulmonary Tuberculosis in HIV Negative Patients: A Systematic Review.

32. Hoeven TA, F. E. (2008). Social distance and spatial distance are not the same, observations on the use of GIS in leprosy epidemiology. Epidemiology and Infection, 136, 1624–1627.

APPENDIX

A1: WinBUGS code for random effects model

```
model {
# Likelihood
for (i in 1 : N) {
tb[i]  ~ dpois(mu[i])
log(mu[i]) <- log(E[i]) + alpha0[sex[i]] + alpha1*(year[i]-2012) + b[county[i]]
}
for (i in 1 : 47) {
RRmale[i]   <-   exp( alpha0[1] + alpha1*(year[i]-2012) + b[i])
RRfemale[i] <-   exp( alpha0[2] + alpha1*(year[i]-2012) + b[i])
RR13[i] <-   exp( alpha0[1] *.5+ alpha0[2] *.5 + alpha1*(1) + b[i])
RR14[i] <-   exp( alpha0[1] *.5+ alpha0[2] *.5 + alpha1*(2) + b[i])
RR15[i] <-   exp(alpha0[1] *.5+ alpha0[2] *.5 + alpha1*(3) + b[i])
b[i] ~ dnorm(0,tau)
}


# Other priors:
alpha0[1]  ~ dflat()
alpha0[2]  ~ dflat()
alpha1 ~ dnorm(0.0, 1.0E-5)
tau  ~ dgamma(0.5, 0.0005)  # prior on precision
```

```
sigma <- sqrt(1 / tau) # standard deviation

}
```

A2: WinBUGS code for Conditional Autoregressive (CAR) model

```
model {
# Likelihood
for (i in 1 : N) {
tb[i]  ~ dpois(mu[i])
log(mu[i]) <- log(E[i]) + alpha0[sex[i]] + alpha1*(year[i]-2012) + b[county[i]]
}
for (i in 1 : 47) {
RRmale[i]   <-   exp( alpha0[1] + alpha1*(year[i]-2012) + b[i])
RRfemale[i] <-   exp( alpha0[2] + alpha1*(year[i]-2012) + b[i])
RR13[i] <-   exp( alpha0[1] *.5+ alpha0[2] *.5 + alpha1*(1) + b[i])
RR14[i] <-   exp( alpha0[1] *.5+ alpha0[2] *.5 + alpha1*(2) + b[i])
RR15[i] <-   exp(alpha0[1] *.5+ alpha0[2] *.5 + alpha1*(3) + b[i])
}

# CAR prior distribution for random effects:
b[1:47] ~ car.normal(adj[], weights[], num[], tau)
for(k in 1:sumNumNeigh) {
weights[k] <- 1
}

# Other priors:
alpha0[1]  ~ dflat()
alpha0[2]  ~ dflat()
alpha1 ~ dnorm(0.0, 1.0E-5)
```

```
tau  ~ dgamma(0.5, 0.0005)  # prior on precision
sigma <- sqrt(1 / tau) # standard deviation

}
```