Bayesian Statistical Frameworks and Quantitative Methods for the Analysis of Imperfect Global Health Datasets

By

Ermias Woldemariam Amene

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy
(Comparative Biomedical Sciences)

at the
UNIVERSITY OF WISCONSIN- MADISON
2015

Date of final oral examination: 08/26/15

The dissertation is approved by the following members of the Final Oral Committee:

Dörte Döpfer, Associate Professor, Medical Sciences
Charles Czuprynski, Professor, Pathobiological Sciences
Christopher W Olsen, Professor, Pathobiological Sciences
Ajay K Sethi, Associate Professor, Population Health Sciences
Bret Largat, Professor, Statistics

Abstract

Data generated from biomedical sciences often lack completeness. Valid inferences from such data need to be developed with great care. Imperfect datasets may contain too many variables given the number of limited observations, and a significant proportion of the data may also be missing. Restricting the analysis to complete information alone can compromise inferences as well. This study investigated a number of methods for analysis of imperfect datasets and geographic representation of risk of disease. Particularly, Bayesian approaches have been used throughout this thesis owing to their strength for analyzing imperfect data and disease mapping. To achieve the three overarching aims of this thesis, i.e., statistical analysis in the presence of missing data, variable selection, and disease mapping three separate datasets about topics from global health were used. These are (1) a WHO vital registration mortality dataset with the aim of predicting missing mortality rates associated with foodborne diseases; (2) The New Zealand campylobacteriosis notification and travelers' database, with the aim of estimating missing travel status of notified campylobacteriosis cases; and (3) The People, Animals and their Zoonoses Project dataset for variable selection and spatial description of the risk of malaria in the Busia region, Western Kenya. First, an important subset of predictors of mortality associated with foodborne diseases have been selected in a stepwise way using a combination of multiple imputation and elastic net regularization approaches. Cluster analysis and Bayesian hierarchical regression then were applied to estimate missing mortality rates for countries lacking the data. Second, missing travel status of campylobacteriosis notifications in New Zealand was estimated using a Bayesian regression model, and the risk of travel associated campylobacteriosis was mapped for New Zealand. Third, a meaningful subset of predictors of malaria was selected, and

the spatial distribution of the risk of malaria was mapped for the Busia region, Western Kenya. The implications of the results obtained can be generalized by expanding the methods to a broader and wider range of problems in addition to applying the methods to assist in the prioritized allocation of resources for the control and prevention of the diseases.

Dedication

I dedicate this thesis to my 75 year old beloved mother, Ehte Ourji, who does not know how to read and write, but strongly believes in education. Her unconditional love and prayer have sustained me throughout my life.

Acknowledgments

First and foremost, I would like to thank God, who has blessed me with patience, health and wisdom to make me who I am today.

I am very privileged to have performed my PhD research work at the School of Veterinary Medicine, University of Wisconsin-Madison, which has an incredibly collaborative environment; therefore, there are many individuals that contributed for my success.

First, I would like to express my sincere gratitude to my supervisor, **Prof Dörte Döpfer**, for the continuous support of my PhD study, for encouraging me to engage in several related research activities and for guiding me to grow as an independent scientist. Her understanding, patience, motivation, feedback, comments and suggestions have all enlightened my way in writing this dissertation. Her understanding and respect of our cultural differences and advice in both research and personal life have been priceless.

Besides my main supervisor, I would like to thank my PhD committee members: **Prof. Chuck Czuprynski, Prof. Chris Olsen**, **Prof. Ajay Sethi** and **Prof. Bret Largat**, for their valuable time, deep and insightful comments, for challenging questions that have widened my perspectives. I would also like to offer my enduring gratitude to the faculty and staff of the School of Veterinary Medicine. Their technical and administrative support has been invaluable to me. Especially I would like to thank **Tom Bennett** and all the IT supporting team of the SVM for helping me maintain my PC. I am particularly indebted to my lab mates, fellow graduate students at the Comparative Biomedical Sciences and at the Population Health Sciences programs who have inspired me to continue my work in this field.

I would like to express my profound regard and gratitude to **Prof Czuprynski** for paving the ways to get the funding for completing my PhD research, without which this achievement would not have been possible. My appreciation also goes to the CBMS program coordinator, **Kathryn Holtgraver,** who has always been keen to remind me of deadlines for course registrations, facilitating paperwork for certification, prelim exams and now my final defense.

I would also like to thank **Prof Peter Muir** for allowing me to use his computer for simulation whenever my PC was short of memory.

I would like to acknowledge the financial support provided to me from the National Institute of Health Ruth L. Kirschstein National Research Service Award Institutional Training Grant, and the UW-Madison Advanced Opportunity Fellowship for completing my PhD graduate study.

Last but not least, I would like to thank my family and friends for their continued support and encouragement. Words cannot express how grateful I am to my family: especially to my mother **Ehte Ourji**, to my sister **Roza Woldemariam**, to my brother-in-law **Tadesse Meskela** and to my fiancée **Tsionawit Melaku** for all of their encouragement and support throughout my life. Your spiritual support and prayer for me was what sustained me thus far.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike Information Criteria
AUC	Area Under the Receiver Operating Characteristic Curve
BGR	Brooks Gelman Rubin
BHM	Bayesian Hierarchical Model
BMA	Bayesian Model Averaging
BMS	Bayesian model Sampling and Averaging
BS	Brier Score
BYM	Besag York and Mollie
CA	Cluster Analysis
CAR	Conditional Autoregressive
CH	Structured Heterogeneity
CorA	Correlation Analysis
CPO	Conditional Predictive Ordinate
DHB	District Health Board
DIC	Deviance Information Criteria
ENR	Elastic Net Regularization
FA	Factor Analysis
FAO	Food and Agricultural Organization
FBD	Foodborne Disease
FERG	Foodborne Epidemiology Reference Group
GEMS	Global Environment Monitoring System
ICD	International Classification of Disease
IRB	Institutional Review Board
KEMRI	Kenya Medical Research Institute
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Linear Discriminant Analysis

LOOCV	Leave-One-Out Cross Validation
MCMC	Monte Carlo Marcov Chain
MAE	Mean Absolute Error
MAR	Missing At Random
MARA	Mapping Malaria Risk in Africa
MCAR	Missing Completely At Random
MICE	Multiple Imputation using Chained Equations
ML	Maximum Likelihood
MNAR	Missing Not At Random
NIH	National Institute of Health
PAZ	Peoples animals and their Zoonoses
PB	Percent Bias
PCA	Principal Components Analysis
PIP	Posterior Inclusion Probabilities
PMP	Posterior Model Probabilities
SD	Standard Deviation
SL	Sub-location
SMR	Standardized Mortality/Morbidity ratio
SW	Silhouette Width
UH	Unstructured Heterogeneity
UPGMA	Unweighted Pair Group Method with Arithmetic mean
WHO	World Health Organization

List of Appendices

Chapter 1: Introduction and Literature review

1.1 Introduction

Datasets generated in biomedical research are often plagued by numerous problems that can hamper the full usage of available information. Such problems include missing data and the presence of too many variables and too few observations, called a wide-problem. Throughout this thesis these challenges are regarded as imperfect data problems. The focus of this thesis is to tackle these problems by the application of a combination of Bayesian methods, and other quantitative approaches, to real-world global health datasets. This thesis is organized into three parts. Chapter 1 lays a framework and motivation for the thesis. It includes the background and context of the study, as well as the scope and objectives of the project. The second section of Chapter 1 describes a general background information about Bayesian and classical methods for data analysis, missing data, variable selection and disease mapping with the emphasis on the Bayesian perspective. The second part of this thesis is composed of four distinct and self-contained chapters (Chapters 2-5). This latter section is built on the analysis of three real world global health datasets that address the problems of missing data and variable selection. Chapter 6 of this thesis comprises a conclusion of the topics discussed in the preceding chapters and future directions for research.

1.1.1 Background and Context

The importance of the field of data analysis is ever increasing as evidenced by the large number of publications regarding new methods and computational tools. Public health epidemiology research frequently generates huge numbers of datasets that are, by their nature, often incomplete

[1, 2]. Conventional analysis of these imperfect datasets usually leads to flawed inferences [3, 4]. This can impact public health decisions because inferences are directly related to the analytic approach. In addition to missing or incomplete information, datasets containing large numbers of variables have hindered the quantification of associations between risk factors and disease. Nonetheless, efforts to improve global health predominantly rely on successful and accurate reporting of information about the status of diseases and their associated factors. Identifying and tackling the sources of the infections and quantifying the underlying aggravating factors significantly helps reduce the disease burden. Lack of resources combined with imperfect data demand efficient quantitative statistical tools to make best usage of the available information while acknowledging data gaps.

Conventional methods for handling missing data are often questionable in terms of results. Most of these methods (and most standard software programs for data analysis) simply ignore the missingness by default and restrict the analysis to complete cases [5]. The major drawback of excluding missing observations from the dataset is that it frequently removes large proportions of the observations resulting in loss of statistical power due to inefficient use of available information [6]. Even under the best of conditions, these methods typically yield biased parameter or biased standard error estimates[7]. On the other hand, selecting the best subset of variables from a pool of potential predictors is one of the hardest and most important problems in biomedical research [8]. Traditional regression methods for subset selection such as stepwise, forward and backward elimination often encounter a problem whenever the number of variables are much larger than the number of observations [9]. Meanwhile, datasets generated in biomedical research may contain several hundred or thousands of variables and limited number of observations; therefore, the

classical methods can no longer be applied [10, 11]. Over the past few decades, penalized likelihood methods, such as the Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net algorithms have been developed to meet the challenges of high dimensional data through data mining (introduced in section 1.2.3). These regularized methods have been used to simultaneously select meaningful predictors and estimate their associations with outcomes [10, 12].

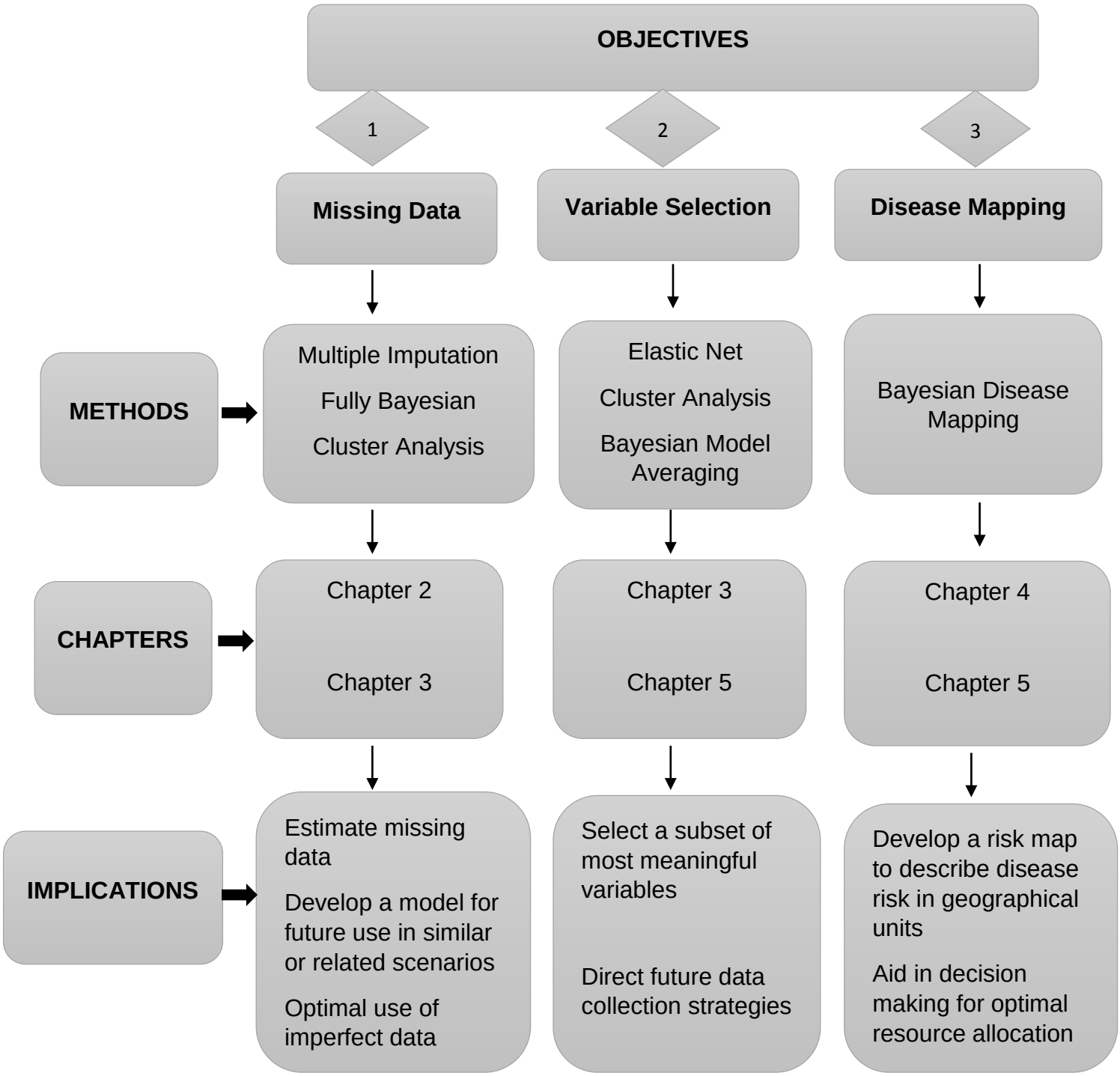
To cope with the challenges of imperfect data, Bayesian statistical methods have been widely used for missing value prediction, disease mapping in time and space, and for source attribution to make optimal usage of imperfect data [13, 14]. The application of Bayesian frameworks to a variety of problems in a number of disciplines has been increasing in recent years. This is mainly due to improved computational capacity that has resulted in increased application of Bayesian methods. These frameworks incorporate existing information (e.g., previous research findings or expert opinion) into the data, which is particularly important when imperfect data are used for inference about health outcomes [13, 15–17].

1.1.2 Scope and Objectives

It is vital to seek and apply methods that make optimal use of imperfect data without compromising inference. This thesis therefore attempts to address three overarching points. Those are missing data, a wide-data problem (variable selection) and spatial description of risk of disease, i.e. disease mapping. The general objective of this research is to apply various Bayesian frameworks and other quantitative approaches to tackle imperfect data problems. The specific objectives of this thesis are **(1)** to apply variable selection methods for high dimensional data (Chapter 1) and (Chapter 5) **(2)** to apply Bayesian and other quantitative methods for analysis of missing data (Chapters 1 and Chapter 3) and **(3)** to describe the spatial and temporal distribution of risk of disease (Chapters 4

and 5). In order to achieve the objectives mentioned above, this thesis explored real-world global health datasets having one or more of these challenges (see Figure 1 below for the schematic representation of the study).

Figure 1.1 Schematic representation of the objectives, methods and implications of the study



1.1.3 Overview of dissertation

This dissertation thesis is organized into six chapters. The first chapter is an introduction to the problem of missing data, variable selection and disease mapping, and it lays out a background for the study and describes imperfect data and their implication for inferences. This chapter reviews sources and types of missing data, commonly used statistical analysis approaches for imperfect data including frequentist and Bayesian frameworks, and introduces the concept of disease mapping. Chapters 2 to 5 are self-contained with their own introduction, methods, results, discussion and conclusions. In Chapter 2, Bayesian frameworks and other quantitative approaches including Multiple Imputation (MI), Cluster Analysis and *Elastic net* regularization methods are implemented for analyzing the World Health Organization's (WHO) Vital Registration dataset regarding mortality associated with foodborne disease. Chapter 3 deals with the application of fully Bayesian modeling and MI methods for predicting missing travel status of notified campylobacteriosis cases in New Zealand. Chapter 4 and 5 focus on Bayesian disease mapping approaches to characterize the spatial and temporal risk distribution of diseases in different geographical regions. While Chapter 4 describes the spatio-temporal distribution of the risk of travel associated campylobacteriosis in New Zealand, Chapter 5 is about the spatial analysis of malaria prevalence in rural homesteads of the Busia region, Western Kenya. Chapter 6 summarizes the findings and derives relevant conclusion and limitations of the study as well as future directions of research.

All the datasets used in this thesis consist of one or more of the problems. Bayesian methods are employed in all the datasets either for handling missing data, for variable selection or for disease mapping.

1.2 Imperfect data

Efficient and valid utilization of information derived from datasets depends on the quality of the information contained and the statistical tools applied. The quality of data is characterized by consistency, completeness, and number of meaningful variables and observations contained in the data [18]. Lack of all or either of these will result in imperfect datasets. Imperfect datasets are ubiquitous across a number of disciplines. The following sections describe shortcomings of datasets and some of the statistical analysis approaches for resolving the consequences of such shortcomings. For the sake of description in this thesis, the term "imperfect data" is classified into two categories, namely missing data and multidimensional data (i.e., the "wide data" problem or a dataset containing many more variables than observations). The following sections discuss missing data (1.2.1), multidimensional data (1.2.2), applicable statistical analysis approaches for imperfect data (1.2.3) and disease mapping (1.2.4).

1.2.1 Missing data

Given the increasing number of computational tools and the high number of incomplete datasets, missing data analysis has been the focus of statistical research in recent years [19–22]. Missing data are ubiquitous in all academic fields including biomedical, social and behavioral sciences, economics and machine learning [23–27]. Any research almost inevitably generates incomplete datasets with missing variables in the dataset or lack of variables per observation. There are several reasons, known and unknown, that can result in missing data. A few examples include: study participants may refuse to complete or skip part of a questionnaire survey, subjects may drop out of the study for a variety of reasons, researchers may not be able to collect complete information

(e.g. bad weather, sickness etc.), data collecting equipment may malfunction and data may be erroneously recorded [5, 28]. Understanding the reason for missingness is crucial, because it determines the choice and performance of the method for analyzing the dataset [29]. In other words, identifying the reason by which part of the data are missing can help the data analyst to select the optimal method for analyzing the imperfect dataset [30]. It has been previously stressed that the reason for missingness is sometimes more important than the amount of missingness in the dataset [31]. In addition, inference using incomplete data should take into account some form of assumption regarding the missing values [32]. According to Rubin and colleagues (1975), there are three categories of missing data with regards to the mechanism by which the missingness was introduced [33, 34]. These are Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). These mechanisms of missing data describe the relationship between measured variables and the probability of unobserved or missing data. A description of these categories is described as follows.

Missing Completely At Random (MCAR)

Missing Completely At Random (MCAR) refers to a mechanism for generating missing data where the missingness does not depend on any of the observed variables in the dataset nor on unobserved values that have not been measured [33]. In other words, missingness in the dataset occurs entirely at random meaning that there is nothing systematic going on that makes some observations more likely to be missing than the others [12, 14]. It can be expressed as follows: *(1.1)*.

$$P(Y/y \text{ is missing}) = P(Y/y \text{ is observed}), \qquad (1.1)$$

which implies that the probability that an observation Y_i is missing is unrelated to its value and the value of other variables in the dataset. If data fulfil the MCAR assumption, then analysis of the

dataset can yield consistent results comparable to analysis of the dataset without missing data. For example, in a study that records information from cases of campylobacteriosis, missing overseas travel status of the cases can be assumed to be MCAR if it occurs because some individuals simply forgot to fill in the travel information section of the questionnaire. Therefore, complete case analysis (i.e. excluding all observations with missing values of a variable/s in the data) of the MCAR data results in valid inference [36].

Missing At Random (MAR)

The fundamental and most widely used assumption about missing data is the MAR model [33]. The term MAR, however, is often confusing because of the use of the word "random". The MAR mechanism is not random per se but it describes a systematic missingness in the dataset where the missing data are correlated with other variables in the study [37]. These other variables provide the mechanism for explaining missing values in the dataset. The MAR mechanism assumes that the missing value of Y_i does not depend on the value of Y_i, after controlling for other variables (*X's*) in the dataset [38] *(1.2)*.

$$P(Y/y \text{ is missing}, X) = P(Y/y \text{ is observed}, X) \qquad (1.2)$$

In other words, the missingness in the system can be explained by the measured variables in the dataset, and not on either the values of the missing observation or the value of unobserved variables [33]. For example, in monitoring the effect of a certain diet on the level of *E. coli* 0157:H7 shedding in a dairy farm, if care takers having insufficient education have difficulties in recording the dietary intake of the cows then the data are MAR because it is the trouble in recording not the diet itself

that is causing the missing data, i.e., the absence of the data depends on the level of education of the care taker.

It is generally very difficult to know the type of missingness in a dataset [38]. However, one way to establish the plausibility of the MAR mechanism is to include as many variables as possible in the analysis that are strong predictors for the outcome variable *Y* [39]. Sometimes MCAR and MAR are called '*ignorable'* missing data mechanisms, which means that there is no need to separately model the missing data mechanism to draw valid inference [40].

Missing Not At Random (MNAR)

If the missing data are not MAR, or if the assumption of MAR is violated, the missing data mechanism is said to be MNAR or *non-ignorable* missing data. Unfortunately, there is no standard empirical test for the MAR assumption and therefore it is impossible to rule out MCAR missingness. Missing Not At Random occurs when the missingness in the system cannot be adequately explained by the measured covariates in the dataset or when it depends on the value of the missing data. In other words, even after conditioning on *X*, the distributions of observed and missing values are not similar as shown below: *(1.3).*

$$P(Y/y \text{ is missing}, X) \neq P(Y/y \text{ is observed}, X) \qquad (1.3)$$

Missing Not At Random frequently occurs whenever the missing values of variable *Y* are related to the values of *Y*. For example, while studying mortality rate associated with foodborne diseases and certain regions that have a high mortality rate did not report those values, then the dataset is regarded to be MNAR. In such cases, the missingness is not ignorable and must be modeled separately to obtain a valid inference from the dataset [38]. This modeling accounts for the missingness mechanism and it will be incorporated into a more complex model to estimate the

missing data [41]. For effective estimation of missing data under MNAR situation, however, a good prior knowledge of the source of missing values is required.

1.2.2 Multidimensional data

Datasets that contain several variables but consisting of relatively few observations per variables are often called multidimensional datasets (i.e., the number of measured covariates is much larger than the sample size) [42]. Massive amounts of multidimensional datasets are continuously being produced by biomedical research and stored at cheaper costs than the recent past. Although these datasets offer rich information, they result in opportunities and challenges regarding data analysis [43]. In a regression model, for example, when the number of predictor variables is too high as compared to the number of observations, the model lacks enough degrees of freedom and therefore the approach is at risk of over-fitting [44]. High dimensional data analysis have become increasingly important in various disciplines such as in science, engineering, and biomedical sciences [45, 46]. Typical examples include microarray data where the number of genes represent the variables and the cases sampled the observations. Here, the number of variables associated with the cases is larger than the number of cases. Extraction of a subset of features that are relevant for a disease outcome from large numbers of potential risk factors is another challenging area in multidimensional data analysis, the so-called "variable selection process" or "dimension reduction process" [47]. Other examples of large data include financial data and texture classification in image processing research [43].

It is frequently unnecessary to assume that all variables in the dataset contribute for the uncertainty associated with predicting the target response variable, such as morbidity or mortality due to a

certain disease. Some of the variables may be simply redundant measurements that can be explained by other variables in the dataset, while others may be completely irrelevant [42]. In real world data collection, however, researchers usually gather as much information as possible. The primary goal of analyzing a dataset with too many variables and too few observations is to retain a subset of relevant variables that can convey as much information as the full dataset while increasing the scope of interpretation [48]. On the other hand, certain dimension reduction approaches produce fewer new 'variables' by means of linear combination of the existing variables to understand the underlying phenomenon of interest. Such methods include Principal Components Analysis (PCA), Factor Analysis (FA) and Linear Discriminant Analysis (LDA). Detailed description of these dimension reduction methods can be found in literature [49, 50]. After selecting the subset of meaningful variables, the appropriate analytic approach can be applied for parameter estimation or prediction of outcomes.

1.2.3 Analysis approaches

In this section, commonly used statistical analysis approaches for missing and multidimensional data are described. Until recently, missing data analysis has heavily relied up on analysis of only complete observations after discarding those with missing values [51]. However, a wide variety of alternative methods for efficient utilization of missing data have been developed [5, 52–54]. The appropriateness of a given method for analyzing missing data is dependent on the mechanism by which the missingness was introduced into the dataset as described in section (1.2.1). It is also important to know whether one is dealing with missing outcome, missing predictor or missing in both, as the analytic approach may depend accordingly. In the statistical literature, there are two frameworks for statistical inference, namely, the classical (Frequentist) and the Bayesian

frameworks. The following section describes various approaches within each framework for dealing with missing and multidimensional data. For detailed technical and theoretical understanding, the reader is directed to the following references [15, 54, 55]

Classical (Frequentist) Inference

The Classical (Frequentist) approach to statistical analysis, which was developed by Ronald Fisher in the early 20th Century, is based on the notion that probabilities are fundamentally related to the frequency of events [55]. Most conventional frequentist approaches do not have the full capacity to handle missing data without losing information. While it would be impossible to adequately summarize all applications of frequentist approaches in this thesis, a selection of commonly used methods related to analysis of imperfect data are described.

Listwise deletion (Complete Case Analysis): Traditional methods for dealing with missing data, which are applied by many commonly used statistical packages, involve deleting any case that has missing values for any variable in the dataset (also called *listwise deletion*) [38]. This is the easiest approach and does not require any special computation method [30]. This method requires the MCAR assumption to hold for valid inference. Ignoring incomplete observations and restricting the analysis to complete cases often compromises the validity of inference derived from the complete cases, particularly when the proportion missing is high, and that the missingness is not MCAR [56]. There are a number of drawbacks to this approach. These include loss of information and lack of representativeness of the complete cases for making reliable inference to the population [30, 57].

Expectation maximization (EM): The most common model-based framework for missing data under the frequentist paradigm is the EM approach. The EM algorithm is a technique often used

to obtain the Maximum Likelihood (ML) estimates of the unknown parameters in a model using an iterative approach [58]. This method iterates through a process of estimating missing data and then parameters using the ML approach. It repeatedly cycles back and forth between two steps, the *Expectation* and the *Maximization* steps. In the *Expectation* step, the expected value of the log-likelihood is estimated for the variables that have missing data and during this step, expectations are computed using the current parameter values. In the *Maximization* step, the expected log-likelihood is maximized to obtain new parameters estimates. These cycles continues until the estimates do not change substantially while cycling through the two steps (i.e., until convergence is reached) [51, 59, 60]. The main disadvantages of the EM method include producing biased parameter estimates, the inability to compute standard errors and the slow rate of convergence particularly whenever the fraction missing is very high[51]. In addition, this method requires a large sample size and the assumption of MAR to hold [61].

Regularization: commonly used classical regression based variable selection methods, such as the backward and forward regression methods, attempt to reduce the number of variables and select the most parsimonious model that still guarantees good predictive performance [62]. However, these methods are not useful if there are larger numbers of variables (p) than observations (n) in a dataset (i.e., $p \gg n$). In such scenarios, these approaches will produce non-identifiable models, high collinearity of variables and can become computationally unstable [63]. The difficulty of determining the parameter values of data is commonly called 'identification problem' [64]. In this instance, *penalized regularization* methods have been proven to be meaningful both theoretically and empirically [65, 66]. Regularization is a process of fine tuning or selecting the preferred level of model complexity by introducing external information (penalty term) to avoid too complex and

over-fitting (or too simple and underfit) models and hence find the optimal predictive model [67]. For model regularization, two things are required: a tuning parameter (lambda), which controls the level of complexity (smoothness) of the models, and a way of checking the predictive performance of the models (cross validation). This technique constrains or regularizes regression coefficient estimates (i.e., shrinks some of the coefficient estimates towards zero, and hence retains a subset of variables) and improves the fit of the model [66].

Several regularization solutions have been proposed in the past that include Ridge regression [68], the LASSO (Least Absolute Shrinkage and Selection Operator) [66] and Elastic net [65]. The ridge regression achieves a better prediction performance by shrinking regression coefficients. It does not produce a parsimonious model as it always keeps all the variables in the model. However, the LASSO, which is an improvement over the Ridge, does both subset selection and shrinkage. Because variable selection is a more important problem in modern data analysis, the LASSO is the more relevant of the two. The Elastic net, on the other hand is an optimized approach that carries both the qualities of LASSO and Ridge, for variable selection and shrinkage, respectively [65]. The application of Elastic net for variable selection is described in section 4.1 of this thesis. An exhaustive survey of regularization methods is beyond the scope of this thesis and the reader can find a comprehensive discussion of the statistical theory behind each method in literature [65–68].

Multiple imputation (MI): Multiple Imputation is the most commonly used model-based approach for handling missing data in many health research datasets. It was first proposed by Rubin (1976) and subsequently developed by several other researchers [33, 69–71]. As it has become a standard

procedure for handling missing data, most modern statistical programs contain tools to perform MI. The method assumes that the missingness is *ignorable* or MAR. Multiple Imputation consists of three steps. The first step is to create multiple datasets (also called "multiply datasets") by imputing missing observations with plausible values using the predictive distribution of the covariates in the dataset [52]. Usually 5-10 multiply datasets are sufficient whenever the proportion missing is not excessive. However, it may be sensible to use 20 or more imputed datasets [72, 73]. Second, each completed dataset is analyzed separately using standard statistical analysis procedures as if each multiply dataset is complete dataset. Finally, individual estimates of each imputed dataset are combined into one overall estimate and variance using Rubin's rule [33]. Since the set of possible imputed values is drawn from the conditional distribution of the missing variables given the observed data, the values inherently contain variation. Although MI is fundamentally a frequentist approach, random sampling from the distribution of the data is derived from Bayesian theory (described below in the next section) therefore it is a blend of both frameworks. There is a variety of resources for detailed information on MI [69, 71, 74, 75].

Bayesian Inference

The Reverend Thomas Bayes (1702-1761) developed what is currently called "Bayes' Theorem" which was the first expression for inverse probability and this theorem is the basis of Bayesian inference. Bayesian inference is a process of learning from data. Bayesian methods are based on the assumption that probability is operationalized as a degree of belief, and not a frequency as is done in classical, or frequentist, statistics. It has recently become a standard tool for data analysis and been widely used in many research areas including experimental research [76], risk assessment

[77], social sciences [78], economics [79], physics [80], chemistry [81] and epidemiology [82]. More specifically, Bayesian statistical approaches have been widely used (among many others) for missing value prediction, disease mapping in time and space and for source attribution to make optimal usage of data [13, 14, 83, 84].

A Bayesian framework offers a formal way for combining two pieces of information using Bayes' rule. These are the observed data described by the likelihood and the prior information which contains the distribution of previous knowledge regarding the event of interest, linked by Bayes' rule as follows *(1.4)*

$$p(\theta/Y) = p(\theta)\, p(Y/\theta)/p(Y) \qquad (1.4)$$

where $p(\theta/Y)$, often called the *posterior distribution*, is the conditional probability of observing θ (which is the parameter of interest, e.g., the mean mortality rate due to foodborne diseases) given Y (the data), $p(\theta)$ is the parameter's *prior distribution*, $p(Y/\theta)$ is the likelihood (i.e. the conditional probability of the data given a particular set of value of θ), and $p(Y)$ is the marginal distribution of the data. In other words, the Bayesian analysis begins with some prior belief, $p(\theta)$, and after learning from the data, $p(Y/\theta)$, the prior belief about θ will change or will be updated to obtain a new information, $p(\theta/Y)$. All inference will be based on this new information.

Because of the fact that Bayesian inference is fully probabilistic, there is no distinction between observed data and unobserved entities in the model. This implies that observed values, missing values, and parameters are all treated in a unified and consistent manner which makes the Bayesian method superior for data analysis in the presence of uncertainty [85]. Application of Bayesian methodologies for a variety of health problems has been increasing in recent years [86, 87]. The main limiting factor to carry out Bayesian inference for routine decision making was

computational challenge. Therefore, the fairly recent increase in the use of Bayesian inference in numerous disciplines is due to the rapid development of computational hardware and software that have made Bayesian analysis feasible for a number of applications [88]. More specifically, the major advances in simulation methods (such as the Marcov Chain Monte Carlo (MCMC) process) and other Bayesian machinery that has been incorporated into easily available software programs such as *WinBUGS* (Windows for Bayesian Inference Using Gibbs Sampling), *JAGS* (Just Another Gibbs Sampling), *INLA* (Integrated Nested Laplace Approximation) and *Stan* have significantly simplified the use of Bayesian models [89–92].

1.2.4 Disease Mapping

One of the fundamental requirements for the control and prevention of disease is to understand the underlying geographical distribution of risk of the disease in a population. Disease maps frequently provide a rapid visual summary of spatial information and may identify patterns in the data that cannot be depicted by other representations. The main focus of disease mapping lays in describing disease in space, for generating potential hypothesis regarding causation, for risk surveillance and devising policy regarding strategic resource allocation [93]. There are two main ways to obtain risk estimates for mapping, namely the Standardized Mortality/Morbidity Ratio (SMR) - the classical approach, and Bayesian Disease Mapping (BDM). In the following section, description of disease mapping in the context of classical and Bayesian frameworks is presented.

Standardized Mortality /Morbidity rates (SMR):

Conventional methods for disease mapping utilize the SMR which is computed as the ratio of observed (O_i) and expected (E_i) number of events (e.g., deaths, diseases etc.) in a given area assuming that the events follow a *Poisson* distribution and that the mean is equal to the variance, as denoted below *(1.5)*[94, 95]

$$SMR\ (\theta_i) = O_i / E_i \qquad (1.5)$$

The expected count (E_i) is usually the number of events in an age-sex adjusted population for a given area. One of the fundamental shortcomings of the classical SMR is that the variance can be large in less populated areas (and hence large E_i) and it can be small in heavily populated areas resulting in unstable and extreme values for risk of the disease[96, 97]. Moreover, assuming that the relative risks (θ_i) are independently drawn from a common distribution is unrealistic in most epidemiological studies [97]. This is due to the fact that θ_i's are typically spatially (temporally) correlated as they reflect spatially (temporally) varying risk of disease and therefore the model requires incorporation of spatial (temporal) dependence [97]. Spatially smoothed estimates are, therefore, more appropriate for identifying true geographical variation of risk of disease which take spatial correlation into account [98, 99]. However, the SMR does not allow for spatial and temporal correlation to be taken into account while estimating the risk of disease [100, 101]. Such inconsistencies in risk estimates are addressed by using the Bayesian approach (discussed below) that can produce smoothed (shrunk) estimates.

Bayesian disease mapping (BDM):

Most disease mapping studies want to identify variation in risk of disease that could be due to unmeasured variables or random effects. The inherent hierarchical structure of the Bayesian

framework provides a convenient platform to incorporate a spatial or temporal correlation across the estimated local disease rates through spatial/temporal random effects [99]. Hence, the BDM method accounts for the spatial correlation between disease risk in neighboring regions which is ignored by the SMR approach [102, 103]. More specifically, the Bayesian structure partitions random effects into several components where each part will be assigned a distinct prior information based on the required inference [104]. For example, the random effect can allow for over-dispersion, also called *Uncorrelated Heterogeneity* (UH) (i.e., whenever the mean is not equal to the variance), and a component for spatial correlation, also called *Correlated Heterogeneity* (CH) [96]. These two components are fitted in the model to capture any extra variation in the data that cannot be explained otherwise. Just as in the case of other parameters in a Bayesian model, priors are required for these two random effect components [104]. The spatially structured component (CH) is usually assigned a so called *Conditional Autoregressive prior* (CAR) which assumes that the mean risk for each area, conditional on the neighboring areas, has a normal distribution with the mean equal to the average risk of the neighboring areas and variance inversely proportional to the number of neighbors [13, 99]. This prior specification assures that neighboring regions are more alike in risk than those that are located farther away. A prior for the amount of spatial similarity or unstructured heterogeneity effect of disease rates, however, is usually difficult to specify [105]. Therefore, a non-spatially structured (also called exchangeable) normal prior distribution is commonly specified for the UH component [106]. It is argued that including both the random effects is more flexible than only the CAR random effect since the former formulation allows the data to decide on the contribution of spatially structured variation and unstructured over-dispersion regarding residual disease risk [97, 99].

The Bayesian framework essentially "borrows" more information from neighboring areas than from areas located further away, smoothing (shrinking) local risk towards local and neighboring values [99, 102]. This gives a more stable estimate of the underlying risk of disease than that produced by the SMR. By doing so, the Bayesian hierarchical structure allows for a model-based estimation of missing data whenever certain regions have less or no data while the neighboring regions contain relatively better information.

The application of BDM is further described in Chapters 3 and 5 of this thesis. In Chapter 3, its application is demonstrated for identifying the temporal and spatial distribution of the risk of travel associated campylobacteriosis in District Health Boards of New Zealand. In Chapter 5, application of BDM is implemented to develop a risk map of malaria prevalence in the Busia region of Western Kenya using the People's Animals and their Zoonoses (PAZ) project dataset.

1.3 References

1. Heitjan DF: Incomplete data: what you don't know might hurt you. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol 2011, 20:1567–1570.

2. Handbook of Statistics: Epidemiology and Medical Statistics. Elsevier; 2007.

3. Lilford RJ, Thornton JG, Braunholtz D: Clinical trials and rare diseases: a way out of a conundrum. BMJ 1995, 311:1621–1625.

4. Gurrin LC, Kurinczuk JJ, Burton PR: Bayesian statistics in medical research: an intuitive alternative to conventional data analysis. J Eval Clin Pract 2000, 6:193–204.

5. He Y: Missing data analysis using Multiple Imputation getting to the heart of the matter. Circ Cardiovasc Qual Outcomes 2010, 3:98–105.

6. Zhen H: Multiple Imputation of missing data based on prediction of conditional quantiles. ProQuest; 2008.

7. Harrington D: Confirmatory Factor Analysis. Oxford University Press; 2008.

8. Ratner B: Variable selection methods in regression: Ignorable problem, outing notable solution. J Target Meas Anal Mark 2010, 18:65–75.

9. Raftery AE, Madigan D, Hoeting JA: Bayesian model averaging for linear models. J Am Stat Assoc 1997, 92:179–191.

10. Friedman J, Hastie T, Tibshirani R: The Elements of Statistical Learning. Volume 1. Springer series in statistics Springer, Berlin; 2001.

11. An H, Huang D, Yao Q, Zhang C-H: Stepwise searching for feature variables in high-dimensional linear regression. 2008.

12. Fan J, Lv J: A selective overview of variable selection in high dimensional feature space. Stat Sin 2010, 20:101–148.

13. Lawson AB: Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology. 1st edition. Chapman and Hall/CRC; 2008.

14. Oba S, Sato M, Takemasa I, Monden M, Matsubara K, Ishii S: A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 2003, 19:2088–2096.

15. Gelman A, Hill J: Data Analysis Using Regression and Multilevel/Hierarchical Models. 1st edition. Cambridge University Press; 2006.

16. Ranta J, Matjushin D, Virtanen T, Kuusi M, Viljugrein H, Hofshagen M, Hakkinen M: Bayesian Temporal Source Attribution of Foodborne Zoonoses: Campylobacter in Finland and Norway. Risk Anal 2011, 31:1156–1171.

17. Winkler RL: Why a Bayesian analysis has not caught on the healthcare decision making. Int J Technol Assess Health Care 2001, 17:56–66.

18. Veregin H: Data quality parameters. Geogr Inf Syst 1999, 1:177–189.

19. Graham JW: Missing data analysis: making it work in the real world. Annu Rev Psychol 2009, 60:549–576.

20. Dong Y, Peng C-YJ: Principled missing data methods for researchers. SpringerPlus 2013, 2:1–17.

21. Rubin LH, Witkiewitz K, Andre JS, Reilly S: Methods for handling missing data in the behavioral neurosciences: don't throw the baby rat out with the bath water. J Undergrad Neurosci Educ 2007, 5:A71–A77.

22. Raghunathan TE: What do we do with missing data? Some options for analysis of incomplete data. Annu Rev Public Health 2004, 25:99–117.

23. Rubright JD, Nandakumar R, Glutting JJ: A simulation study of missing data with multiple missing X's. Pract Assess Res Eval , 19:2.

24. Holt B, Benfer RA: Estimating missing data: an iterative regression approach. J Hum Evol 2000, 39:289–296.

25. Graham JW: Missing Data. New York, NY: Springer New York; 2012.

26. Enders CK: Applied Missing Data Analysis. Guilford Press; 2010.

27. Marlin BM: Missing Data Problems in Machine Learning. 2008.

28. McKnight PE, McKnight KM, Sidani S, Figueredo AJ: Missing Data: A Gentle Introduction. Guilford Press; 2007.

29. Curran D, Bacchi M, Schmitz SFH, Molenberghs G, Sylvester RJ: Identifying the types of missingness in quality of life data from clinical trials. Stat Med 1998, 17:739–756.

30. Buhi ER, Goodson P, Neilands TB: Out of Sight, Not Out of Mind: Strategies for Handling Missing Data. Am J Health Behav 2008, 32:83–92.

31. Tabachnick BG, Fidell LS: Using Multivariate Statistics. 6 edition. Pearson; 2012.

32. Seaman S, Galati J, Jackson D, Carlin J: What Is Meant by "Missing at Random"? Stat Sci 2013, 28:257–268.

33. Rubin DB: Inference and missing data. Biometrika 1976, 63:581–592.

34. Little RJA, Rubin DB: Statistical Analysis with Missing Data. 2nd ed. Hoboken, N.J: Wiley; 2002. [Wiley Series in Probability and Statistics]

35. Widaman KF: Iii. Missing Data: What to Do with or Without Them. Monogr Soc Res Child Dev 2006, 71:42–64.

36. Fox J: Applied Regression Analysis and Generalized Linear Models. SAGE Publications; 2015.

37. Baraldi AN, Enders CK: An introduction to modern missing data analyses. J Sch Psychol 2010, 48:5–37.

38. Allison PD: Missing Data. SAGE Publications; 2001.

39. Collins LM, Schafer JL, Kam CM: A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods 2001, 6:330–351.

40. Holman R, Glas CAW: Modelling non-ignorable missing-data mechanisms with item response theory models. Br J Math Stat Psychol 2005, 58:1–17.

41. Schafer JL, Graham JW: Missing data: Our view of the state of the art. Psychol Methods 2002, 7:147–177.

42. Schmid M, Potapov S, Pfahlberg A, Hothorn T: Estimation and regularization techniques for regression models with multidimensional prediction functions. Stat Comput 2010, 20:139–150.

43. Fan J, Han F, Liu H: Challenges of Big Data Analysis. Natl Sci Rev 2014, 1:293–314.

44. Hawkins DM: The Problem of Overfitting. J Chem Inf Model 2004, 44:1–12.

45. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, Fine MJ, Glymour C, Gordon G, Hanusa BH, Janosky JE, Meek C, Mitchell T, Richardson T, Spirtes P: An evaluation of machine-learning methods for predicting pneumonia mortality. Artif Intell Med 1997, 9:107–138.

46. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinforma Oxf Engl 2000, 16:906–914.

47. Zhang Q, Abel H, Wells A, Lenzini P, Gomez F, Province MA, Templeton AA, Weinstock GM, Salzman NH, Borecki IB: Selection of models for the analysis of risk-factor trees: leveraging biological knowledge to mine large sets of risk factors with application to microbiome data. Bioinforma Oxf Engl 2015.

48. Brown PJ, Vannucci M, Fearn T: Bayes model averaging with selection of regressors. J R Stat Soc Ser B Stat Methodol 2002, 64:519–536.

49. Jolliffe IT: Principal Component Analysis. Springer Science & Business Media; 2002.

50. Song F, Mei D, Li H: Feature selection based on linear discriminant analysis. In 2010 International Conference on Intelligent System Design and Engineering Application (ISDEA). Volume 1; 2010:746–749.

51. Roth PL: Missing Data: A Conceptual Review for Applied Psychologists. Pers Psychol 1994, 47:537–560.

52. van Buuren S: Flexible Imputation of Missing Data. 1st edition. Boca Raton, FL: Chapman and Hall/CRC; 2012.

53. Chen F: Missing no more: Using the MCMC procedure to model missing data. In Proceedings of the SAS Global Forum 2013 Conference, Cary NC: SAS Institute Inc. Available at http://support.sas.com/resources/papers/proceedings11/436-2013. pdf. Citeseer; 2013.

54. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB: Bayesian Data Analysis, Third Edition. CRC Press; 2013.

55. Fisher RA: On the mathematical foundations of theoretical statistics. Philos Trans R Soc Lond Ser Contain Pap Math Phys Character 1922:309–368.

56. Acock AC: Working With Missing Values. J Marriage Fam 2005, 67:1012–1028.

57. O'Rourke T: Methodological techniques for dealing with missing data. Am J Health Stud 2003, 18:165–168.

58. Dempster AP, Laird NM, Rubin DB: Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B Methodol 1977, 39:1–38.

59. Horton NJ, Kleinman KP: Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am Stat 2007, 61:79–90.

60. Pigott TD: A review of methods for missing data. Educ Res Eval 2001, 7:353–383.

61. Allison PD: Handling Missing Data by Maximum Likelihood. 2012.

62. Blanchet FG, Legendre P, Borcard D: Forward selection of explanatory variables. Ecology 2008, 89:2623–2632.

63. Núñez E, Steyerberg EW, Núñez J: Regression Modeling Strategies. Rev Esp Cardiol Engl Ed 2011, 64:501–507.

64. Jacquez JA, Greif P: Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. Math Biosci 1985, 77:201–227.

65. Zou H, Hastie T: Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 2005, 67:301–320.

66. Tibshirani R: Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B 1994, 58:267–288.

67. Breiman L: Heuristics of instability and stabilization in model selection. Ann Stat 1996, 24:2350–2383.

68. Hoerl AE, Kennard RW: Ridge Regression: biased estimation for nonorthogonal problems. Technometrics 1970, 12:55.

69. Schafer JL: Multiple imputation: a primer. Stat Methods Med Res 1999, 8:3–15.

70. Rubin DB: Frontmatter. In Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc.; 1987:i–xxix.

71. Rubin DB: Multiple Imputation After 18+ Years. J Am Stat Assoc 1996, 91:473–489.

72. Graham JW, Olchowski AE, Gilreath TD: How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci Off J Soc Prev Res 2007, 8:206–213.

73. Kenward MG, Carpenter J: Multiple Imputation: current perspectives. Stat Methods Med Res 2007, 16:199–218.

74. Craig Enders, Samantha Dietz, Marjorie Montague, Jennifer Dixon: Modern alternatives for dealing with missing data in special education research. In Applications of Research Methodology. Volume 19. Emerald Group Publishing Limited; 2006:101–129. [Advances in Learning and Behavioral Disabilities, vol. 19]

75. Schafer JL, Olsen MK: Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivar Behav Res 1998, 33:545–571.

76. Ibrahim JG, Chen M-H, Chu H: Bayesian methods in clinical trials: a Bayesian analysis of ECOG trials E1684 and E1690. BMC Med Res Methodol 2012, 12:183.

77. O'Hagan A, Stevens JW: Bayesian methods for design and analysis of cost-effectiveness trials in the evaluation of health care technologies. Stat Methods Med Res 2002, 11:469–490.

78. Jackman S: Bayesian Analysis for the Social Sciences. 1st edition. Chichester, U.K: Wiley; 2009.

79. Poirier DJ, others: The growth of Bayesian methods in statistics and economics since 1970. Bayesian Anal 2006, 1:969–979.

80. Toussaint U von: Bayesian inference in physics. Rev Mod Phys 2011, 83:943–999.

81. Armstrong N, Hibbert DB: An introduction to Bayesian methods for analyzing chemistry data: Part 1: An introduction to Bayesian theory and methods. Chemom Intell Lab Syst 2009, 97:194–210.

82. Greenland S: Bayesian perspectives for epidemiological research: I. Foundations and basic methods. Int J Epidemiol 2006, 35:765–775.

83. Schrödle B, Held L: Spatio-temporal disease mapping using INLA. Environmetrics 2011, 22:725–734.

84. Ranta J, Matjushin D, Virtanen T, Kuusi M, Viljugrein H, Hofshagen M, Hakkinen M: Bayesian temporal source attribution of foodborne zoonoses: *Campylobacter* in Finland and Norway. Risk Anal Off Publ Soc Risk Anal 2011, 31:1156–1171.

85. Bousquet O, Luxburg U von, Rätsch G (Eds): Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003 [and] Tübingen, Germany, August 4-16, 2003: Revised Lectures. Berlin ; New York: Springer; 2004. [Lecture Notes in Computer Science, Lecture Notes in Artificial Intelligence, vol. 3176]

86. Etzioni RD, Kadane JB: Bayesian statistical methods in public health and medicine. Annu Rev Public Health 1995, 16:23–41.

87. Gandhi M, Mukherjee B, Biswas D: A Bayesian approach for inference from a bridging study with binary outcomes. J Biopharm Stat 2012, 22:935–951.

88. Malakoff D: Bayes offers a “new”way to make sense of numbers. Science 1999, 286:1460–1464.

89. Robert C, Casella G: A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. Stat Sci 2011, 26:102–115.

90. Lunn DJ, Thomas A, Best N, Spiegelhalter D: WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 2000, 10:325–337.

91. Carpenter B, Gelman A, Hoffman M: Stan: a probabilistic programming language. .

92. Plummer M, others: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In Proceedings of the 3rd international workshop on distributed statistical computing. Volume 124. Technische Universit at Wien; 2003:125.

93. Elliott P, Wartenberg D: Spatial Epidemiology: Current Approaches and Future Challenges. Environ Health Perspect 2004, 112:998–1006.

94. Liddell FD: Simple exact analysis of the standardized mortality ratio. J Epidemiol Community Health 1984, 38:85–88.

95. Jones ME, Swerdlow AJ: Bias in the Standardized Mortality Ratio when using general population rates to estimate expected number of deaths. Am J Epidemiol 1998, 148:1012–1017.

96. Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, Schlattmann P, Divino F: Disease mapping models: an empirical evaluation. Disease Mapping Collaborative Group. Stat Med 2000, 19:2217–41.

97. Gilks WR, Richardson S, Spiegelhalter D: Markov Chain Monte Carlo in Practice. CRC Press; 1995.

98. Waller LA, Carlin BP, Xia H, Gelfand AE: Hierarchical spatio-temporal mapping of disease rates. J Am Stat Assoc 1997, 92:607–617.

99. Besag J, York J, Mollié A: Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 1991, 43:1–20.

100. N. G. Best RAA: Bayesian models for spatially correlated disease and exposure data. 1999.

101. Veugelers PJ, Hornibrook S: Small area comparisons of health: applications for policy makers and challenges for researchers. Chronic Dis Can 2002, 23:100–110.

102. Clayton D, Kaldor J: Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987, 43:671–681.

103. Assunção R, Krainski E: Neighborhood dependence in Bayesian spatial models. Biom J 2009, 51:851–869.

104. Lawson AB, Browne WJ, Rodeiro CLV: Disease Mapping with WinBUGS and MLwiN. John Wiley & Sons; 2003.

105. Bernardo JM: Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting. Clarendon Press; 1999.

106. Waller LA, Gotway CA: Applied Spatial Statistics for Public Health Data. John Wiley & Sons; 2004.

Chapter 2: Variable selection and regression analysis for the prediction of mortality rates associated with foodborne diseases

Ermias Amene[1], Laura A Hanson[2], Elizabeth A Zahn[2], Sommer R Wild[2],*

Dörte Döpfer[1]

[1] University of Wisconsin-Madison, WI, USA

[2] Saint Olaf College, Northfield, MN, USA

*Correspondence:
Ermias Amene
University of Wisconsin-Madison
2015 Linden Drive
Madison 53706 WI
Email: amene@wisc.edu

Summary

The purpose of the current study was to apply a model-based approach for filling data gaps in mortality rates associated with foodborne disease using the WHO Vital Registration mortality dataset. *Correlation Analysis* and *Elastic net regularization* methods were applied to drop redundant variables and to select the most meaningful subset of predictors. Whenever predictor data were missing, multiple imputation was used to fill in plausible values. Secondly, *Cluster Analysis* was applied to identify similar group of countries based on the values of the predictors. Finally, a *Bayesian hierarchical regression model* was fitted to this dataset for predicting mortality rates. From 113 potential predictors, 32 variables were retained after Correlation Analysis. Out of these 32 variables, 8 predictors with non-zero coefficient were selected using *Elastic Net regularization*. Based on values of these variables, four clusters of countries were identified. The uncertainty of predictions was large for countries within the cluster lacking mortality rates and it is low for a cluster that has some information. A data-driven cluster of countries and a meaningful subset of predictors can be used to fill data gaps in mortality rate using predictions from Bayesian hierarchical regression models. The inherent uncertainty around the resulting mortality rates predictions is a consequence of the data quality.

Key words: Food borne diseases, Bayesian hierarchical regression, Cluster Analysis, Mortality, Prediction, Elastic Net

Disclaimer: The views expressed in this document are solely those of the authors and do not represent the views of the World Health Organization.

2.1 Introduction

Foodborne diseases (FBD) remain a growing concern for high levels of morbidity and mortality in the human population worldwide [1]. There are many indictors hinting at an increase in global incidence of FBDs [2, 3]. For industrialized countries in general, it has been estimated that one third of the population suffers from foodborne illness every year [4]. A recent study estimated 37.2 million illnesses, 228,744 hospitalizations, and 2,612 deaths each year due to FBD in the United States alone [5]. These figures are assumed to be extremely high in resource limited regions of the world although solid data are lacking in these regions [6].

Currently no precise information exists about the global burden of FBDs, although their effect on both development and trade is considered enormous [7]. This is substantially attributed to lack of suitable data about mortality and morbidity rates associated with FBD in many countries and regions of the world [8]. Challenges associated with imperfect data have been emphasized in studies about estimates for the global burden of pathogen-specific FBDs such as non-typhoidal *Salmonella* gastroenteritis and typhoid fever [9, 10]. This lack of data, particularly from developing countries, makes it difficult to calculate global estimates of disease burden. This has been a challenge for appropriate allocation and prioritization of resources for food safety control and intervention efforts [11]. In 2006, as part of the FBD prevention and control program, the World Health Organization's (WHO) Initiative to Estimate the Global Burden of Foodborne Diseases (FERG) was launched to fill this gap[1, 12].

This report represents a first attempt to cluster WHO countries based on average values of selected FBD predictors and use the variables and the clusters in a Bayesian hierarchical modeling framework to predict FBD mortality rates for countries missing the data.

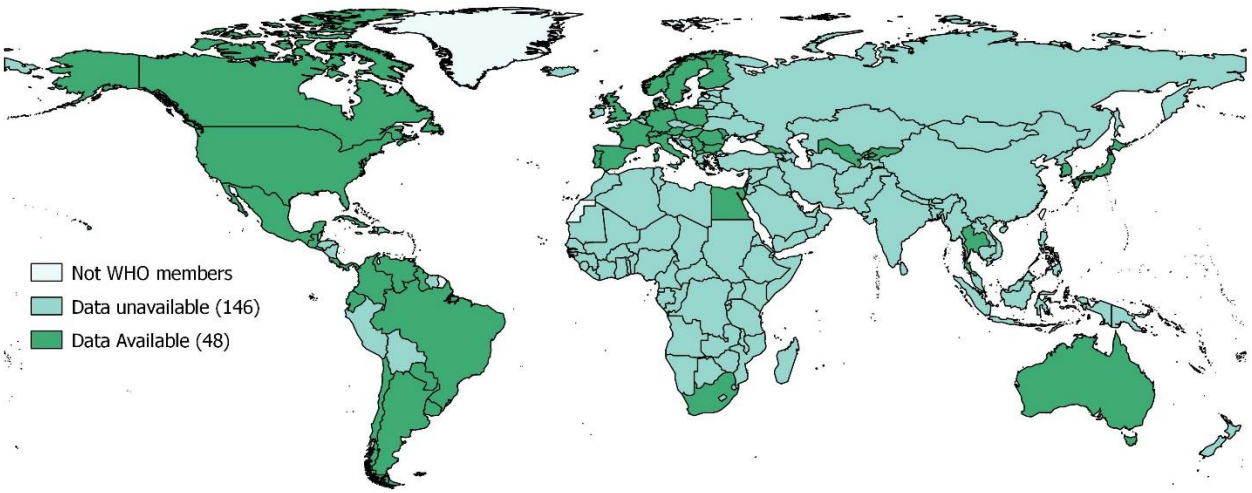
2.2 Methods

Dataset

Mortality data

We obtained the data regarding mortality rates associated with FBD from the WHO Vital Registration database (2000-2005) at the launch of FERG [12]. Foodborne diseases included in the database are bacterial and viral gastroenteritis, parasitic diseases and hepatitis A and E. International Classification of Disease coding system (ICD-10) was used to represent those diseases; but FBDs that are associated with chemicals and biotoxins were ruled out due to lack of specific ICD codes [13, 14]. The mortality rates were averaged over the available years and the mean rate was expressed per 100,000 population based on the 2005 population census. Since the mortality rate data were positively skewed, we log-transformed them to stabilize the variance and make the data normally distributed. Out of 194 WHO countries, only 48 have complete data on national mortality rates associated with FBD and the remaining 146 lack this information (see Figure 2.1)

Figure 2.1. The geographical distribution of WHO countries based on the availability of data for mortality rate due to foodborne diseases.



Note: numbers in parenthesis are the number of countries

Predictor set

We obtained potential predictors of mortality associated with FBD from publicly available databases (Food and Agricultural Organization (FAO), World Bank) in two steps. First, we obtained 113 predictors for countries with complete mortality rate data. After we selected the important subset of predictors, we searched values of the selected predictors for all WHO countries. The search criteria for the predictors included an established direct and/or indirect association with FBD mortality, a potential to be modifiable and having a global impact upon FBD mortality [2, 14, 15]. Not all 194 countries had a complete set of predictors therefore missing values ranged from 0 to 95%.

Statistical analysis

Analysis of the data was performed using the freeware statistical tools R version 3.1.2 and JAGS (Just Another Gibbs Sampler) version 3.4.0 [16, 17]. The models were specified and parameterized in R and the analysis was done by calling JAGS from *R* using *R2jags* package [17]. A stepwise approach was followed for variable selection and estimation of missing mortality rates associated with FBD as follows:

Correlation analysis

First, among a total of 113 variables, 67 contained at least one missing observation and therefore were excluded from further analysis. The remaining 46 variables were subjected to pairwise correlation analysis (CorA) to avoid highly correlated and redundant variables. Those pairs of variables with a correlation coefficient, r>=0.85 [18], were identified and one member of a pair with high correlation was retained based on biological plausibility.

Elastic Net regularization (ENR)

Following CorA, we applied *Elastic Net regularization* path (ENR). It offers a statistically appealing regression approach to select meaningful subset of predictors of mortality associated with FBD for the 48 countries with complete mortality data. *ENR* is a relatively new variable selection method proposed by Zou and Hastie (2005), which was developed to overcome the flaws of the commonly used Ordinary Least Squares approach with regard to prediction accuracy [19]. The basic form of linear regression model used to perform variable selection with an *ENR* is *(4.1)*:

$$\mathbf{Y} = \mathbf{X}\beta + e, \; e \sim N(0, \sigma^2) \qquad (4.1)$$

Where **Y** is a vector of log-total mortality rates (response variable), **X** is an *n x p* matrix of predictors, β is *p* vector of regression coefficients and *e* is the vector of residual errors. The *ENR* uses a mixture of the $\mathbf{l}_1$ (Least Absolute Shrinkage and Selection Operator (LASSO)) and $\mathbf{l}_2$ (ridge regression) penalties, which does both automatic variable selection and shrinkage respectively [19]. *ENR* has two parameters, α and λ. We set α at 0.5 (1=LASSO and 0=Ridge) and performed cross-validation to find the optimal value of regularization parameter λ. The λ value was used for variable selection. The *glmnet* package in R was used to fit the *ENR* procedure [20]. A detailed description of regression–based *ENR* as a data mining technique can be found in literature [19–21].

Imputation of missing values

After we selected variables using *ENR*, we searched values of these variables for the remaining 146 countries from publicly available and validated databases (FAO, World Bank). Whenever multiple values for a given country were available, we took the value for the latest year. Since not all countries had full information of the selected predictors, Multiple Imputation (MI) was performed to fill-in missing predictor values using the *MICE* (Multiple Imputation using Chained Equations) package in R [22]. Multiple Imputation helps to handle missing data, where missing values are replaced by random draws from the predictive distribution of the missing data given the observed data [22, 23]. The procedure generates *m* numbers of complete datasets (also called "multiply" datasets) ready for further analysis. Optimum number of *m* varies across studies and may depend on the study design and the proportion of values missing. Literature reports that 5-10 multiply datasets are sufficient for generating valid estimates [24, 25]. We used twenty multiply datasets in this study. The imputed values were averaged across the number of multiply datasets

to fill values for a given missing value. Convergence of the imputation process was assessed by visually examining density plots of each variable to evaluate the plausibility of imputed values across the number of iterations during the imputation process.

Cluster analysis

Following the imputation step, we carried out Cluster Analysis (CA). The purpose of CA is to aggregate countries into groups based on similar values of predictors such that countries within a cluster have homogenous mortality rates. Although several types of clustering methods exist, we compared four commonly used hierarchical clustering methods to identify the appropriate clustering solution for our dataset. These are *single linkage*, *complete linkage*, *UPGMA* (Unweighted Pair Group Mean Average) and *Ward's minimum variance* methods [26]. We used visual examination of the resulting dendrograms, Gower's distance [27] and Cophenetic correlation to select the method of choice [28]. Based on established rules, smaller Gower's dissimilarity coefficient and larger Cophenetic correlation indicate that the preferred clustering solution fits the data well. We selected UPGMA as the clustering method of choice for our data. We decided on the optimal number of clusters using the *gap-statistic* which is one of the most popular methods for estimating the number of clusters in a dataset [29]. In addition we evaluated the average Silhouette Width (SW) which is a composite index reflecting the compactness and separation of clusters (a high SW indicates the clusters are homogenous). A detailed technical description of these methods can be found elsewhere [27, 28, 30].

Bayesian hierarchical Regression

After having developed a dataset for all 194 countries, we fitted a Bayesian Hierarchical Regression Model (BHM) for predicting log-total mortality rate associated with FBDs *(2.2)*. We incorporated the clusters obtained from the CA as random effects into our BHM. The regression model fitted to the data is as follows:

$$\boldsymbol{Y}_i = N(\boldsymbol{\alpha}_j[i] + \boldsymbol{\beta}_k X_{ki}, \boldsymbol{\sigma}_y^2), \text{ for } i=1,\ldots n \text{ ; } k=1,2..K \qquad (2.2)$$
$$\boldsymbol{\alpha}_j = N(\mu_\alpha, \sigma_\alpha^2), \text{ for } j=1,..J$$

where $\boldsymbol{Y_i}$ denotes the response variable (log-total mortality); $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the intercept and the regression coefficients, respectively; n= total number of countries; $\boldsymbol{X_{ki}}$ denotes the predictors (K=8); J= number of clusters; the variance ($\boldsymbol{\sigma^2}$), $\boldsymbol{\alpha}$*'s* and $\boldsymbol{\beta}$*'s* are parameters to be estimated from the data. (The *R-JAGS* code is provided in Appendix 2.1). In addition to the model constructed using our four cluster solution, we evaluated a new model incorporating the WHO's Global Environmental Monitoring System/Food (GEMS) cluster for comparing the results. The GEMS/Food cluster categorizes the WHO countries into 17 groups based on food consumption and dietary intake of various chemicals [31]. A non-hierarchical Bayesian framework was also fitted to the data, which doesn't take any clustering of the data into account.

Valid inference from the above model assumes that the missingness in the system is *Missing At Random (MAR).* Missing data is considered *MAR* whenever the missingness can be explained by one or more predictors in the dataset. Although it is not possible to directly test the *MAR* assumption based on a data alone, it can be demonstrated by showing association between

predictors and missingness of the response variable [32]. Therefore, we created a dummy variable for whether mortality rate is missing or not, and run a logistic regression to statistically test if any of the variables are associated with missingness. A strong statistical association indicates that the MAR assumption can hold. A detailed description of missing data mechanisms can be found in Chapter 1 (section 1.2.1) [33, 34].

Some of the predictors in our dataset were not normally distributed and therefore we log-transformed them to stabilize their distribution before applying the regression approach. These subset of variables are "Per Capita animal calorie consumption", "Birth per Adolescent", "Fertility rate", "Maternal death risk" and "Kilo calorie per day".

In a Bayesian framework all the parameters in the model must have a prior distribution, which is a way of quantifying lack of knowledge about the parameters [34]. We assigned all the coefficients to have uninformative prior distribution (i.e a normal distribution with mean 0 and a precision of 0.01). This implies that the magnitudes of the regression coefficients are expected to lie between -10 and 10. The prior for the precisions, i.e. the inverse variances, were defined in terms of the standard deviation parameters and given uniform prior distribution on the range (0, 10) (Appendix 2.1).

We ran the model for 50,000 iterations with a burn-in of 5000 (i.e., discard the first 5000 iterations). We assessed convergence by running two chains of dispersed initial values and then by observing autocorrelation and density plots of the parameters from the models' outputs. Whenever more than one model is to be evaluated for fit, we used the Deviance Information

Criteria (DIC) and the effective number of parameters (pD) as model fit comparison tools [35]. The DIC is a Bayesian alternative of the Akaike Information Criteria for comparing competing models, and the *pD* is a measure of the complexity of the model [35]. A difference in the DIC of more than 5-10 units is regarded as strong evidence in favor of the model with smaller DIC [36].

Model Validation

In order to assess the predictive performance of our model, we carried out cross validation by using part of the dataset with complete information on mortality rates. We implemented the leave-one-out cross-validation method (LOOCV) used to estimate the generalizability of a model in the absence of external data [37]. This method takes one observation out of the data, sets it aside as a 'testing set', and fits the model using the remaining data, called the 'training set' to assess statistical predictions of the model. The resulting coefficients are then applied to the 'test set', to generate predicted values which are compared to the observed value of that single case. This procedure is performed repeatedly for all observations of the data and the Mean Absolute Error of prediction (MAE) is calculated *(2.3)* and compared with the baseline MAE (i.e the MAE computed without cross validation). This comparison allows to assess 'out-of-sample' predictive performance of the model whenever no external data exists [38].

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \qquad (2.3)$$

where, $\boldsymbol{n}$=total number of test sets, $\boldsymbol{f_i}$ =predicted log-total mortality and $\boldsymbol{y_i}$ =observed log-total mortality. Ninety five percent Confidence Intervals for the MAEs were computed by a non-

parametric bootstrapping method with 2000 replications using the *boot.ci* procedure in R. Smaller MAE indicates a better out of sample prediction of the model.

Sensitivity analysis

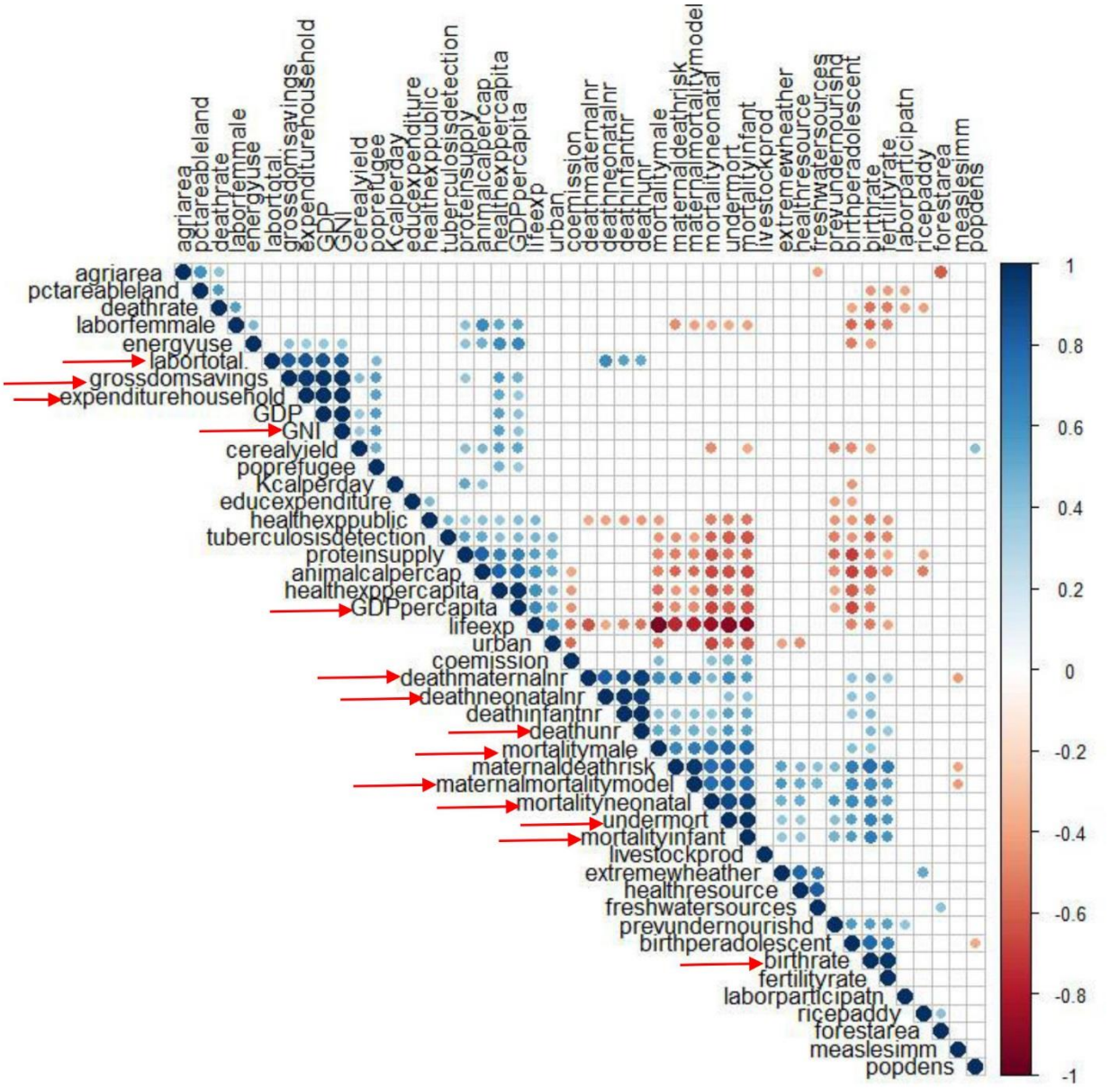
We assessed robustness of the results to model specification by changing the priors. Sensitivity analysis regarding priors was conducted by first assigning uninformative priors to means and inverse variances, and then changing the priors of the precision by a factor of 10 as shown in Table 4. Additionally we investigated the stability of predictions by randomly deleting mortality rates and fitting the model, and also by randomly adding plausible hypothetical mortality rates for a subset of countries missing the data and refitting the model.

2.3 Results

Correlation Analysis

Out of 46 variables screened by means of pairwise CorA, we retained 32 variables for further analysis as shown in the correlation matrix in Figure 2.2.

Figure 2.2. The correlation matrix of 46 potential predictors of mortality associated with Foodborne Diseases.



Notes: The values on the right of the plot represent correlation coefficient between a pair of variables. Lighter shades show strong positive correlation while darker shades indicate strong negative correlation between pair of variables. Variables that are indicted by the red arrows are those variables dropped after correlation analysis (n=14).

Elastic Net Regularization

The remaining 32 variables which were retained after CorA were subjected to *ENR*. Eight non-zero coefficient variables were selected as the final subset of predictors. We used these variables for CA and regression analysis as indicated in the next sections. Description of the eight variables and the proportion of missing values for these variables for all 194 countries are indicated in Table 2.1.

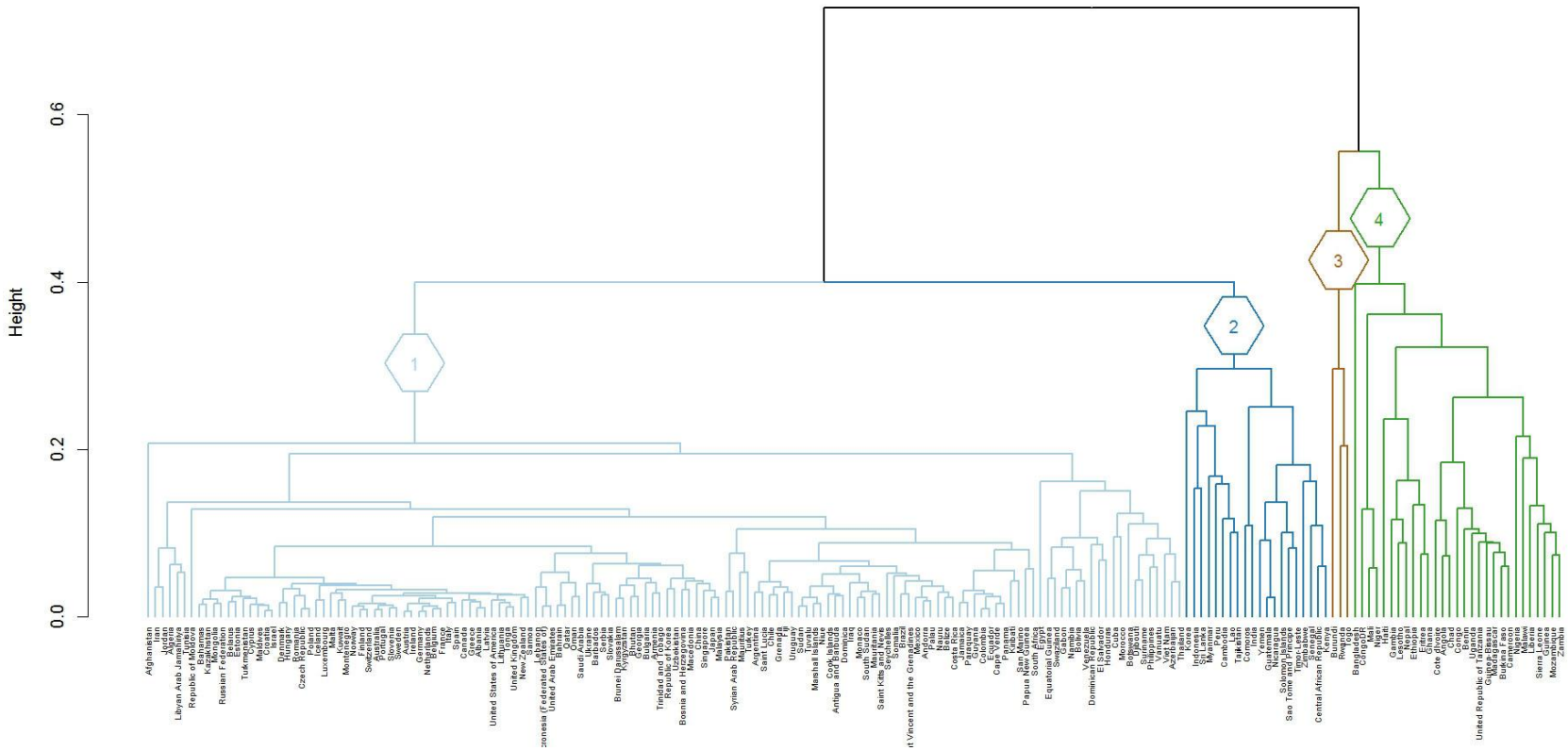
Table 2.1. Description and percent missing of the eight selected variables for predicting log-total mortality associated with Foodborne Diseases.

Variable name	Percent missing	Description (Source)
Life expectancy	11.8	Life expectancy at birth, total (Years). Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life http://data.worldbank.org/indicator/SP.DYN.LE00.IN
Animalcalpercap	11.8	Average Calorie Supply from Animal Products - per Capita http://faostat3.fao.org/search/*/E
Birthperadolescent	7.7	Adolescent fertility rate, the number of births per 1,000 women ages 15-19 http://data.worldbank.org/indicator/SP.ADO.TFRT
Pctareableland	3.0	Percent arable land (http://data.worldbank.org/indicator/AG.LND.ARBL.ZS)
Fertilityrate	5.1	Total fertility rate represents the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with current age-specific fertility rates. (http://data.worldbank.org/indicator/SP.DYN.TFRT.IN)
Maternaldeathrisk	7.7	Maternal mortality ratio (national estimate, per 100,000 live births) (http://data.worldbank.org/indicator/SH.STA.MMRT.NE)
Laborfemmale	9.2	Labor force participation rate for females ages 15-24: the proportion of the population ages 15-24 that is economically active. (http://data.worldbank.org/indicator/SL.TLF.ACTI.1524.FE.NE.ZS)
Kcalperday	12.0	Calorie supply per capita per day http://faostat.fao.org/

Cluster Analysis

The average Silhouette width (SW=0.59) and the Mantel optimal cluster methods resulted in two and three clusters respectively, while the *gap statistic* suggested four clusters to be optimum. As the gap statistic is the most recommended approach, we decided to partition our dataset into four clusters as shown in the dendrogram (Figure 2.3) [29].

Figure 2.3. Hierarchical Cluster Analysis of all 194 WHO countries using the Unweighted Pair Group Mean Average (UPGMA) method.



Notes: (1) Eight predictors of mortality associated with foodborne diseases were used to construct the dendrogram. The numbers shown on the top of the dendrogram indicate the cluster identification.

Regression and Model validation

Model validation and fit

The result of a statistical test for checking validity of the MAR assumption is shown in Table 2.2. Three of the eight predictors are significantly associated with missingness in the data indicating that the MAR assumption holds for the current analysis.

Table 2.2. Logistic regression of missing indicator (1=missing log-total mortality, 0=observed logtotal mortality) on predictors of mortality associated with Foodborne diseases to test the plausibility of Missing At Random Assumption.

Coefficients	**Estimate**	**Pr(>|z|)**
(Intercept)	21.14	0.002
Life expectancy in Years	0.18	0.001**
Per Capita calorie Supply (animal origin)*	-0.16	0.76
Birth per adolescent*	-0.89	0.01**
Percent of arable land	-0.01	0.49
Fertility rate*	1.25	0.20
Maternal mortality ratio*	0.05	0.87
Female labor[1]	-0.02	0.05**
Kilo calorie per day per Capita*	-0.57	0.71

*log-transformed variables; ** significantly associated at 0.05 level
[1] Labor force participation rate for females ages 15-24

Table 2.3 shows the results of goodness-of-fit assessment (DIC and *pD*) and the MAEs for the three models having different structures. No substantial difference was observed regarding both

DIC and MAE among the three models. However, the model fitted with the *GEMS* cluster (Hierarchical B in Table 2.3) has the highest *pD* due to large number of clusters in the data. This model also did not converge well even after a higher number of iterations. We selected the BHM because its structure will allow to 'borrow strength' (i.e., to pool information) across clusters. The latter characteristic is very helpful whenever data are lacking within clusters. Assessment of autocorrelation, density plots and trace plots showed that convergence criteria for the BHM were met.

Table 2.3. Goodness-of-fit and Mean Absolute Errors of three Bayesian models for predicting mortality rates associated with Foodborne diseases.

Model	Model fit		MAE[5]	MAE[6]
	DIC[3]	*pD*[4]	(95% CI)[7]	(95% CI)
Non-Hierarchical	123.9	11.76	0.53(0.43,0.69)	0.65(0.53,0.82)
Hierarchical A[1]	123.7	12.70	0.52(0.42,0.69)	0.66(053,0.83)
Hierarchical B[2]	126.1	15.04	0.51(0.41,0.67)	0.66(0.53,0.84)

Notes

[1] Four cluster random effect; [2] GEMS cluster random effect; [3] Deviance Information Criteria;

[4] Effective number of parameters (measure of model complexity); [5]Mean Absolute Error obtained from the fitted model; [6]MAE obtained from the model after 'Leave One Out' Cross validation; [7]Bootstrap 95% confidence interval

Sensitivity Analysis and Prediction of Mortality Rates

Table 2.4 indicates the range and specifications of priors for the variance and mean components to evaluate robustness of the BHM. Varying the priors did not substantially change the predictions

of log-total mortality rates for countries within Cluster 1 (Figure 2.6a) and Cluster 4 (Figure 2.6d). These clusters contain countries with information regarding mortality rates. On the other hand, the priors has substantially influenced the uncertainty of predictions within clusters lacking any mortality rate data, i.e., Cluster 2 (Fig 2.6b) and Cluster 3 (Fig 2.6c).

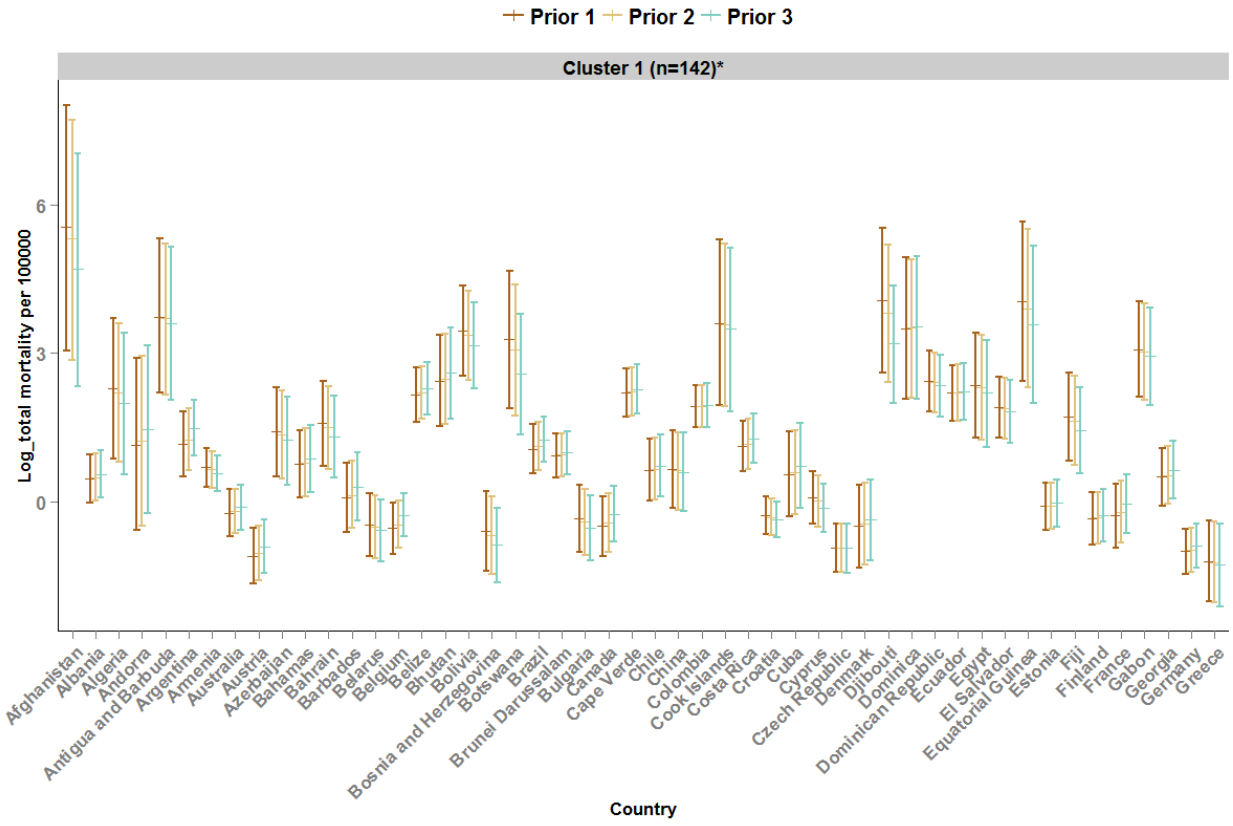
Table 2.4. Prior parameter values employed on the Bayesian hierarchical model for sensitivity analysis.

Priors	Variances	Means	
Set 1	$1/\sigma^2$~dgamma(10^{-3},10^{-3})	b0~dnorm(0, 10^{-3})	b~dnorm(0, 10^{-3})
Set 2	$1/\sigma^2$~dgamma(10^{-2},10^{-2})	b0~dnorm(0, 10^{-2})	b~dnorm(0, 10^{-2})
Set 3	$1/\sigma^2$~dgamma(10^{-1},10^{-1})	b0~dnorm(0, 10^{-1})	b~dnorm(0, 10^{-1})

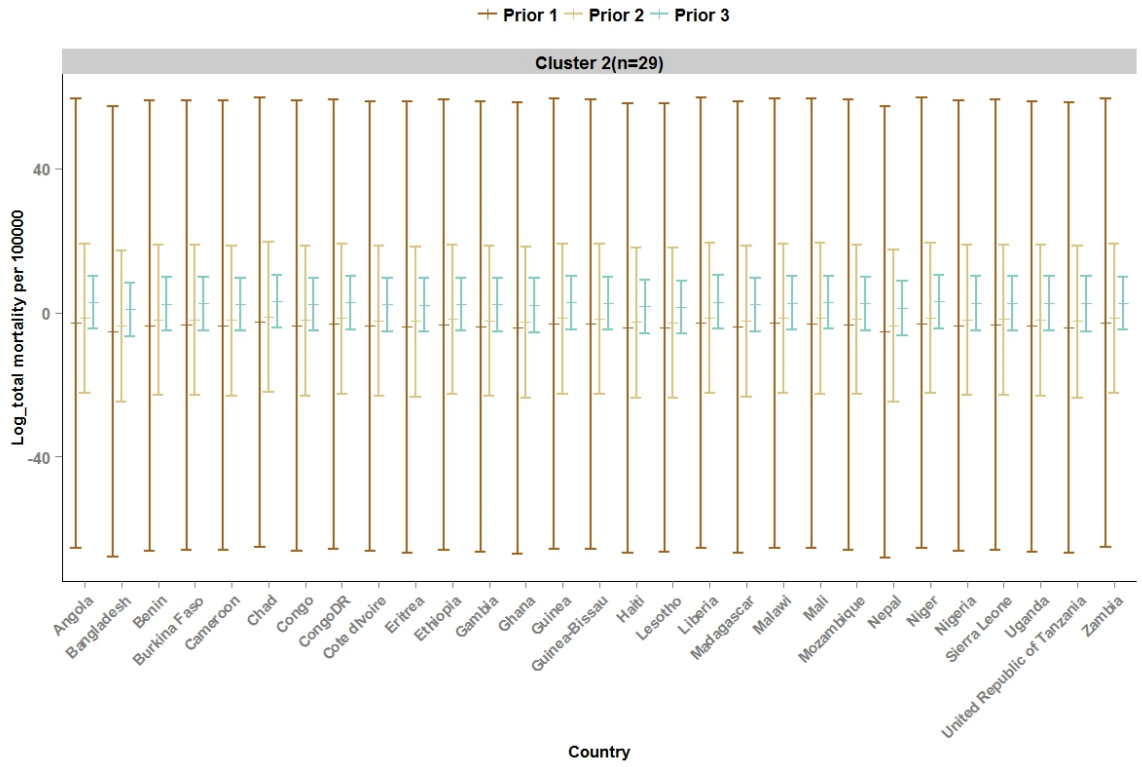
Notes: $1/\sigma^2$: precision (inverse of the variance). In JAGS model, priors for variances are specified by precision. Gamma distribution is frequently used to specify priors for precision. *b0*: Average log-total mortality rate (intercept); *b*: priors for regression coefficients.

Figure 2.4. (a, b, c, d). Sensitivity analysis of the median and 95% Credible Intervals of log-total mortality predictions of the Bayesian Hierarchical Model using three priors.

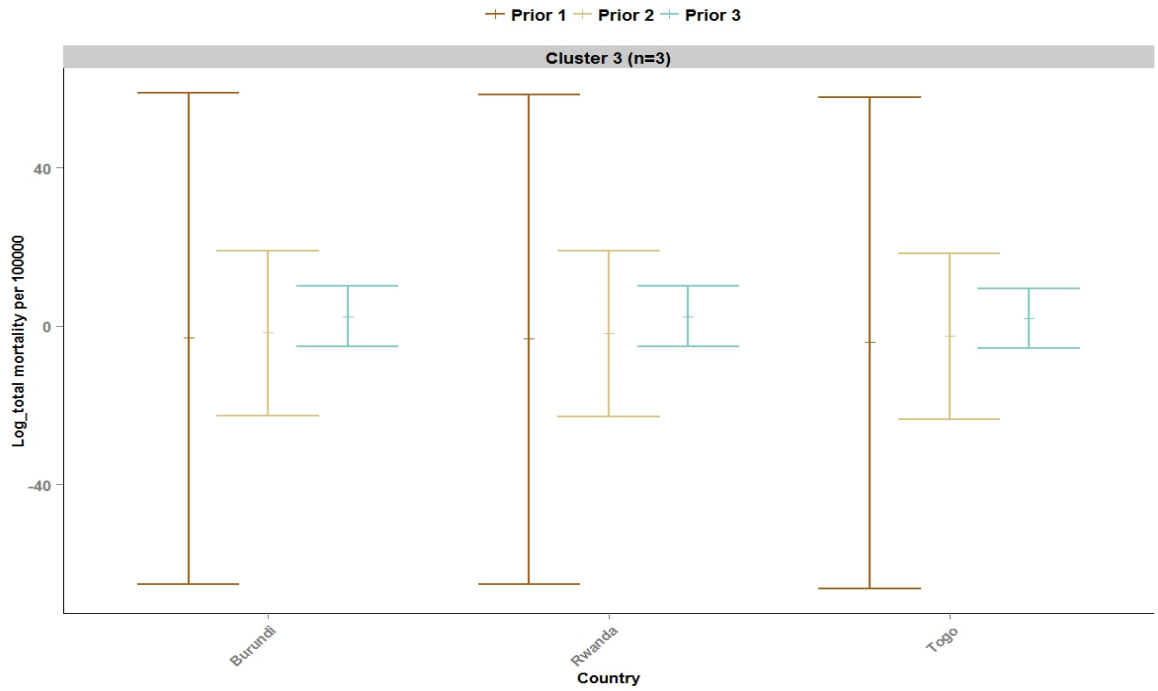
(a)



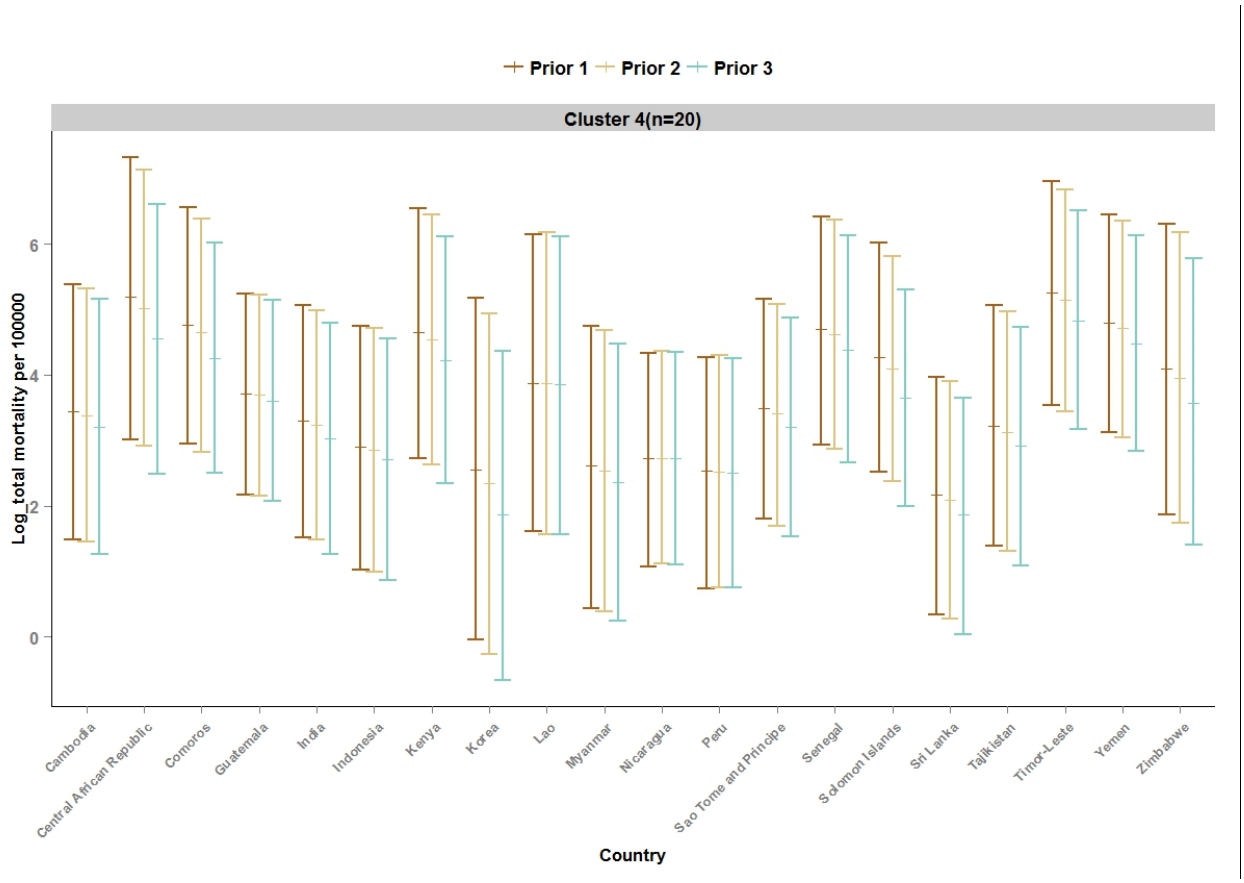
(b)


Prior 1
Prior 2
Prior 3
Cluster 2(n=29)
Log_total mortality per 100000
40
0
-40
Angola
Bangladesh
Benin
Burkina Faso
Cameroon
Chad
Congo
CongoDR
Cote dIvoire
Eritrea
Ethiopia
Gambia
Ghana
Guinea
Guinea-Bissau
Haiti
Lesotho
Liberia
Madagascar
Malawi
Mali
Mozambique
Nepal
Niger
Nigeria
Sierra Leone
Uganda
United Republic of Tanzania
Zambia
Country

(c)


Prior 1
Prior 2
Prior 3
Cluster 3 (n=3)
Log_total mortality per 100000
40
0
-40
Burundi
Rwanda
Togo
Country

(d)



Note: The four figures represent the four clusters. Since the predictions are stable for all countries in Cluster 1 (Fig 2.4a) while using the three priors, only the values of 50 countries are shown for optimal display of this Cluster.

*The numbers in parenthesis indicate the number of countries that belong to the cluster
Prior 1: precision ~ dgamma(10^{-3},10^{-3}); mean~dnorm(0, 10^{-3})
Prior 2: precision ~ dgamma(10^{-2},10^{-2}); mean~dnorm(0, 10^{-2})
Prior 3: precision ~ dgamma(10^{-1},10^{-1}); mean~dnorm(0, 10^{-1})

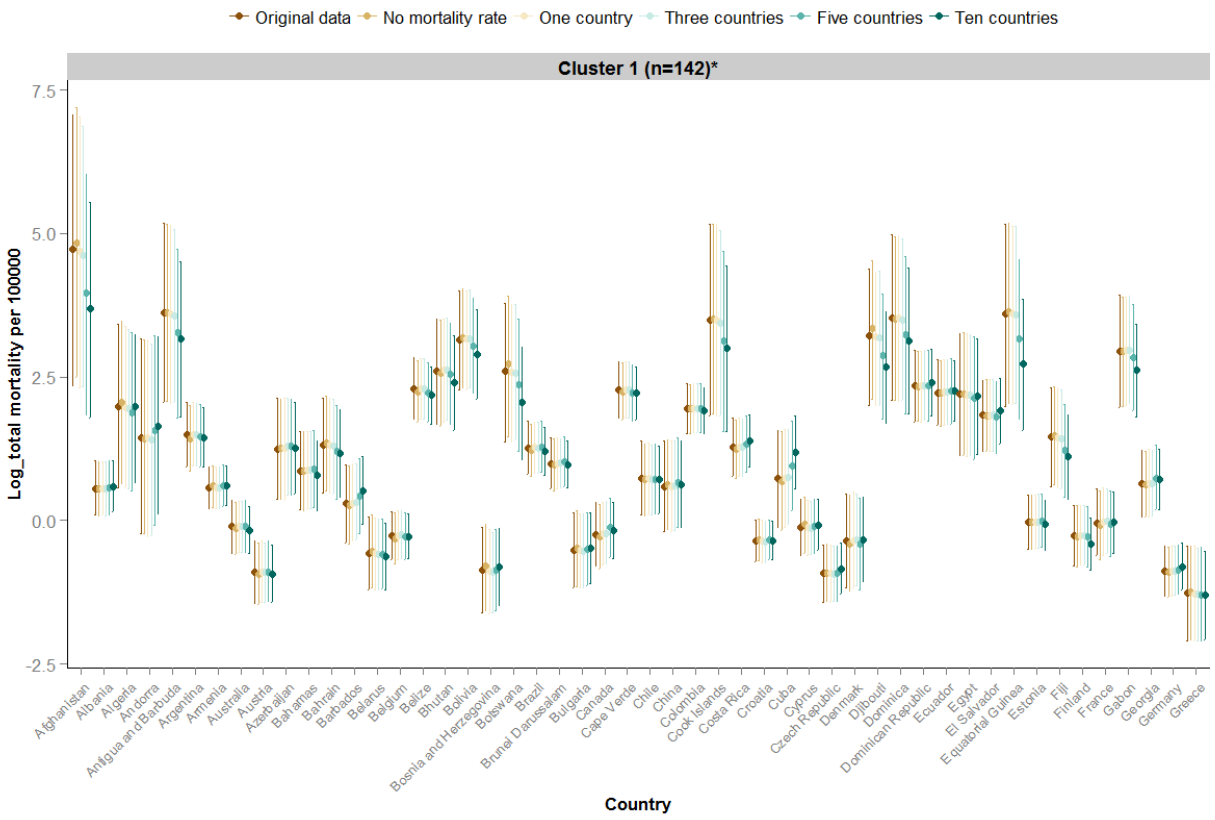
We set the same prior distribution for all clusters (also called exchangeable priors) and specified a smaller values for the variance of the means and precision parameters (Prior set 3) for the final prediction. Constraining the parameter values to be within – 10 and 10 (for example, specifying the prior of *b0* at *dnorm(0,0.01)*) is not a serious restriction. Since the model is on a log-scale, it is

not possible that we will see values as extreme as -10 or 10, which corresponds to a mortality rate of e^{-10} or e^{10} per 100, 000 population.

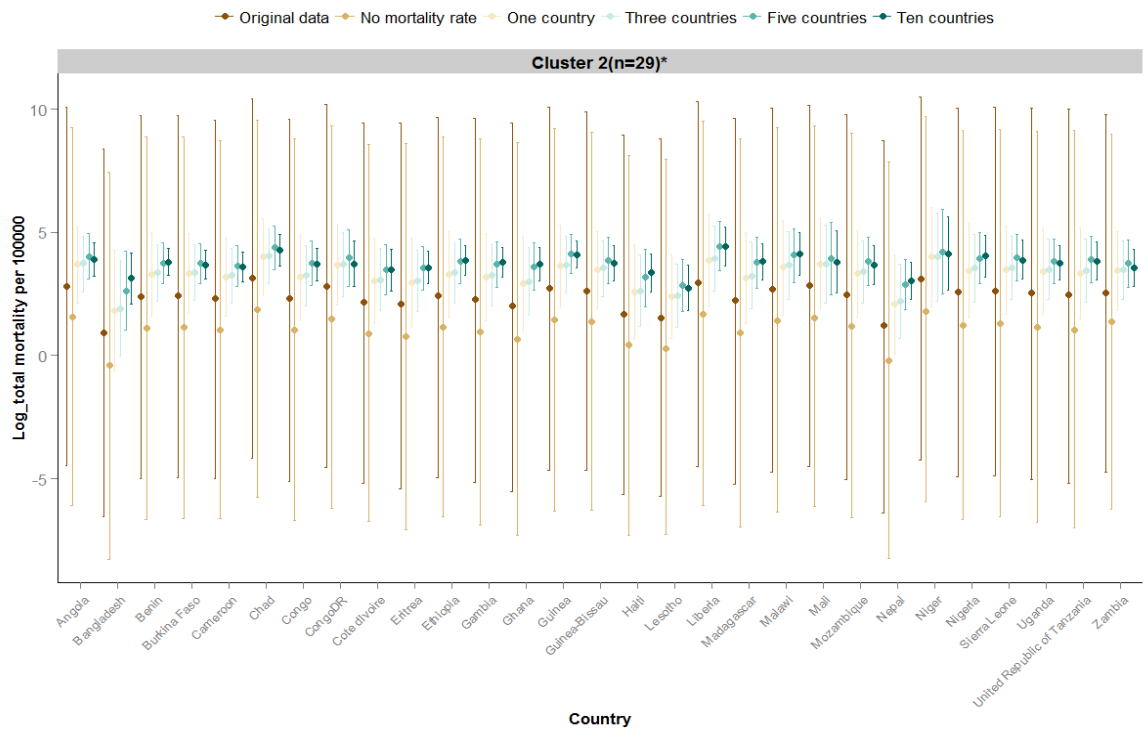
Deleting the observed mortality rate from Cluster 4 (i.e. removing the mortality rate values of Guatemala) and refitting the model resulted in very large uncertainty of predicted log-total mortality rates for all countries that belong to this cluster (Fig 2.5d). On the other hand, randomly adding plausible values for a subset of countries in Cluster 2 (Fig 2.5b) and Cluster 3 (Fig 2.5c), i.e., clusters that lack any observed mortality rate data, considerably reduced the uncertainty of predictions. This indicates that uncertainty is highest for clusters with little or no information and the predicted mortality rate for countries lacking the data can be substantially improved if mortality rate values are obtained for a few countries in the cluster. The change in predicted log-total mortality was minimal for countries in Cluster 1 (Fig 2.5a) whenever mortality rate values were added or deleted from the other clusters as part of the sensitivity analysis.

Figure 2. 5. (a, b, c, d): Comparison of the median and 95% Credible Intervals of log-total mortality predictions of the Bayesian Hierarchical Model with regard to deleting, and randomly adding mortality rates for a subset of countries.

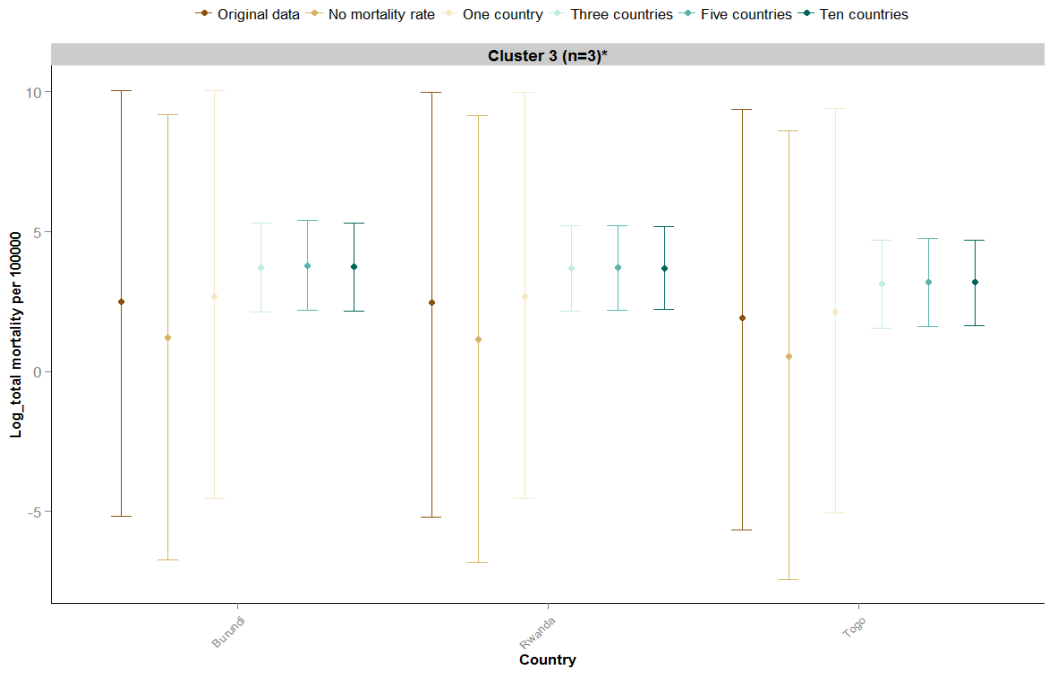
(a)



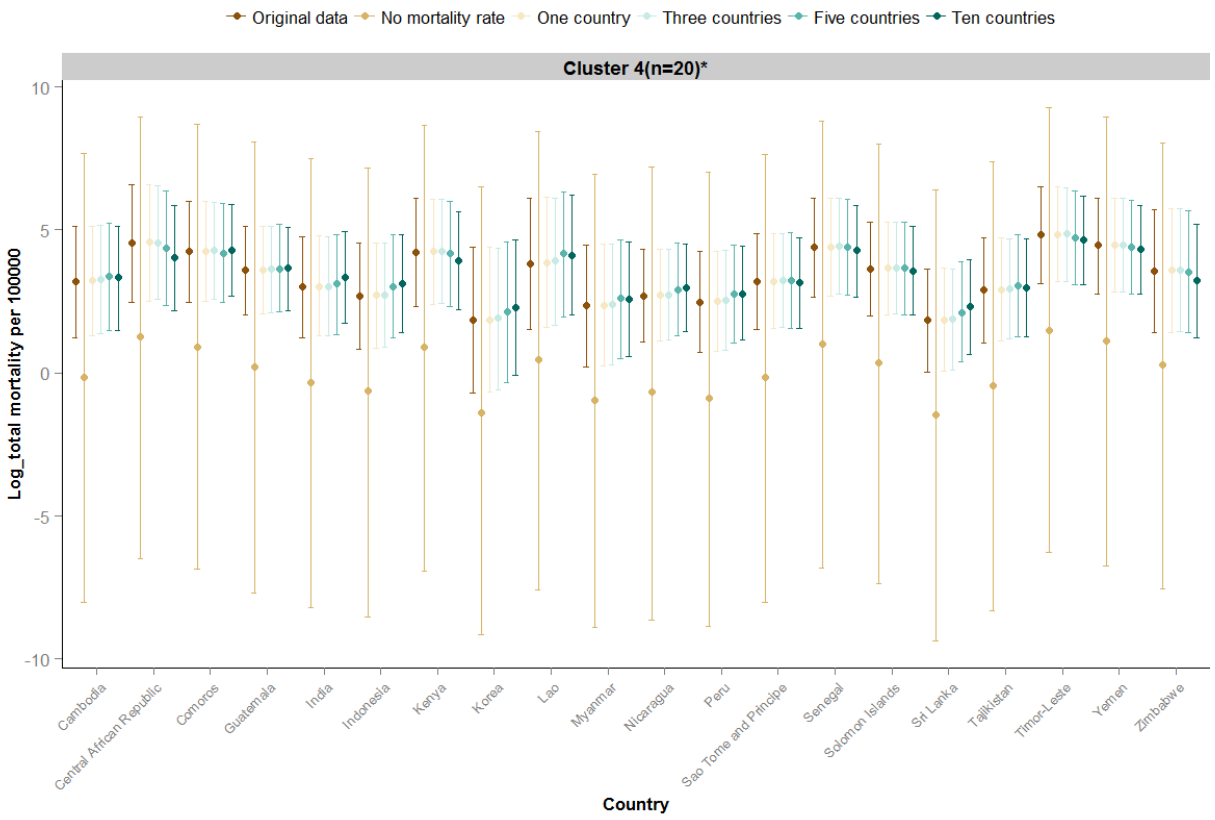
(b)


Original data
No mortality rate
One country
Three countries
Five countries
Ten countries
Cluster 2(n=29)*
Log_total mortality per 100000
10
5
0
-5
Angola
Bangladesh
Benin
Burkina Faso
Cameroon
Chad
Congo
CongoDR
Cote d'Ivoire
Eritrea
Ethiopia
Gambia
Ghana
Guinea
Guinea-Bissau
Haiti
Lesotho
Liberia
Madagascar
Malawi
Mali
Mozambique
Nepal
Niger
Nigeria
Sierra Leone
Uganda
United Republic of Tanzania
Zambia
Country

(c)


Original data
No mortality rate
One country
Three countries
Five countries
Ten countries
Cluster 3 (n=3)*
Log_total mortality per 100000
10
5
0
-5
Burundi
Rwanda
Togo
Country

(d)



Notes: (1)The four figures represent the four clusters. Since the predictions are stable for all countries in Cluster 1 (Fig 2.5a) regardless of deleting and adding values, only the values of 50 countries out of 142 are shown for optimal display of this Cluster. These four figures depict the change in uncertainty of predictions whenever new information is added or some information is deleted from the dataset. (2) Subsets of countries were randomly selected from Cluster 2 (Fig 2.5b) and Cluster 3 (Fig 2.5c) that lack mortality rate data. Then a log-total mortality rate of 3.75 was assigned to each of them before a model was fitted. The 3.75 value is an observed mortality rate of Guatemala, which has the largest value among all countries that has information on mortality rates. Note that each plot in the above figure has different scales on the y-axis to optimally display the 95% Credible Intervals.

Keys to the legend:
Original data: all observed data included in model predictions.
No mortality rate: model predictions after deleting the value of Guatemala from Cluster 4.
One country: model predictions after a hypothetical mortality rate was assigned to one country (Angola)
Three countries: model predictions after a hypothetical mortality rate was assigned to three countries (Angola, Guinea and Rwanda)
Five countries: model predictions after a hypothetical mortality rate was assigned to five countries (Angola, Guinea, Rwanda, Uganda and Nepal)
Ten countries: model predictions after a hypothetical mortality rate was assigned to ten countries (Angola, Guinea, Rwanda, Uganda, Nepal, Benin, Bangladesh, Madagascar, Chad, and Burkina Faso)

*numbers in parenthesis indicate the number of countries that belong to the cluster

The final predicted log-total mortality rates for all countries using the BHM is shown in Appendix 2.2. None of the countries in Cluster 2 and 3 had observed mortality rates. For countries within these clusters, the BHM predicted wide 95% CIs for the median of log-total mortality rates indicating large uncertainty. On the other hand, the uncertainty of predictions were very small for countries within Cluster 1 and Cluster 4. The overall median and 95% CI of the predicted log-total mortality rate ranges from -1.23 (-2.03, -0.44) for Greece to 5.04 (2.68, 7.36) for Afghanistan, which when exponentiated will yield the median mortality rate of 0.29 (0.13, 0.63) and 155.19 (14.66, 1572.85) per 100,000 population, respectively. As indicated in Appendix 2.2, some of the 95% CIs of predictions did not contain the observed mortality rates. This could be due to the unusually high rate of missing mortality rate in the dataset.

2.4 Discussion

Lack of sufficient and complete data on mortality and morbidity from many countries have been addressed previously [15]. This scarcity of data is an ongoing challenge for the estimation of global burden of FBDs. Although most of the observed mortality rates in our dataset are from developed countries, which have a comparatively low burden of FBD, it is possible that similar potential risk factors (predictors) may be shared between developed and developing countries. The eight predictors we selected in this report are proxy attributes to capture the socio-economic, food-production, hygiene and health status of countries. This is in agreement with a study regarding variable selection to estimate the missing incidence of specific foodborne diseases [40]. In

addition, a frequentist-based analysis of part of the current dataset highlighted that both health and non-health related variables can be used as a proxy predictors to measure mortality associated with FBDs [14].

Grouping countries based on the values of these predictors is a novel attempt to create data-driven clusters of countries with a homogenous mortality rate. WHO countries have been previously grouped based on geographical attributes (e.g., WHO sub-regions) or other parameters depending on the goal of the classification scheme. For example, the GEMS Food Cluster is designed to group countries based on food consumption and risk assessment [31]. Comparison of predictions and model fits using the GEMS cluster and our data-driven clusters as random effects indicated that the model with the GEMS cluster lacks convergence while our four cluster solution has converged and fits the data well. In a recent study to estimate missing national level incidence of specific food foodborne diseases, the WHO sub-regions and the food cluster regions were used as random effects [40]. In our study, 47 of the 48 countries with complete mortality rate data have been grouped into Cluster 1, which indicates that those countries which routinely report mortality rates have similarities in the eight predictor values.

While analyzing missing data, it is important that the mechanism of missingness in the data is understood [41]. Although the MAR assumption, as such, is not testable, we justified its validity for our dataset by demonstrating association of one or more of the predictors with the missingness of mortality rates. This implies that the missingness can be partly explained by the predictors in the dataset. Meanwhile, it is also important to note that any other missing data analysis approaches require assumptions that are just as difficult to justify.

In this study, the choice of the best fit model to use for prediction of mortality was based on comparison of model fit, out of sample prediction performance and the method's suitability for analysis of missing data with hierarchical structure. To evaluate predictive performance of a model, using the same data as was initially used to build the model may introduce over-fitting problems [42]. In addition, collecting new data to validate the model for predictive strength is not practically feasible, and therefore the LOOCV method solves an over-fitting problem and helps to assess the predictive accuracy. Although we did not observe a substantial difference between the three models regarding MAEs, we preferred the BHM for a number of reasons. The structure of a BHM enables 'borrowing strength' across clusters that improves prediction of mortality rates which is particularly essential whenever data are missing [43]. Moreover, a BHM facilitates the estimation of several parameters over similar units (for example, countries within clusters) in order to improve the precision of the estimated effects for each unit [44]. It has also been described previously that a Bayesian approach allows for a more efficient use of data as the method does not depend on the asymptotic theory of large sample approximation [44]. This is essential whenever there are few observations and a high proportion of missing values in the dataset.

Part of this dataset has been analyzed previously using a classical frequentist framework [14]. Our Bayesian approach, however, has several important advantages over this likelihood based frequentist method. The development of Bayesian models in general offers an opportunity to assign pertinent information (prior) to unknown parameters (including missing values distribution) [34]. This is particularly useful for analysis of a dataset with missing values. Secondly, Bayesian models can be easily updated rationally as new data becomes available. This will mean that future research on FBD can directly utilize the current results as priors. The other reason is that BHM

provide a convenient setting for a dataset with inherent hierarchical structure. In our dataset, countries within clusters are assumed to have more similar mortality rates as compared to the rates between countries across clusters.

Implementing BHM allows pooling of information across clusters, such that clusters with little or no data 'borrow strength' about the log-mortality rates from other clusters. In our analysis, the predicted median mortality rates for countries in Clusters 2 and Cluster 3 were smoothed towards the overall average population estimate (Figure 2.4 and Figure 2.5). Although the predicted median mortality rates are close to the overall average log-mortality rate, the uncertainty is large. This large degree of uncertainty in the prediction is a direct result of data quality or lack thereof. The reduction in uncertainty of the predictions achieved by adding hypothetical but plausible data for a subset of countries has a practical implication. For example, data collection strategy for mortality rates can be based on cluster information. If mortality data can be obtained from a proportion of countries from a properly defined cluster, we can use BHM to predict mortality rate for the remaining countries missing the data, making the best possible usage of all data.

2.5 Conclusions

The extent of the global burden associated with FBD is still unknown and therefore tackling mortality associated with FBD is a continuous global challenge. The difficulty of estimating the burden of FBD is partly due to the lack of information about mortality and morbidity rates associated with FBD in many countries of the world. Therefore, it will be a compromise to use only a fraction of countries with complete information to generalize for the overall global

foodborne mortality rates without using models. A subset of the most meaningful predictors of mortality associated with FBDs can be used to implement a BHM that predicts mortality rates for countries lacking the data. An informative clustering of countries based on this subset of variables can help 'borrow strength' across similar countries using a BHM to predict mortality rates associated with FBD. The high proportion of missing values in the dataset might be the cause for some of the predictions to be outside the observed range. Therefore, the predictions obtained from the final model in this report should always be interpreted with caution. Finally, when resources are limited, the selected variables can provide suggestions for future data collection regarding risk factors of FBD mortality.

Financial Support

This work was supported by the NIH Ruth L. Kirschstein National Research Service Award Institutional Training Grant T32 RR023916 and T32 OD010423

Conflict of interest

The authors declare that they have no any conflict of interests

2.6 References

1. Havelaar AH, Cawthorne A, Angulo F, Bellinger D, Corrigan T, Cravioto A, Gibb H, Hald T, Ehiri J, Kirk M, Lake R, Praet N, Speybroeck N, de Silva N, Stein C, Torgerson P, Kuchenmüller T: WHO Initiative to Estimate the Global Burden of Foodborne Diseases. *The Lancet* 2013, 381:S59.

2. Todd EC: Epidemiology of foodborne diseases: a worldwide review. *World Health Stat Q Rapp Trimest Stat Sanit Mond* 1997, 50:30–50.

3. Käferstein FK: Actions to reverse the upward curve of foodborne illness. *Food Control* 2003, 14:101–109.

4. WHO | Fact sheets [http://www.who.int/mediacentre/factsheets/en/]

5. Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson M-A, Roy SL, Jones JL, Griffin PM: Foodborne illness acquired in the United States--major pathogens. *Emerg Infect Dis* 2011, 17:7–15.

6. Yasmine Motarjemi & Fritz K. Käferstein / Global estimation of foodborne diseases [http://apps.who.int/iris/handle/10665/54779]

7. Kuchenmüller T, Hird S, Stein C, Kramarz P, Nanda A, Havelaar AH: Estimating the global burden of foodborne diseases--a collaborative effort. *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull* 2009, 14.

8. WHO (World Health Organization) | Initiative to estimate the Global Burden of Foodborne Diseases [http://www.who.int/foodsafety/foodborne_disease/ferg/en/index.html]

9. Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, Jones TF, Fazil A, Hoekstra RM: The Global Burden of Nontyphoidal Salmonella Gastroenteritis. *Clin Infect Dis* 2010, 50:882–889.

10. Crump JA, Luby SP, Mintz ED: The global burden of typhoid fever. *Bull World Health Organ* 2004, 82:346–353.

11. Torgerson PR, de Silva NR, Fèvre EM, Kasuga F, Rokni MB, Zhou X-N, Sripa B, Gargouri N, Willingham AL, Stein C: The global burden of foodborne parasitic diseases: an update. *Trends Parasitol* 2014, 30:20–26.

12. WHO (World Health Organization) | First formal meeting of the Foodborne Disease Burden Epidemiology Reference Group (FERG), 26-28 November 2007 [http://www.who.int/foodsafety/foodborne_disease/ferg1/en/]

13. ICD-10: International Statistical Classification of Disease and Health Related Problems. World Health Organization; 2004.

14. Hanson LA, Zahn EA, Wild SR, Döpfer D, Scott J, Stein C: Estimating global mortality from potentially foodborne diseases: an analysis using vital registration data. *Popul Health Metr* 2012, 10:5.

15. Jahan S: Epidemiology of Foodborne Illness. *Lancet* 2012, 336.

16. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria [http://www.r-project.org/]

17. Plummer M, others: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*; 2003:20–22.

18. Taylor R: Interpretation of the Correlation Coefficient: A Basic Review. *J Diagn Med Sonogr* 1990, 6:35–39.

19. Zou H, Hastie T: Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol* 2005, 67:301–320.

20. Friedman J, Hastie T, Tibshirani R: Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010, 33:1.

21. Sill M, Hielscher T, Becker N, Zucknick M: c060: Extended Inference for Lasso and Elastic-Net Regularized Cox and Generalized Linear Models. *R Package Version 02 URL HttpCRAN R-Proj Orgpackage C060* 2013.

22. van Buuren S: *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press; 2012.

23. Schafer JL: Multiple imputation: a primer. *Stat Methods Med Res* 1999, 8:3–15.

24. Graham JW, Olchowski AE, Gilreath TD: How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prev Sci* 2007, 8:206–213.

25. Azur MJ, Stuart EA, Frangakis C, Leaf PJ: Multiple Imputation by Chained Equations: What is it and how does it work? *Int J Methods Psychiatr Res* 2011, 20:40–49.

26. Borcard D, Gillet F, Legendre P: *Numerical Ecology with R*. Springer; 2011.

27. Gower JC: A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 1971, 27:857–871.

28. Saraçli S, Doğan N, Doğan İ: Comparison of hierarchical cluster analysis methods by cophenetic correlation. *J Inequalities Appl* 2013, 2013:1–8.

29. Tibshirani R, Walther G, Hastie T: Estimating the number of clusters in a data set via the gap statistic. *J R Stat Soc Ser B Stat Methodol* 2001, 63:411–423.

30. Rousseeuw PJ: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987, 20:53–65.

31. Global Environmental Monitoring System (GEMS) Food : Report of the WHO working group on Collection Of Food Consumption data (COFOCO), 2012.

32. Mason A, Best N, Richardson S, PLEWIS I: Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. 2010.

33. Rubin DB: Inference and missing data. *Biometrika* 1976, 63:581–592.

34. Gelman A, Hill J: *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 1st edition. Cambridge University Press; 2006.

35. Spiegelhalter DJ, Best NG, Carlin BP: *Bayesian Deviance, the Effective Number of Parameters, and the Comparison of Arbitrarily Complex Models*. Division of Biostatistics, University of Minnesota; 1998.

36. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A: Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol* 2002, 64:583–639.

37. Stone M: Cross-Validatory Choice and Assessment of Statistical Predictions. *J R Stat Soc Ser B Methodol* 1974, 36:111–147.

38. Willmott CJ, Matsuura K: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res* 2005, 30:79–82.

39. Kilian R, Matschinger H, Loffler W, Roick C, Angermeyer MC: A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment. *J Ment Health Policy Econ* 2002, 5:21–32.

40. McDonald SA, Devleesschauwer B, Speybroeck N, Hens N, Praet N, Torgerson PR, Havelaar AH, Wu F, Tremblay M, Amene EW, Döpfer D: Data-driven methods for imputing national-level incidence in global burden of disease studies. *Bull World Health Organ* 2015, 93:228–236.

41. Little RJA, Rubin DB: *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, N.J: Wiley; 2002. [*Wiley Series in Probability and Statistics*]

42. Arlot S, Celisse A: A survey of cross-validation procedures for model selection. *Stat Surv* 2010, 4:40–79.

43. Hong H, Carlin BP, Chu H, Shamliyan TA, Wang S, Kane RL: A Bayesian Missing Data Framework for Multiple Continuous Outcome Mixed Treatment Comparisons. 2013.

44. Congdon PP: *Bayesian Statistical Modelling*. John Wiley & Sons; 2007.

Chapter 3: Filling gaps in notification data: a model-based approach applied to travel related campylobacteriosis cases in New Zealand

Amene, E.[1], Horn, B.[2], Pirie, R.[2], Lake, R.[2] and D.,Döpfer [1]

[1]University of Wisconsin-Madison, School of Veterinary Medicine, Department of Medical Sciences, USA,

[2]Institute of Environmental Science and Research, Christchurch, New Zealand

Abstract

Background: Information from notified cases of disease has the potential to inform the epidemiology and risk factors for the disease. However, the value of this information is often compromised by incomplete or partial information related to individual cases. In an effort to enhance the value of information from enteric disease notifications in New Zealand, this study explored the use of fully Bayesian and Multiple Imputation models to fill risk factor data gaps. As a test case, overseas travel as a risk factor for infection with campylobacteriosis has been examined.

Methods: Eleven years of short term international travel and campylobacteriosis notification data were obtained from New Zealand national databases. Two methods, namely Fully Bayesian Specification (FBS) and Multiple Imputation (MI), were compared regarding predictive performance for various levels of artificially induced missingness of overseas travel status in the dataset. Predictive performance of the models was assessed through Brier Score, Area Under the ROC Curve, and Percent Bias of regression coefficients. Finally, the best model was selected and applied to predict missing overseas travel status of campylobacteriosis notifications.

Results: There was no difference in the predictive performance of the FBS and MI methods at a lower rate of missingness (<10%), but the FBS approach performed better than MI at a higher rate of missingness (50%, 65%, 80%). The estimated proportion (95% CI) of cases associated with overseas travel was greatest in highly urban District Health Boards (DHBs) in Counties Manukau, Auckland and Waitemata, at 0.37 (0.12, 0.57), 0.33 (0.13, 0.55) and 0.28 (0.10, 0.49), whereas the lowest proportion was estimated for more rural West Coast, Northland and Tairawhiti DHBs at 0.02 (0.01, 0.05), 0.03 (0.01, 0.08) and 0.04 (0.01, 0.06), respectively. The added advantage of using a Bayesian approach is the model's prediction can be improved whenever new information becomes available.

Conclusion: We propose the use of FBS, which offers a flexible approach for data augmentation particularly when the missing rate is very high and when the Missing At Random (MAR) assumption holds. High rates of travel associated cases in urban regions of New Zealand predicted by this approach are plausible given the high rate of travel in these regions, including destinations with higher risk of infection.

Key words: Campylobacteriosis, Fully Bayesian Specification, Multiple Imputation, missing value

3.1 Background

Information originating from investigation of notified cases of an infectious disease has the potential to inform about the epidemiology and risk factors associated with the disease. Aggregating demographic and risk factor information from surveillance systems can help to set policy, monitor trends, and develop risk management options. However, the value of this information is often compromised by incomplete or partial information related to individual cases.

In New Zealand, cases of notifiable diseases are reported by general practitioners, laboratories and public health workers and the information is stored in the *EpiSurv* database. *EpiSurv* is operated by the Institute of Environmental Science and Research (ESR) on behalf of the Ministry of Health. A series of case report forms (https://surv.esr.cri.nz/episurv/index.php) are used to collect information about cases, disease diagnosis and clinical course, risk factors for the disease and case management.

Campylobacteriosis is the most common cause of acute gastrointestinal illness in developed countries including New Zealand. The principal species causing the disease in humans are *C. jejuni* and to a lesser extent *C. coli*. The complex epidemiology of campylobacteriosis and the presence of less understood risk factors such as overseas travel, has hindered the development of successful control programmes in many industrialized nations. Campylobacteriosis has been a notifiable disease in New Zealand since 1980. Data from notified cases are reported annually in surveillance summaries and have been analyzed for trends and to assess the effect of specific interventions [1, 2]. These analyses are primarily based on demographic information, since for a variety of reasons the risk factor information is not supplied for all cases. However, the value of complete information on cases has been demonstrated by a sentinel site study in the Manawatu region of

New Zealand, which has made a special effort to complete risk factor reporting, alongside microbial subtyping [3].

In an effort to enhance the value of information from campylobacteriosis notifications in New Zealand, we have explored the use of models to fill risk factor data gaps. As a test case, we examined overseas travel as a risk factor for campylobacteriosis. Identifying the proportion of cases of campylobacteriosis where infection was acquired overseas is important to properly understand and measure domestic risk factors and the success of any risk management interventions [4]. International travel as a risk factor is important, as the rate of overseas travel by New Zealanders is high (e. g. 46 trips per 100 per year as compared to the international average of 14 per 100 in 2008) [5, 6]. However, whether (or not) cases had travelled overseas as a potential risk factor is reported for less than half of the notified cases of campylobacteriosis, and the reporting of this factor varies considerably across the 20 District Health Boards (DHBs) in New Zealand (see map of New Zealand in Appendix 3.1). One approach to adjusting for this lack of data, as currently used in annual surveillance reports, is to apply the proportion travel related from the cases for which information is available. This approach estimates that approximately 7% of campylobacteriosis notifications nationally over the period 2000 to 2010 were acquired overseas. However this information may be biased for a variety of reasons and does not fully reflect regional variation.

As an alternative, we applied Multiple Imputation (MI) [7] and Fully Bayesian Specification (FBS) [8] models, seeking to adjust rates of travel associated illness and fill data gaps using covariates derived from demographic characteristics and travel rates in the general New Zealand population.

3.2 Methods

Empirical data

Campylobacteriosis Notifications

Campylobacteriosis notification records were obtained from the *EpiSurv* database [9]. All case notifications were completely anonymized to conceal the identity of individuals. The database registers a number of demographic and risk factor characteristics of the cases in addition to clinical features. This project extracted the following factors; *if case had travelled overseas during the incubation period*, *countries visited during travel*, *reporting region* in New Zealand (at District Health Board, DHB, level) , *report date*, *age* and *gender*.

Overseas Travel

A custom data extract for short term international travel of New Zealand residents between 2000 and 2010 was obtained from Statistics New Zealand [10]. The data included the travel patterns of New Zealand residents to international destinations including: countries visited, month of arrival back in New Zealand, and demographic characteristics of travelers. Short term travel is defined as international departures of New Zealand residents for an intended period of less than 12 months. The annual number of overseas trips for a region was calculated at the DHB level. The total number of trips during the study period (2000-2010) was divided by the average DHB population estimates during the same period as extracted by Statistics New Zealand [10] to provide a population rate (TRAVEL RATE) which was used in the modelling. Individuals may have traveled multiple times,

however, this information could not be taken into account because the database does not identify individuals who have traveled more than once.

Urban or rural lifestyle

Statistics New Zealand [10] used seven definitions to define urban and rural New Zealand in 2006; main urban area, satellite urban community, independent urban community, rural area with high urban influence, rural area with moderate urban influence, rural area with low urban influence and highly remote/rural area. The 2006 census data have been analyzed by Statistics New Zealand to provide population estimates in each of these urban, rural categories. The proportion of a DHB's population living in one of the three urban areas, i.e. *main urban*, *satellite urban* and *independent urban* (URBAN) is used as a DHB level variable in the modelling. The *main urban* areas are those having a minimum population of 20,000 and the *satellite urban* areas identifies towns and settlements with strong links to main urban areas through employment. The *independent urban* areas on the other hand are those whose population is without a significant dependence on main urban centers through employment location [10].

Deprivation Index

New Zealand Deprivation Index (DI) is a measure of socioeconomic deprivation which combines certain variables from the 2006 census reflecting eight dimensions of deprivation. These include income, home ownership, need of support, employment, educational qualifications, the amount of living space, access to communication and access to transport [11]. The *DI* is available for a subpopulation of a DHB in an ordinal scale from 1 to 10, where 1 represents the least deprived

areas and 10 the areas with the most deprived scores [11]. For the subsequent modeling, a DHB level *DI* is required. For this reason, we weighted the *DI* with the subpopulation size to get a weighted median of *DI* for the total DHB population.

Poultry Intervention

In 2007 interventions were put into place by the New Zealand Poultry Industry to reduce *Campylobacter* in retail poultry alongside the introduction of a regulatory *Campylobacter* Performance Target. These activities include routine monitoring and testing of poultry processors for the prevalence of *Campylobacter* spp by using cecal testing and strict hygienic practices throughout the production and processing stages. Mandatory *Campylobacter* performance targets were introduced based on test results and escalated regulatory responses were put in place if the targets were not met [2]. This resulted in a drop in campylobacteriosis notifications from 15,728 in 2006 to 6,594 in 2008[2]. As a domestic intervention, the total decrease in campylobacteriosis cases would not have affected the number of cases resulting from overseas travel. A binary variable (INTERVENTION) was added to the model to allow the sudden drop in notifications to be incorporated into the model.

A complete description of predictor variables in the notification and travelers' dataset which were used for this analysis is shown in Table 3.1. While *Deprivation index*, *Urban* (population under urban influence), *DHB* and *Travel Rate* are variables at a reporting region level; *Age*, *Sex*, *Season* and *Intervention* (whether the case was recorded before or after 2006) are case specific variables.

Table 3.1. The description of variables in New Zealand campylobacteriosis notification and short term international travelers' dataset (2000-2010).

Variables	**Details**
DEPRIVATION INDEX	Categorical, 1-10 scale (1=least deprived, 10=most deprived)
URBAN	Numeric, Proportion of DHB population under urban influence
DHB	Categorical, Residence District Health Board
TRAVEL RATE	Numeric, Residence DHB's rate of short term international travel
REPORT DATE	Year of campylobacteriosis notification, 2000-2010
AGE	Four categories, <5, 5-19, 20-65 and 65+ Years
SEX	Two categories, Male and Female
SEASON	Four categories, Spring (Sep-Nov), Summer (Dec-Feb), Autumn (Mar-May) & Winter (Jun-Aug)
OVERSEAS TRAVEL	Three categories, Yes, No, Unknown [The status of short term overseas travel, 62% of the cases did not have travel information.]
INTERVENTION	A binary indicator variable to identify before and after the 2006 poultry intervention period.

Notes: (1) *Deprivation index*, *Urban*, *DHB* and *Travel Rate* are DHB level variables, whereas *Report Date, Age, Season, Overseas Travel* and *Intervention* are measured at an individual case level.

Statistical Analysis

Model development

Initially, analysis was restricted to *Campylobacter* notifications with complete information on overseas travel status and the predictors. This subset of the dataset included 38% (44,285) of all notifications reported between 2000 & 2010. The remaining 62% (72,436) lack travel information. The reason for performing this restricted analysis was to select the best prediction model based on cases with complete data.

Next, we investigated performance of the two approaches, namely MI and FBS for different rates of artificially introduced missingness into the dataset (10%, 50%, 65% & 80% of the data were randomly deleted).

Application of MI and FBS requires that missingness in the data is Missing At Random (MAR). MAR assumes the probability of missingness only depends on the observed variables in the data, i.e. in our analysis the probability of missing overseas travel depends on measured covariates in the dataset, but does not depend on whether or not an individual has actually made overseas travel. A detailed description of types of missingness can be found in literature [12]. While the MAR assumption, as such, is not testable, it can be supported by demonstrating association of predictors with the missingness of the variable of interest. We performed the MAR test by creating a dummy variable for missing overseas travel (i.e. missing overseas travel=1, otherwise= 0), fitting a logistic regression model of this variable on the predictors and checking for a statistical significance of the association. A strong association indicates that the missingness in the system can be MAR.

In order to achieve the required percentage of missing values, we deleted a proportion of the travel status from the complete dataset. This is carried out by taking into account the association between the predictors and travel status to assure the validity of the MAR assumption in the resulting missing data. Since adding as many predictors as possible into the imputation model helps strengthen the MAR assumption, we used all of the variables in the imputation model. However,

we dropped residence *DHB* from the analysis model due to its high correlation with DHB level factors such as *URBAN, DI and TRAVEL.* Finally, based on the models' performance parameters, we selected the best model and applied it for predicting overseas travel status in the full dataset.

Complete case analysis

Complete case (CC) analysis refers to analysis restricted to cases with fully reported travel status (disregarding missing values). The subsequent missing data analysis was compared to the results of the CC analysis. We fitted a multiple logistic regression model to predict *overseas travel* on the above predictors. Frequentist and Bayesian logistic regression frameworks were applied to these data. The generalized form of the logistic model is shown below *(3.1)*.

$$\text{Log}\left(p(x)/1\text{-}p(x)\right) = \beta_0 + X_i\beta_i \quad i=1,\ldots, n \ (n\text{=number of predictors}) \qquad (3.1)$$

Where $p(x)$ is the probability that a case made short term overseas travel, β_o and β_i are the regression coefficients and X_i denotes the predictors.

Multiple Imputation

Multiple Imputation is a principled way of handling incomplete data where missing observations are replaced by draws from the predictive distribution of the missing data given the observed data [13, 14]. All potential predictors available in the dataset (Table 3.1) were incorporated into the imputation model. Including all variables predictive of overseas travel will help the MAR assumption to be increasingly plausible, in addition to producing unbiased results [15, 16]. This is because subjects with missing data based on (other) known characteristics, i.e. MAR- are by definition a random subset from the sample given these known characteristics [17]. Among all the

predictors in the data, 0.6% of *Age* and 1.6% of *Sex* have missing values (Table 3.1). Since the proportion of missing values in these two variables is very low (i.e. less than 2%), we deleted records associated with missing observations, and restricted the analysis to the remaining fully reported case records [13].

Implementation of MI on each category of missingness was performed using the *R* package, *MICE* (Multiple Imputation using Chained Equations) [7]. The MI procedure creates *m* numbers of so called complete multiply datasets. Simulation studies have shown that as few as 3 multiply datasets are adequate for a dataset with 20% missing values [18]. Other studies have shown that 5-10 multiply datasets are usually optimum depending on the proportion missing [7]. In this study, we have used 20 multiply datasets. Convergence of the imputation models was visually examined by assessing density plots of predicted overseas travel status [7]. Each complete dataset was analyzed separately with a standard logistic regression method and then point estimates and standard errors were pooled according to Rubin's rule [19]. We used the pooled regression coefficients to construct a logistic regression equation for predicting the probability of overseas travel.

Fully Bayesian Specification (FBS)

While MI was derived from within a Bayesian framework (sampling from the posterior distribution of missing values, conditional on observed values), Bayesian approaches have been applied more generally [20]. Bayesian full probability modelling provides a flexible method for incorporating different assumptions about the missing data mechanism and accommodating different patterns of missing data in the model [21].

As in the MI approach, the FBS also assumes that missingness in the data is MAR. A separate logistic regression model of overseas travel on the predictors indicated above was fitted to each dataset. For Bayesian analysis, we used the *JAGS 3.4.0.* program (Just Another Gibbs Sampler), which is called into the R environment through *rjags* package[22]. The use of a Bayesian method requires that the priors of unknown parameters to be specified properly [23]. This is a way of incorporating uncertainty about the parameters into the model. For our analysis, all regression coefficients and the intercept were assigned uninformative priors (a normal distribution with mean 0 and standard deviation of 100, i.e. each with an inverse variance of 10^{-4}). This implies that the regression coefficients are expected to lie within a range from -100 to 100 (Appendix 3.2). For computational reasons, Bayesian models in JAGS require the variance to be specified in terms of the precision (inverse of the variance). The models were run for 30,000 iterations with the first 3000 iterations discarded as burn-ins. All models were initialized with two chains. For realistic starting values, we set the initial values for each chain obtained from the fitted regression coefficients (see Appendix 3.2).

Bayesian inference including FBS relies on Markov Chain-Monte Carlo (MCMC) algorithm to draw samples from the posterior distribution. This implies that convergence of the algorithm has to be assessed. Convergence indicates that the samples from the MCMC process are, in fact, drawn from the actual joint posterior distribution of the parameters. We have assessed convergence of the Bayesian models through visually evaluating density plots, autocorrelation and BGR statistic of the parameters in the model. Furthermore, we plotted the *jags* output to check the values of 'Rhat'. The 'Rhat' value is a measure of convergence that takes into account the different starting values

for the different chains [23]. It should have a value close to 1 for declaring convergence. More iterations were run whenever the model appeared not to be converging.

Model evaluation and performance

We evaluated our models by comparing *PB* and *BS* of regression coefficients and predictions, respectively. The PB indicates the percent deviation of the regression coefficients of models fitted to the missing data as compared to those estimated by the fully observed dataset (i.e. Complete Cases) *(3.2)*. Note that, the description of bias used here is slightly different to the usual definition (the expectation of difference between parameter estimates) [24].

$$PB= (b_m\text{-}b_f/b_f)*100 \qquad (3.2)$$

where b_f is the regression coefficient estimated from the models fitted to the complete cases, and b_m is the regression coefficient estimated from the other models (i.e. missing data models). The BS, on the other hand, is an overall measure of predictive performance, i.e. a combination of discrimination and calibration [25] *(2.3)*. The *BS*, or average prediction error is defined as follows:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \quad \text{t=1,…,N} \qquad (3.3)$$

in which f_t is predicted probabilities by the model, O_t is the observed outcome (0 or 1), and N is the total number of observations. A *BS* value close to 0 indicates the model performs well, whereas larger scores show poorly fitting models [26].

Additionally, we evaluated our models using the Area Under the Receiver Operating Characteristic (AUC) curve. The AUC is often used to summarize and compare the discriminatory accuracy of a diagnostic test or modality, and to evaluate the predictive power of statistical models for binary outcomes [27]. We used the AUC to evaluate how accurate our logistic regression models were in predicting overseas travel. Accordingly, we selected the FBS approach as a method of choice to apply to the original dataset.

Prediction of Overseas Travel

A fully Bayesian logistic regression model was fitted to the original dataset to predict missing overseas travel status of notified campylobacteriosis cases. The priors for all parameters in the model were specified as uninformative as before. We ran the model for 30,000 iterations, used 2 chains and 3000 iteration burn-ins. Finally, we investigated model fit by examining density plots, autocorrelation and trace plots of a subset of parameters in the model for a visual graphical assessment. After a reasonable convergence was achieved (i.e., smooth density plots and no autocorrelation), we extracted the individual predicted summary measures of probability of overseas travel for the cases (mean, median and standard deviation) from the posterior distribution. These values describe the distribution of the probability of overseas travel for individual cases. Since the prior distribution and the likelihood (data) are normal, the posterior distribution is approximately normal [28]. Therefore the means and standard deviations (SD) of individual cases were used to simulate a normal distribution from which an estimate and 95% CI of the predicted proportion of travel associated cases for each DHB were computed as follows *(3.4)*:

$$X_j = N(n=10000, \text{mean}=m_i, \text{sd}=sd_i),\ i=1,2,..,I \text{ and } j=1,2,..,J \qquad (3.4)$$

where X_j is a normal distribution of a vector of means and SDs of individual cases in DHB$_j$, m_i denotes the mean value for each case, sd_i is the SD of the mean value for each case, indexes *I* and *J* stand for the number of individual cases in a DHB and the number of DHBs, respectively. Our model's prediction of the status of overseas travel was compared to the observed proportion of campylobacteriosis notifications.

3.3 Results

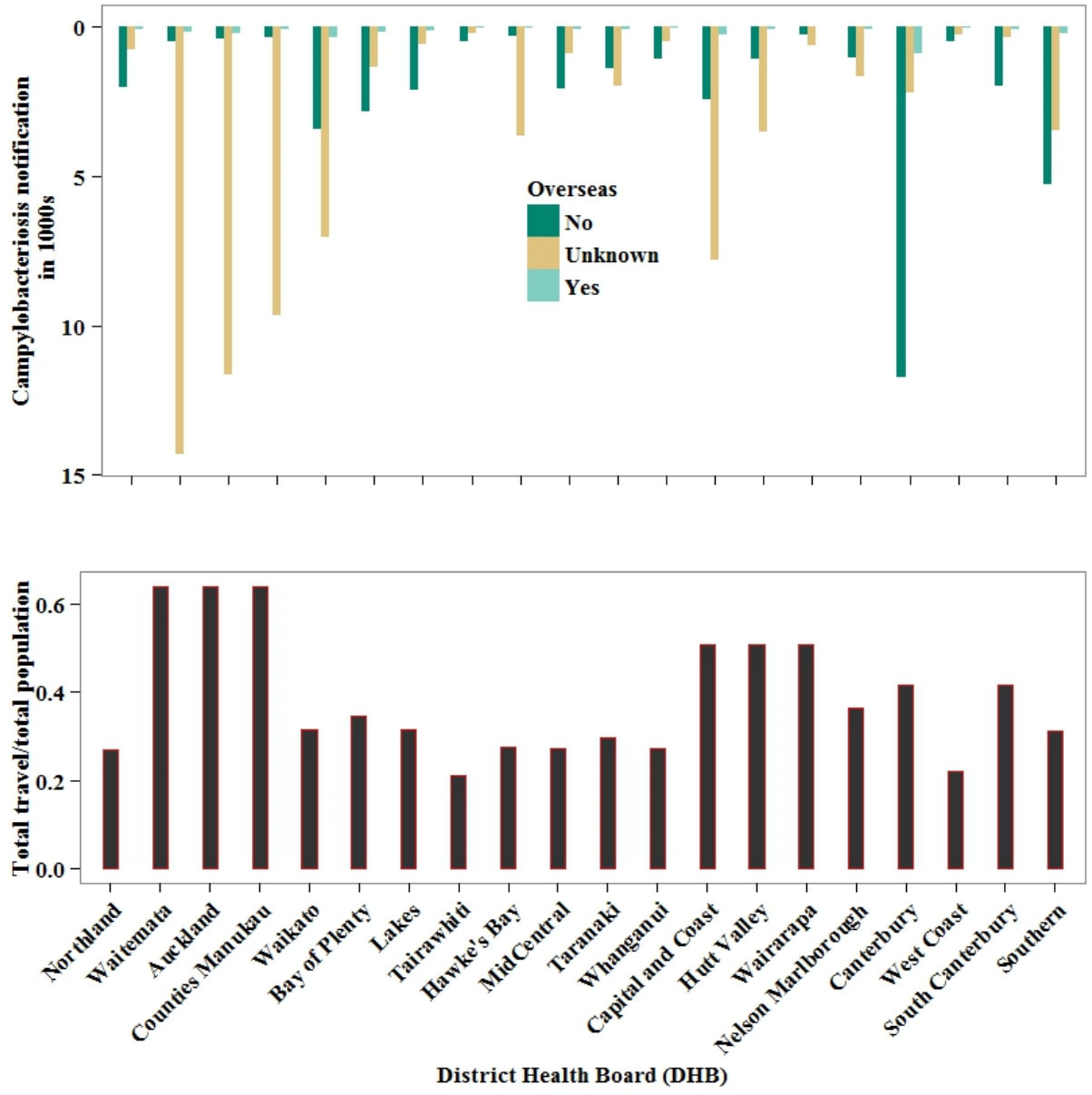
There were 121,764 notifications of campylobacteriosis in New Zealand reported between 2000 and 2010. Of these most were culture confirmed (Confirmed) or epidemiologically linked to confirmed cases or outbreak sources (Probable) (Table 3.2). Deducting those for which the notification status was "Unknown" or "Under Investigation" gave 119,375 cases for the primary dataset. Table 3.2 shows the proportion of all campylobacteriosis cases notified between 2000 and 2010 with and without travel information. Among 119,375 cases, 44,285 (37.1%) had complete information in the travel section of the *EpiSurv* questionnaire, and 3107 (7%) of cases who had filled this section had made short term international travel. As there are no definitive results for the cases with a case status of 'under investigation' and 'Unknown', we excluded them from the analysis. Furthermore, 0.6% of *Age* and 1.6% of *Sex* variables were missing in the primary dataset and the associated records were excluded, which makes the total number of cases available for analysis as 116,721.

Table 3.2. The total number of campylobacteriosis notification in New Zealand residents categorized by information on overseas travel (2000-2010).

Travel status	***Campylobacter* status**				
	Confirmed	Probable	Under Investigation	Unknown	Total
No	41617	60	52	416	42145
Unknown	74481	110	222	1653	76466
Yes	3100	7	7	39	3153
Total	119198	177	281	2108	121764

Figure 3.1 displays the total number of notified *Campylobacter* cases between 2000 and 2010 which are categorized by the status of overseas travel reporting and annual rate of overseas travels per person in each DHB. Most of the cases reported from Auckland, Waitemata, and Counties Manukau DHBs lack travel information. However, the majority of reported cases and more than 55% of all travel between 2000 and 2010 originated from residents in these DHBs [10]. As shown in Figure 3.1, more than 60% of all cases come from six DHBs, namely Waitemata (12.8%), Canterbury (12.7%), Auckland (10.6%), Waikato (9.3%), Capital and Coast (8.9%), and Counties Manukau (8.7%) (Fig 3.1).

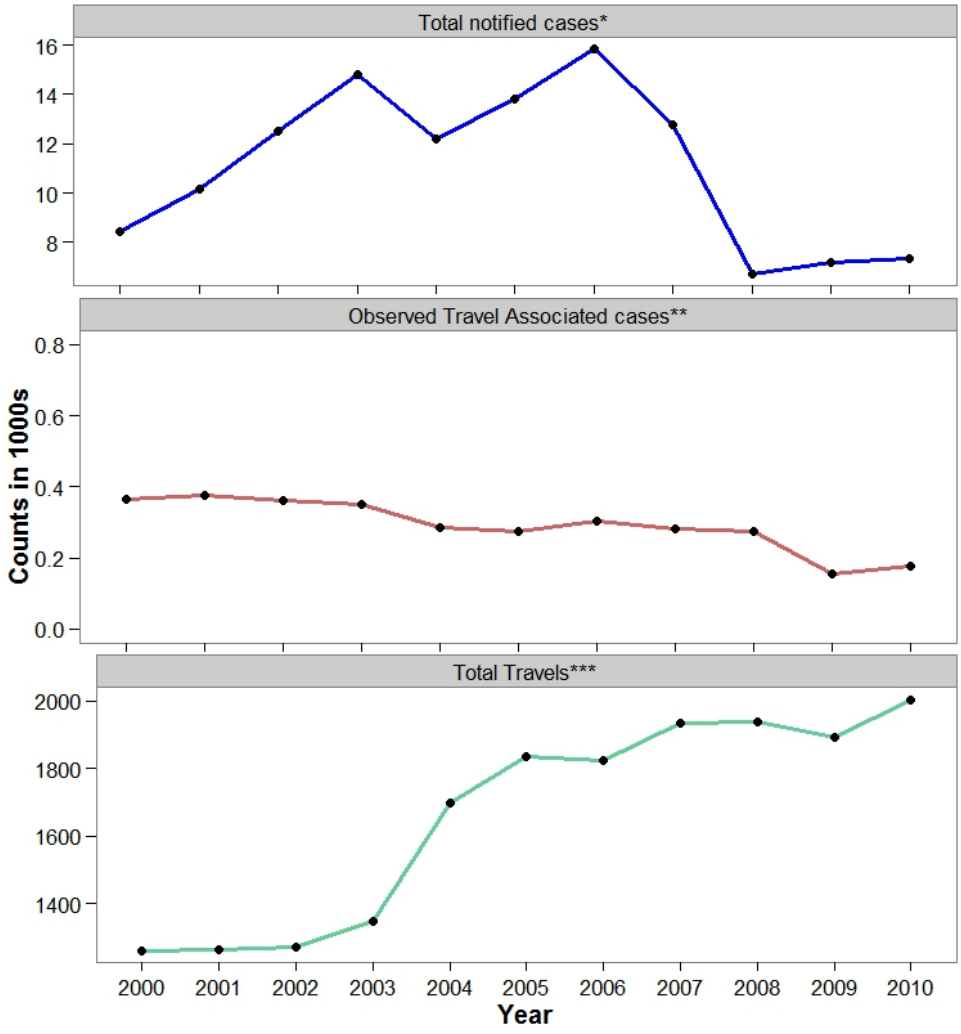
Figure 3.1. Distribution of campylobacteriosis notification categorized by the status of overseas travel (upper panel) and the annual proportion of short term international travels (lower panel), in DHBs of New Zealand (2000 – 2010).



The number of short term international trips by New Zealanders consistently increased between 2000 and 2010 (bottom panel in Figure 3.2). As evident from Figure 3.2, total campylobacteriosis notification in New Zealand had been increasing until 2006 except a slight decrease in 2003-2004.

After 2006, the total number of notifications declined considerably. The middle panel of Figure 3.2 indicates that the total number of reported travel associated cases has slightly declined over time.

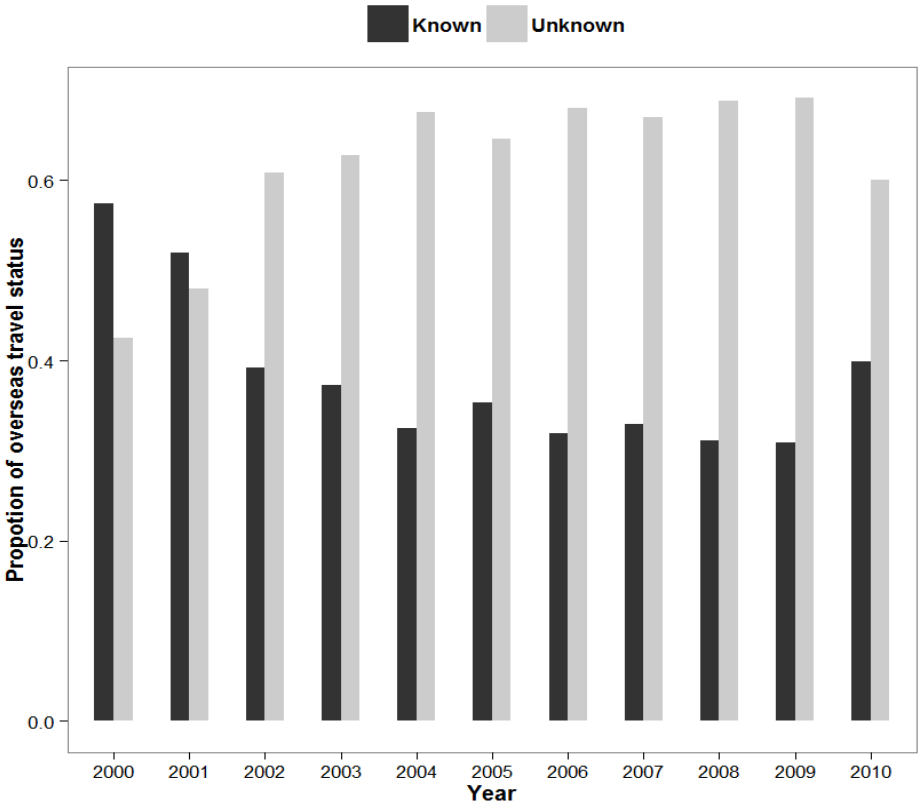
Figure 3.2. Annual short term international travel and campylobacteriosis notification of New Zealand residents (2000-2010).



*Total notified cases: total number of campylobacteriosis cases notified between 2000 and 2010
**Observed travel associated cases: campylobacteriosis cases that had confirmed overseas travel during the incubation period of the disease
***Total travels: total number of short term international travels between 2000 and 2010. Short term international travel is defined as international departures of New Zealand residents for an intended period of less than 12 months (*Statistics New Zealand* [http://www.stats.govt.nz/]).

The overall trend of the availability of information on travel status for the notified campylobacteriosis cases has been consistently decreasing throughout the study period (Fig 3.3).

Figure 3.3. The proportion of campylobacteriosis notifications in New Zealand with known and unknown status of overseas travel information (2000-2010).



The MAR assumption was examined by statistically testing the association between the missing value indicator and predictors of overseas travel in the dataset. Table 3.3 shows that majority of the predictors are strongly associated with missingness in overseas travel. This strong association

implies that missingness in the data can at least partially be explained by the fully observed variables in the model, which supports the MAR assumption (Table 3.3).

Table 3.3. A Logistic regression of missing indicator (1=missing overseas travel information, 0=observed overseas travel information) on predictors of overseas travel status to test the validity of Missing At Random assumption.

Coefficients	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-8.757	0.089	<0.001
Urban*	2.992	0.103	<0.001
DepIndex**	0.525	0.006	<0.001
Travel Rate***	0.081	0.001	<0.001
Age(5-19)	0.154	0.027	<0.001
Age(20-59)	0.033	0.023	0.145
Age(60+)	-0.142	0.027	<0.001
Summer	0.014	0.018	0.443
Autumn	-0.002	0.021	0.94
Winter	0.035	0.021	0.085
Male	0.153	0.014	<0.001
Intervention****	0.345	0.016	<0.001

Keys: * proportion of DHB population under urban influence; ** Deprivation index (scale 0-10, 0 being least deprived and 10 being most deprived DHB; *** Short term international travel per 100 residents of a DHB; ****a binary indicator variable to identify pre and post 2006 intervention. Age (<5), Spring, and Female sex are reference categories.

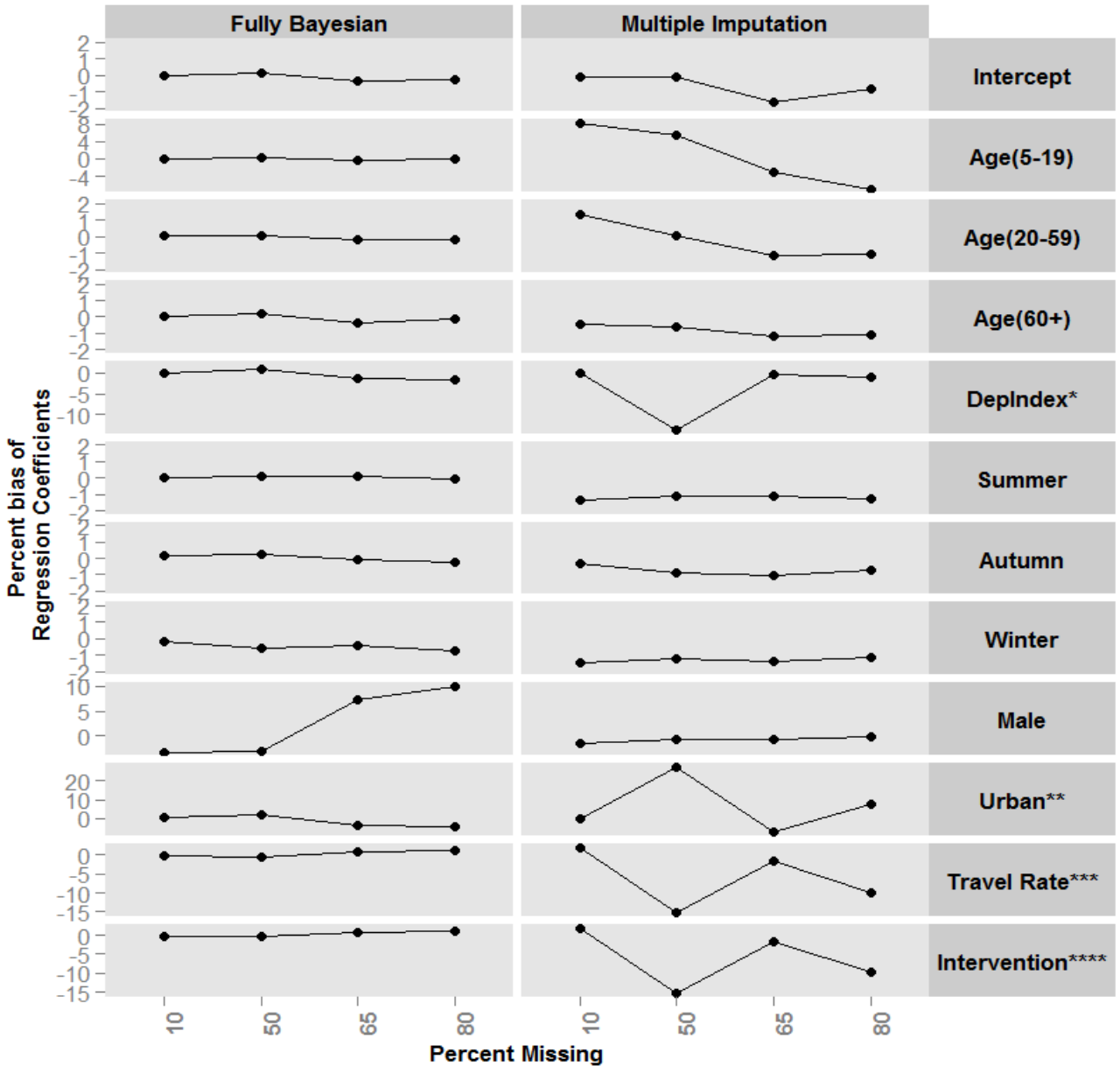
The outcomes of applying MI and FSB models to the datasets with artificially induced missingness is given in Table 3.4 and Figure 3.4. Comparison of BS and AUC to select the best predictive model shows that the FBS model is more robust than MI as the rate of missingness increases (Table 3.4). At 10% MAR, there was no difference between MI and FBS. However at 50%, 65% and 80% MAR cases, the FBS approach resulted in relatively higher AUC and smaller BS than MI (Table 3.4). Furthermore, PB of regression coefficients were consistently low across all categories of missingness for most of the variables in the FBS models as compared to MI (Figure 3.4). This evidence suggests the FBS performs relatively better for a dataset with a high rate of missing values.

Table 3.4. Comparison of Brier Score and Area Under the Curve (AUC) between Fully Bayesian and Multiple imputation models for the prediction of overseas travel status of campylobacteriosis cases.

Accuracy measure	Complete data[1]		Missing data							
	Frequentist	Bayesian	Multiple iputation				Fully Bayesian			
			10pct[3]	50pct	65pct	80pct	10pct	50pct	65pct	80pct
Brier Score[2]	0.062	0.062	0.067	0.24	0.18	0.19	0.062	0.063	0.062	0.063
AUC	0.67	0.67	0.64	0.49	0.42	0.49	0.67	0.67	0.65	0.64

[1] n=44,285
[2]AUC=Area Under the Receiver Operating Characteristic Curve
[3] pct = percent

Figure 3.4. The comparison of Fully Bayesian and Multiple Imputation models regarding Percent Bias of regression coefficients for different proportion of missing overseas travel status of campylobacteriosis cases.

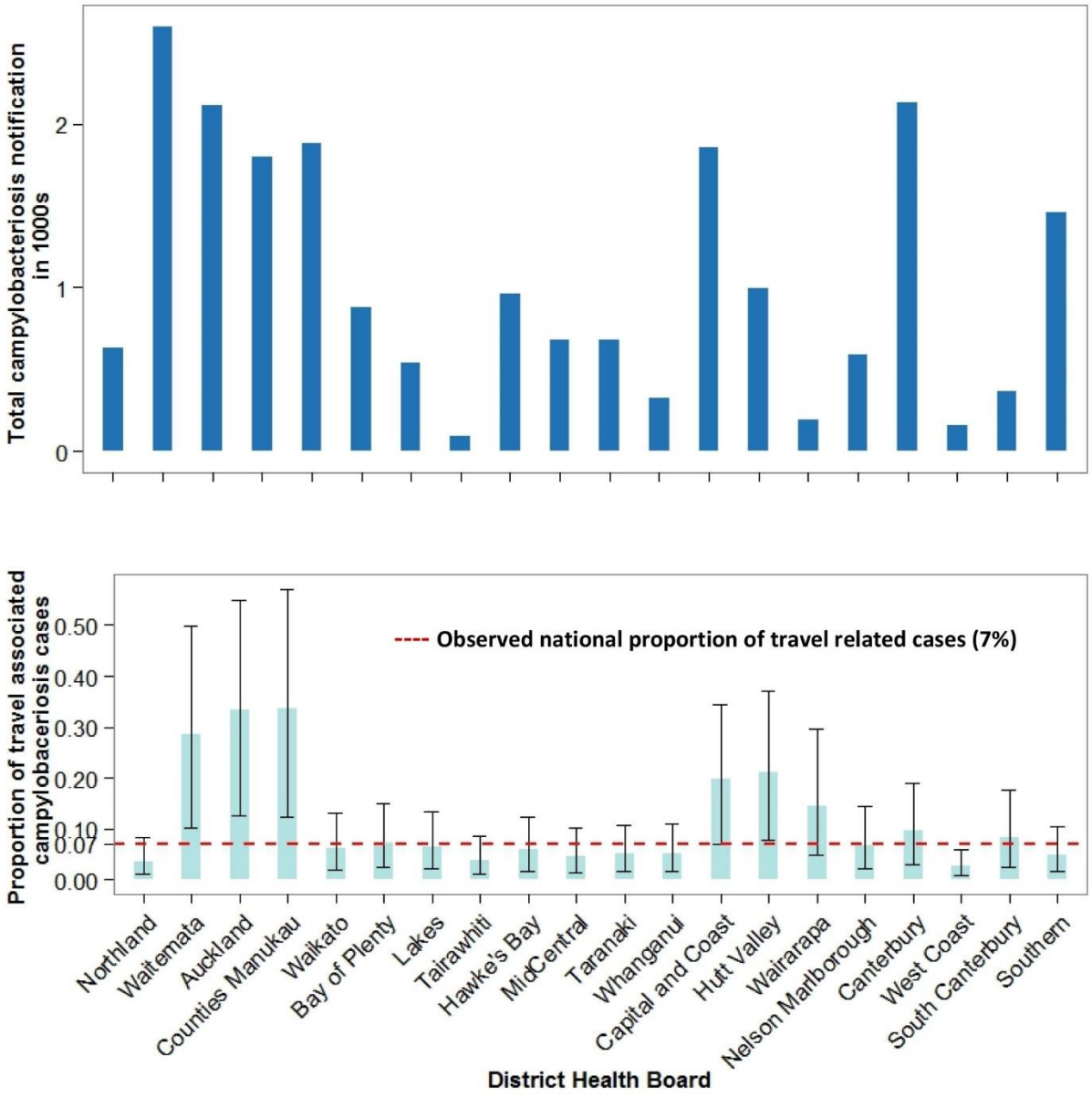


Keys: * Deprivation index (scale 1-10, 1 being least deprived and 10 being most deprived DHB; **proportion of DHB population under urban influence;*** Short term international travel per 100 residents of a DHB; ****a binary indicator variable to identify cases reported before or after 2006 poultry intervention period.

The FBS model was chosen to be applied to the original dataset to estimate the proportion of cases due to overseas travel in each DHB during the period 2008 to 2010. During this period the number of campylobacteriosis notifications and travel rates were relatively stable. Figure 3.5 shows the total number of notified campylobacteriosis cases (upper panel) and the estimated proportion of travel related cases as predicted by our model (lower panel). The horizontal dashed line in the bottom panel is drawn to indicate the percent of reported travel associated cases (7%) among all cases that have provided travel information.

In many of the DHBs with a high rate of campylobacteriosis notification (see upper panel of Figure 3.5) and high rate of travel (see bottom panel of Figure 3.1), such as Auckland, Counties Manukau and Waitemata, our model predicted a high proportion of campylobacteriosis cases to be associated with overseas travel. For example, the proportion of travel associated cases was higher in Counties Manukau, Auckland and Waitemata DHBs, at 0.34 (0.12, 0.57), 0.33 (0.13, 0.55) and 0.28 (0.10, 0.49), whereas the lowest proportions were estimated for West Coast, Northland and Tairawhiti at 0.02 (0.01, 0.06), 0.03 (0.01, 0.08) and 0.04 (0.01, 0.08) respectively. Except for Auckland, Counties Manukau, West Coast and Waitemata DHBs, the 95% CI of the predicted proportion of travel associated cases included the observed national proportion of travel related cases (horizontal dashed line in bottom panel of Figure 3.5). Accordingly, the national estimate and 95% CI of the proportion of travel related cases based on our model is 0.16 (0.02, 0.48).

Figure 3.5. The total number of campylobacteriosis notification (upper panel) and the proportion of travel related cases (lower panel) for each DHB of New Zealand (2008-2010).



Notes: (1) Bottom panel: proportion of travel related cases predicted by the Bayesian model. The error bars are 95% CIs of the proportion of overseas travel. (2) The dashed horizontal line is the proportion of travel related campylobacteriosis cases for which travel history is available nationally (7%).

3.4 Discussion

Data gaps in notification data have been a continuous public health challenge for identifying the source of infection and preventing infectious diseases, including campylobacteriosis. The increase of overseas travel by New Zealanders and the established risk of overseas travel for *Campylobacter* infection emphasize the need to study travel associated illnesses.

A total of 18.3 million short term international trips by New Zealand residents were recorded between 2000 and 2010. Most travel was to the Pacific region, East Asia and North America, while the least travel was recorded for the regions of West and Central Africa and Antarctica. This is in agreement with previous reports that New Zealanders travel to more than 150 countries, of which countries in the Pacific region and North America are the most popular destinations [6]. International travel has been increasing since 2004 (see Figure 3.2).

In contrast, a substantial reduction of incidence of notified campylobacteriosis cases occurred after 2006 (Figure 3.2). The significant changes in notifications post 2006 were believed to be the result of interventions targeting poultry [2]. Despite this overall decline in notifications of campylobacteriosis in New Zealand, the change attributable to cases associated with overseas travel is not well understood. Although the outbound travel rate of New Zealand residents has been increasing, we noticed a slight decline in notified travel associated cases throughout the study period (middle panel of Fig 3.2). This could be due to the decrease in reporting of travel status for the cases throughout the study period (Figure 3.3) that may have confounded conclusions on the origin of the disease.

In addition, there is a consistently low reporting rate of detailed travel information in urban areas of New Zealand such as in Auckland and Wellington regions. A case control study in the New Zealand regions with high notification rates, including Auckland region, suggests that recent overseas travel was a significant risk factor for the occurrence of campylobacteriosis in this region [29]

The majority (62%) of campylobacteriosis case reports in New Zealand lack travel history during the incubation period prior to disease. The level of completeness of travel history for notified cases has been a challenging task as is reported by some other studies [30–32]. It is therefore necessary to estimate travel associated cases based on imperfect data.

Among the total number of notifications with known travel history in the eleven years span of our study, only 3107 (7% of notifications with known travel status) had reported overseas travel. As New Zealanders are prolific travelers, this proportion of cases may underestimate the true contribution of travel as a risk factor for campylobacteriosis in New Zealand. For this reason, model-based methods such as MI and FBS can be useful to fill the data gaps, using covariates that predict overseas travel. The FBS model resulted in an estimate of the national proportion of notifications due to overseas travel of 16%, a higher value compared to the 7% estimate using only known values. Higher rates of travel related campylobacteriosis have been reported in other developed countries such as in Canada (21.6%) [33], England (17%) [30], USA (18%) [34], Denmark (18%) [35] and Switzerland (46.1%) [36].

Both MI and FBS have become popular for data augmentation in recent years due to their generic application and availability of large variety of computational tools [7, 14, 20]. Several other

methods have been described in the literature for filling data gaps, such as maximum likelihood estimation (ML) and Weighted Estimating Equations(WEE) [37, 38]. However, we focused on MI and FBS, because these are regarded as the most applicable and widely used methods [7, 21, 37, 39]. When the proportion missing in the data is small, and MAR assumption holds, both methods resulted in similar predictions. This is supported by previous studies indicating that MI and FBS are asymptotically equivalent whenever the data is MAR [40]. At a higher rate of missingness, though, the FBS was more robust than MI. This is indicated in Table 3.4 and 3.5, that the Brier Score, AUC and PB of regression coefficients remained stable for the FBS case. We therefore chose the FBS model for predicting overseas travel status of the cases using demographic and other predictors.

Our model predicted a high proportion of travel associated cases in major urban areas of New Zealand, such as in Auckland, Counties Manukau and Waitemata DHBs. This could be due to high rates of travel of their residents to the Pacific Islands and South East Asia regions, which is partially driven by the comparatively high proportion of Asian ethnicity (23.8%) and Pacific Peoples (14.6%) in the Auckland region [6, 10]. It has been previously established that individuals traveling to these world regions are at a higher risk of travel associated illnesses, including campylobacteriosis [41]. On the other hand, the DHBs with a smaller proportion of model-predicted travel related cases (e.g., Northland, West Coast and Tairawhiti) are those with a lower outbound travel rate.

The use of FBS and MI methods provides a methodology to calculate uncertainty bounds around the estimates of travel associated cases. The degree of uncertainty of the predicted proportion of

travel associated cases can be attributed to variation in the risk of travel associated illnesses among individuals that have different covariate values. Such variation in the risk of campylobacteriosis with respect to age, sex and season is in agreement with previous reports in literature [5, 42].

If the MAR assumption holds, which is usually difficult to achieve, our Bayesian prediction model provides a plausible way for predicting missing overseas travel of campylobacteriosis cases. [21]. It is also important to note that any other missing data analysis approaches require assumptions that are just as difficult to justify [43]. However, the FBS procedure should not be viewed as the 'gold standard' for filling data gaps for every situation, although it offers a flexible approach for data augmentation. Priors can be enhanced if data regarding association of risk factor–outcome become available. In addition, the Bayesian model specification can be modified if the MAR assumption is thought to be violated.

Better notification reporting, particularly for areas with high outbound travel and high notification of cases such as in highly urban areas can improve our understanding of the epidemiology of travel associated campylobacteriosis in New Zealand. However, reporting completeness is limited by the resources available in each DHB. Use of alternative data collection approaches such as web based applications, cross tabulation of Customs data with *EpiSurv* data, and creating awareness in the population regarding the importance of the information for the public health databases may improve reporting completeness. Although the emphasis in this report is on predicting travel information of *Campylobacter* cases in New Zealand, the method can be implemented for other diseases of public health significance which have similar data gaps.

3.5 Conclusions

The apparent challenge of data gaps regarding risk factors for campylobacteriosis suggests the use of model-based approaches for estimating missing values. Filling data gaps is particularly important for regions with a high rate of incomplete data. The fully Bayesian modelling approach offers a flexible alternative for data augmentation particularly when the missing rate is very high. Due to the strong MAR assumption necessary for the prediction, the FBS, on the other hand, may not be the best approach for predicting missing values in imperfect datasets.

Competing interest

The authors declare that they have no any competing interests

Acknowledgments

This work was supported by the NIH Ruth L. Kirschstein National Research Service Award Institutional Training Grant T32 RR023916 and T32 OD010423. Parts of this material are based on data and information provided by the Institute of Environmental Science & Research Ltd on behalf of the Ministry of Health. However, the analyses, conclusions, opinions and statements expressed herein are those of the authors, and not necessarily those of the Institute of Environmental Science & Research Ltd or the Ministry of Health.

3.6 References

1. Baker MG, Sneyd E, Wilson NA: Is the Major Increase in Notified Campylobacteriosis in New Zealand Real? *Epidemiol Infect* 2007, 135:163–170.

2. Sears A, Baker MG, Wilson N, Marshall J, Muellner P, Campbell DM, Lake RJ, French NP: Marked Campylobacteriosis Decline after Interventions Aimed at Poultry, New Zealand. *Emerg Infect Dis* 2011, 17:1007–1015.

3. Mullner P, Shadbolt T, Collins-Emerson JM, Midwinter AC, Spencer SEF, Marshall J, Carter PE, Campbell DM, Wilson DJ, Hathaway S, Pirie R, French NP: Molecular and spatial epidemiology of human campylobacteriosis: source association and genotype-related risk factors. *Epidemiol Infect* 2010, 138:1372–1383.

4. Horn BJ, Lake RJ: Incubation period for campylobacteriosis and its importance in the estimation of incidence related to travel. *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull* 2013, 18.

5. Ekdahl K, Andersson Y: Regional risks and seasonality in travel-associated campylobacteriosis. *BMC Infect Dis* 2004, 4:54.

6. Outbound Travel by New Zealand Residents. New Zealand | SERIES D1 | September 2009. The Ministry of Tourism. [http://www.med.govt.nz/about-us/pdf-library/tourism-publications/]

7. van Buuren S: *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press; 2012.

8. Gelman A, Hill J: *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press; 2006.

9. EpiSurv: Public Health Surveillance, Information for New Zealand Public Health Action [https://surv.esr.cri.nz/episurv/index.php]

10. Home - Statistics New Zealand [http://www.stats.govt.nz/]

11. Salmond C, Crampton P, Atkinson J: NZDep2006: New Zealand Index of Deprivation, 2007. 2007.

12. Rubin DB: Inference and missing data. *Biometrika* 1976, 63:581–592.

13. Schafer JL: Multiple imputation: a primer. *Stat Methods Med Res* 1999, 8:3–15.

14. Harel O, Zhou X-H: Multiple imputation: review of theory, implementation and software. *Stat Med* 2007, 26:3057–3077.

15. Horton NJ, Lipsitz SR: Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am Stat* 2001, 55:244–254.

16. Liu M, Taylor JM, Belin TR: Multiple imputation and posterior simulation for multivariate missing data in longitudinal studies. *Biometrics* 2000, 56:1157–1163.

17. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM: Review: A gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006, 59:1087–1091.

18. Van Buuren S, Boshuizen HC, Knook DL, others: Multiple imputation of missing blood pressure covariates in survival analysis. *Stat Med* 1999, 18:681–694.

19. Rubin DB: *Multiple Imputation for Nonresponse in Surveys*. Wiley; 1987.

20. Horton NJ, Kleinman KP: Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Stat* 2007, 61:79–90.

21. Mason A, Best N, Richardson S, PLEWIS I: Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. 2010.

22. Plummer M, others: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. *Volume 124*. Technische Universit at Wien; 2003:125.

23. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB: *Bayesian Data Analysis, Third Edition*. 3 edition. Boca Raton: Chapman and Hall/CRC; 2013.

24. Alexina Mason NB: Insights into the use of Bayesian models for informative missing data. International Biometric Conference, Dublin, 13-18 July 2008

25. Brier GW: Verification of Forecasts Expressed in terms of Probability. *Mon Weather Rev* 1950, 78:1–3.

26. Steyerberg E: *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer Science & Business Media; 2008.

27. Cleves MA: Comparative assessment of three common algorithms for estimating the variance of the area under the nonparametric receiver operating characteristic curve. *Stata J* 2002, 3:280–9.

28. Pratt JW, Raiffa H, Schlaifer R: *Introduction to Statistical Decision Theory*. MIT Press; 1995.

29. Eberhart-Phillips J, Walker N, Garrett N, Bell D, Sinclair D, Rainger W, Bates M: Campylobacteriosis in New Zealand: results of a case-control study. *J Epidemiol Community Health* 1997, 51:686–691.

30. Zenner D, Gillespie I: Travel-associated salmonella and campylobacter gastroenteritis in England: estimation of under-ascertainment through national laboratory surveillance. *J Travel Med* 2011, 18:414–417.

31. Guzman-Herrador B, Vold L, Nygard K: Surveillance of travel-associated gastrointestinal infections in Norway, 2009-2010: are they all actually imported? *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull* 2012, 17:20294.

32. Lake R, Horn B, Ball A, New Zealand, Ministry of Agriculture and Forestry, MAF Biosecurity New Zealand, New Zealand Food Safety Authority, New Zealand, Ministry for the Environment: *Campylobacter in Food and the Environment, Examining the Link with Public Health Pathway Attribution*. Wellington [N.Z.]: Ministry of Agriculture and Forestry; 2011.

33. Ravel A, Nesbitt A, Marshall B, Sittler N, Pollari F: Description and burden of travel-related cases caused by enteropathogens reported in a Canadian community. *J Travel Med* 2011, 18:8–19.

34. Ricotta EE, Palmer A, Wymore K, Clogher P, Oosmanally N, Robinson T, Lathrop S, Karr J, Hatch J, Dunn J, Ryan P, Blythe D: Epidemiology and antimicrobial resistance of international travel-associated Campylobacter infections in the United States, 2005-2011. *Am J Public Health* 2014, 104:e108–114.

35. Neimann J, Engberg J, Mølbak K, Wegener HC: A case-control study of risk factors for sporadic campylobacter infections in Denmark. *Epidemiol Infect* 2003, 130:353–366.

36. Schorr D, Schmid H, Rieder HL, Baumgartner A, Vorkauf H, Burnens A: Risk factors for Campylobacter enteritis in Switzerland. *Zentralblatt Für Hyg Umweltmed Int J Hyg Environ Med* 1994, 196:327–337.

37. Joseph G Ibrahim M-HC: Missing-Data Methods for Generalized Linear Models. *J Am Stat Assoc - J Amer Stat Assn* 2005, 100:332–346.

38. Lipsitz SR, Ibrahim JG, Zhao LP: A Weighted Estimating Equation for Missing Covariate Data with Properties Similar to Maximum Likelihood. *J Am Stat Assoc* 1999, 94:1147–1160.

39. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009, 338.

40. Chen Q, Ibrahim JG: A note on the relationships between multiple imputation, maximum likelihood and fully Bayesian methods for missing responses in linear regression models. *Stat Interface* 2014, 6:315–324.

41. Dupont HL: Systematic review: prevention of travellers' diarrhoea: Systematic Review: Prevention of Travelers' Diarrhoea. *Aliment Pharmacol Ther* 2008, 27:741–751.

42. Unicomb LE, Dalton CB, Gilbert GL, Becker NG, Patel MS: Age-specific risk factors for sporadic Campylobacter infection in regional Australia. *Foodborne Pathog Dis* 2008, 5:79–85.

43. Enders CK: *Applied Missing Data Analysis*. New York: Guilford Press; 2010.

Chapter 4: Bayesian spatio-temporal analysis of travel associated campylobacteriosis in New Zealand

Amene, E.[1], Horn, B.[2], Pirie, R.[2] Lake, R.[2] and D., Döpfer [1]

[1]University of Wisconsin-Madison, School of Veterinary Medicine, Department of Medical Sciences, USA,

[2]Environemental Science and Research, Christchurch, New Zealand

Abstract

Background: campylobacteriosis is a notifiable disease in New Zealand. Overseas travel is one of the major risk factors for acquiring this disease. A better understanding of the spatial and temporal distribution of travel associated campylobacteriosis is crucial for control and prevention efforts to be effective.

Methods: Bayesian predictive inference was applied by implementation of the *INLA* package in R, which makes usage of Integrated Laplace Approximation for prediction of risk. A Poisson lognormal regression model was fitted to the data with both space structured and unstructured random effects. For model selection and Goodness of Fit evaluation, the Deviance Information Criteria was used. The selected model resulted in travel associated campylobacteriosis risk maps of New Zealand and the main travel destinations.

Results: Based on a combination of Deviance Information Criteria, pD, Conditional Predictive Ordinate and parsimony, a model consisting of both spatially structured and unstructured random effects, and an unstructured temporal random effect fitted the data best. A map depicting the spatial and temporal distribution of travel related campylobacteriosis was developed. Based on this map, major urban areas have consistently higher risk of travel related cases than other regions of New Zealand throughout the study period. The risk map constructed using observed travel associated

cases with known travel destination showed similar pattern of risk as compared to previously established campylobacteriosis risk zones of main global destinations.

Conclusion: Mapping local area specific estimates of campylobacteriosis risk informs strategic allocation of public health resources. Spatial and temporal description of disease risks in different areas subject to surveillance strengthens the surveillance of the disease. The result provides a better understanding of the spatial risk and the variability of risk due to spatial and non-spatial components. Further investigation of areas with high risk of travel related campylobacteriosis in New Zealand yields clues toward the understanding of disease causation.

Key words: Campylobacteriosis, New Zealand, Bayesian, travel

4.1 Background

Campylobacter infections are the most commonly reported bacterial causes of food borne gastrointestinal illnesses, accounting for more than 10% of all diarrheal cases in humans worldwide [1]. Particularly in developed countries, infection with *Campylobacter* species is the leading cause of human bacterial gastroenteritis [2–5].

Although there has been a significant reduction in the incidence of campylobacteriosis in New Zealand after 2006, the infection rate is fairly high compared to other developed countries [6]. For example, the notified rate of campylobacteriosis in New Zealand in 2011 was 151.9 per 100,000 population compared to the USA (13.02) and the EU (45.2) [7, 8]. A number of risk factors for acquiring campylobacteriosis have been reported in literature. The most common ones are consumption and handling of poultry, unpasteurized milk and dairy products; consumption of

untreated water; contact with domestic animals; and, international travel [6–9]. Poultry and poultry products have been implicated as the source of infection in up to 80% of the cases and as the risk factor associated with more than 41% of transmissions to humans [5, 6]. On the other hand, the rates of infection and epidemiology of *Campylobacter* infections associated with other risk factors such as international travel remain unclear [13]. With increasing travel and trade opportunities across the globe, there has been a greater risk of acquiring campylobacteriosis across borders. A better understanding of the spatial and temporal distribution of travel associated campylobacteriosis is crucial for improved control and prevention efforts. Meanwhile, no information exists regarding model-based travel associated campylobacteriosis risk in New Zealand that takes into account the spatial and temporal dependence of the infection risk. Maps of spatial predictions and corresponding uncertainties in model outputs can allow informed decision-making with regard to targeted disease control [14, 15]. In this study, we used a Bayesian disease mapping (BDM) framework to develop a model for predicting spatial and temporal distribution of risk of travel associated campylobacteriosis [16, 17]. The aim of this analysis is to classify District Health Boards (DHBs) of New Zealand based on distribution of the risk of travel associated campylobacteriosis.

4.2 Methods

Data

Campylobacteriosis notification and short term international travel data were obtained from New Zealand national databases (*EpiSurv* and *Statistics New Zealand*). The *EpiSurv* (Notifiable Disease Surveillance Database) (https://surv.esr.cri.nz/episurv/index.php) records information about national and local notifiable cases and their associated demographic structures. *Statistics New*

Zealand (http://www.stats.govt.nz/), on the other hand, is the public service department of New Zealand responsible for collecting and producing data including census counts and migration patterns for New Zealand residents (such as departure regions and main travel destinations). In this database, short-term travel is defined as international departures of New Zealand residents for an intended duration of stay outside of New Zealand of less than 12 months. The original notification dataset, however, lacks travel information for a significant proportion of the cases. In our previous study, we used a covariate-driven model-based approach to fill-in notification data-gaps including overseas travel status for the campylobacteriosis cases (Chapter 3). We used the imputed dataset from our previous study for the BDM approach toward the analyses of travel associated cases. The summary of the dataset used for the current analysis is shown in Table 4.1 below.

Table 4.1. The short term international travels and the total campylobacteriosis notifications of New Zealand residents (2000-2010).

District Health Board	Total Travels	Total Notifications[1]	Proportion of travel associated cases[2] Mean(95% CI)
Northland	410364	2878	0.04(0.01,0.07)
Waitemata	3138067	14963	0.18(0.07,0.28)
Auckland	2679979	12273	0.22(0.09,0.33)
Counties Manukau	2811854	10095	0.23(0.1,0.36)
Waikato	1104560	10859	0.05(0.02,0.1)
Bay of Plenty	687786	4342	0.06(0.02,0.11)
Lakes	377901	2789	0.06(0.02,0.1)

Tairawhiti	98764	726	0.04(0.01,0.07)
Hawke's Bay	423987	3970	0.05(0.02,0.09)
MidCentral	451491	3060	0.04(0.01,0.08)
Taranaki	324005	3453	0.05(0.02,0.08)
Whanganui	240727	1578	0.05(0.02,0.09)
Capital and Coast	1410679	10482	0.13(0.05,0.2)
Hutt Valley	549565	4696	0.15(0.06,0.23)
Wairarapa	204285	879	0.11(0.04,0.19)
Nelson Marlborough	497308	2753	0.06(0.02,0.1)
Canterbury	2054498	14874	0.07(0.03,0.12)
West Coast	72688	760	0.03(0.01,0.06)
South Canterbury	208955	2340	0.07(0.02,0.11)
Southern	946026	8951	0.04(0.02,0.08)

Notes: New Zealand has 20 District Health Boards (DHBs). The DHBs are organizations established by the New Zealand Public Health and Disability Act 2000, which are responsible for ensuring the provision of health and disability services to populations within a defined geographical area.

[1] Total campylobacteriosis notifications are either culture confirmed or epidemiologically linked (probable) cases. [2] Travel associated cases are estimated from a Bayesian hierarchical model implementing predictive covariates (Deprivation Index, proportion of the population under urban influence, DHB's Travel rate, age of the case, season of travel and whether the case was reported before or after 2006 intervention period). The mean value was used to compute the estimated travel associated cases form the total notification.

Statistical Analysis

We performed a descriptive analysis regarding travel departures and total campylobacteriosis notifications in the DHBs and the geographic distribution of main travel destinations. For this

analysis, we grouped the main countries of destination into twenty 'World Regions' based on geographic proximity [13].

The distribution of *Campylobacter* cases are assumed to follow a *Poisson* process, as the disease usually is a sporadic rare event in a large population, mean and variance are equal, and this distribution gives a good approximation to the true underlying process [18]. A Bayesian spatio-temporal model with spatially structured and unstructured random effects was fitted to the data. Incorporating random effects into a regression model absorbs extra-variation in the data that could be attributable to unmeasured and clustered variables [19]. The Besag, York and Mollié model (BYM) was chosen to fit the data for the reason that it easily accommodates area specific random effects and covariate adjustments *(4.1)* [20]. This model corrects for spatial correlations between infection risks in nearby areas. This method utilizes the spatial structure of the data to "borrow" more information from neighboring areas compared to from those areas located far away and it adjusts local infection risks from local, neighboring values. A number of Bayesian hierarchical models have been proposed that extend the BYM model for analysis of spatially referenced data [21–23]. For the DHB level analysis, we adopted the model formulated to incorporate both spatial and temporal random effects [17]. The generalized form of the spatio-temporal Bayesian regression model has the following structure:

$$Y_{ij} \sim Poisson\ (\theta_{ij} E_{ij}), \qquad (4.1)$$

where $\boldsymbol{Y_{ij}}$ and $\boldsymbol{E_{ij}}$ are observed and expected number of cases, respectively, for each area *i* and time unit *j*; $\boldsymbol{\theta_{ij}}$ denotes the relative risk in area $i = 1,...n$ and time $j=1,...J$. The expected count

($\boldsymbol{E_{ij}}$) denotes the expected number of cases in region *i*, which is frequently age-standardized or adjusted for possible confounders [23]. Since there is no information regarding this value for travel associated campylobacteriosis cases in New Zealand, we used the current observed national proportion (i.e., 7%) and applied it to the total campylobacteriosis cases of a DHB to obtain the expected counts of travel associated cases (E_i). The expected counts serve as an offset in the model *(4.2)*. The above formulation *(4.1*) is the standard generalized linear mixed model formulation with Poisson response [25]. Following which, the relative risk (θ_{ij}) of travel associated campylobacteriosis is defined on a logarithmic scale as the linear combination of covariates and random effects *(4.2)*.

$$log(\theta_{ij}) = log(E_{ij}) + \alpha + ui + vi + \beta_j + \gamma_j + \delta_j + \zeta_{ij} \qquad (4.2)$$

The intercept $\boldsymbol{\alpha}$ quantifies the average risk in all regions estimated from the data whereas u_i and v_i are area specific and non-spatial random effects, respectively, to capture the residual or unexplained source of variation in risk. The term u_i *represents* a spatially structured random effect and vi is unstructured over-dispersion. The entity β represents the global time effect (in years); γ and δ are temporally structured and unstructured random effects, respectively*;* and ξ stands for the interaction between time and space; *n* stands for the number of regions, and *J* identifies the year of travel (*J*=11 years). In order to determine the risk with regard to the main travel destinations, we used the observed travel associated cases that have confirmed their travel destination. Since these proportion of cases constitute only a small proportion from the total notification (2.5%), we did not have enough sample to include temporal effect. Therefore we summarized all cases with complete information regarding travel history to develop a risk map for the entire eleven year duration.

Bayesian analysis requires prior information to be specified for all unknown parameters in the model, a way to incorporate uncertainty about the parameters [26]. We specified a Conditional Autoregressive (CAR) normal prior for the spatial component (u_i) to take into account the neighborhood structure such that regions sharing the same border are more alike than those located far apart. The advantage of the CAR prior for spatial structure is that it allows one to incorporate spatial correlation into the model. The accommodation of spatial structure specifies the distribution of the random effects for the DHB regions as dependent on the collection of random effects of all the other neighboring DHBs while those regions share same border [27]. An exchangeable normal prior was used for the unstructured random effects (*vi*) [27]. A first order 'random walk' prior (RW1) was defined for the temporally structured effect γ. The RW1 prior reflects a belief that sudden jumps in the risk of travel associated campylobacteriosis between consecutive time points is unlikely, i.e successive time points are correlated. A detailed technical description of the CAR prior and the RW1 priors can be found in literature [28, 29]. Uninformative priors with mean 0 and a small precision (10^{-3}) are specified for all other unknown parameters in the model. This ensures that the results are dependent on the data.

We compared goodness-of-fit of models using the Deviance Information Criteria (DIC), pD [30] and the Conditional Predictive Ordinate (CPO) [31]. The DIC is a generalization of Akaike Information Criteria (AIC) for Bayesian models [32]. It is a composite measure of model fit that trades off a measure of model adequacy against a measure of model complexity. The DIC is given by the equation *(4.3)*:

$$DIC = D + pD, \qquad (4.3)$$

where *D* is the deviance of the model representing the fit of the model to the data, and *pD* is the effective number of parameters linked to model complexity. A difference of more than 5-10 units is regarded as strong evidence in favor of the model with smaller DIC [30]. The CPO, on the other hand, is a leave-one-out cross validation score that measures the predictive performance of a model [31] defined as *(4.4):*

$$CPO_i = \pi\,(Y_i^{obs} / Y_{-i}), \qquad (4.4)$$

where, $\boldsymbol{Y_i^{Obs}}$ refers to the observed value $\boldsymbol{i}$, and $\boldsymbol{Y_{-i}}$ denotes the observations $\boldsymbol{Y}$ with the i^{th} component removed. It measures the posterior probability of observing the value of $\boldsymbol{Y_i}$ when the model is fitted to all data except $\boldsymbol{Y_i}$. We used the cross validated logarithmic score (-mean (log (CPO)) for comparing competing models, where smaller value indicates a better predictive performance [33]. Both the DIC and the CPO can be obtained automatically from the INLA output by setting the option `cpo=TRUE` in the `control.compute` statement within the `inla(.)` call (see Appendix 4.1). We selected the model with the smallest DIC for predicting relative risk estimates in the maps. All the analysis was performed in *R* version 3.1.3 and the *INLA* package (Integrated Laplace Approximation) [34, 35].

Ethical consideration

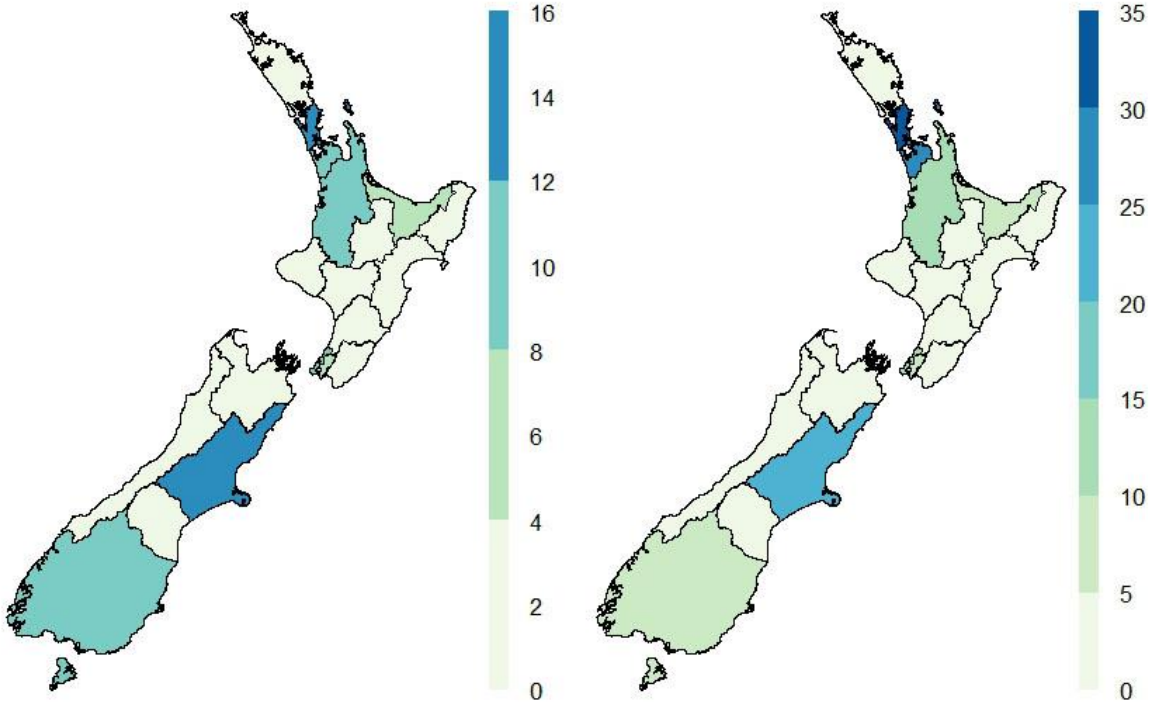
This study has been reviewed and approved by the New Zealand Southern Heath and Disability Ethics Committee (Ethics Reference: **MEC/12/EXP/029/AM03**).

4.3 Results

Descriptive results

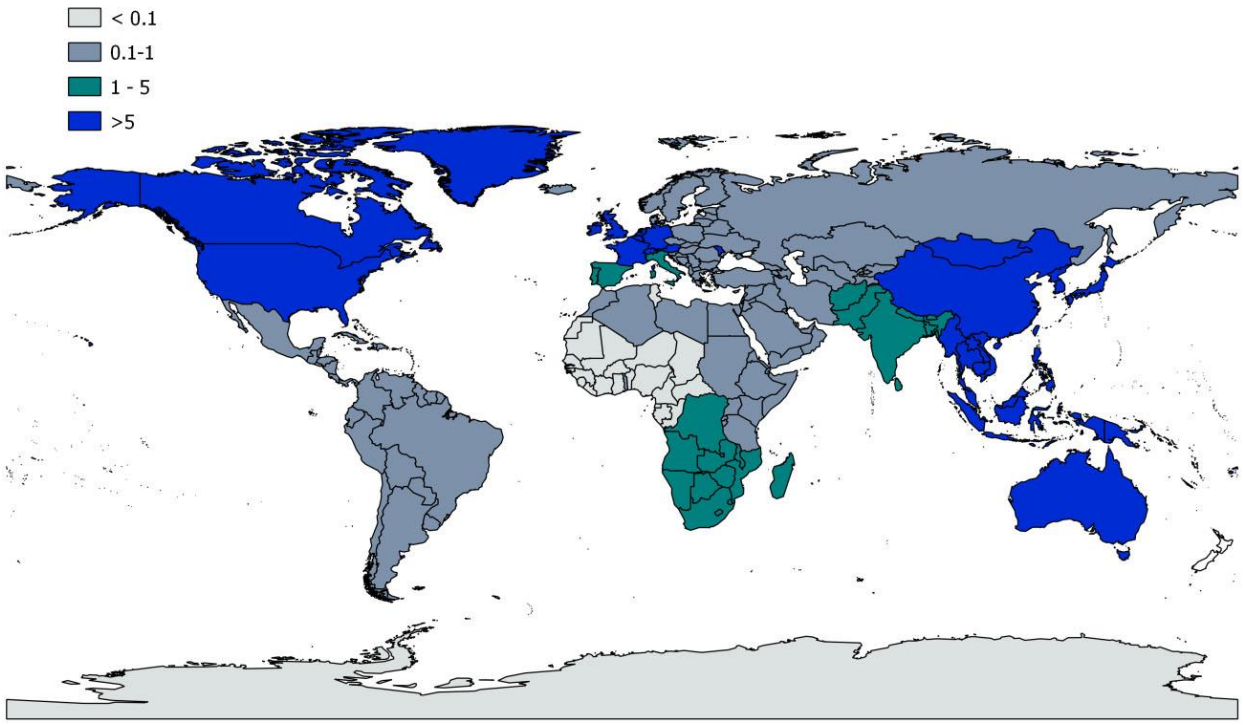
The distribution of total campylobacteriosis notifications and short term international travels from the 20 DHBs of New Zealand is shown in Fig 4.1. Majority of the campylobacteriosis notification was reported from Waitemata (12.8%), Canterbury (12.7%), Auckland (10.6%), Waikato (9.3%), Capital and Coast (8.9%) and Counties Manukau (8.6%). On the other hand about 56% of all travels were from Waitemata (16.7%), Counties Manukau (15.0%), Auckland (14.3%) and Canterbury (10.9%).

Figure 4.1. The total campylobacteriosis notification in 1000s (left panel) and short term international travels in 100,000s (right panel) of New Zealand residents (2000-2010).



The main travel destinations of New Zealand residents during the study period is shown in Figure 4.2. International destinations are categorized into twenty world regions based on geographical proximity of the countries of arrival and then travel counts were aggregated accordingly (see Figure 4.2). There were 18.3 million New Zealanders who have made a short-term international travels between 2000 and 2010. Most travel was to the Pacific region (68,074), East Asia (12,375), and Western Europe (7,500) per 100,000 travelers, while the least travel was recorded for Antarctica (7.6), Central and West Africa (16.8) per 100,000 travelers.

Figure 4.2. The main short term international travel destinations[1] of New Zealand residents in 100,000s (2000-2010).



[1]Notes: the main short term international travel destinations of New Zealanders is summarized into 20 world regions: **Antarctica:** Antarctica, Bouvet Island, French Southern and Antarctic Lands, Heard Island and McDonald Islands, South Georgia South Sandwich Islands; **Nordic countries** = Denmark, Finland, Iceland,

Norway; **Western Europe** = Austria, Belgium, France, Germany, Ireland, Luxembourg, The Netherlands, Switzerland, United Kingdom; **Southern Europe** = Italy, Malta, Monaco, Portugal, Spain; **Eastern Europe** = Bulgaria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia; **Eastern Mediterranean** = Albania, Cyprus, Former Yugoslavia, Greece, Israel, Turkey; **Russia and former USSR** = Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Russia, Tajikistan, Turkmenistan, Ukraine, Uzbekistan; **Arab countries and Iran** = Bahrain, Iraq, Iran, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, Syria, United Arab Emirates, Yemen; **Indian Subcontinent** = Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, Sri Lanka; **East Asia** = Brunei, Burma, Cambodia, China, Hong Kong, Indonesia, Japan, Laos, Malaysia, Mongolia, North Korea, Philippines, South Korea, Singapore, Taiwan, Thailand, Tibet, Viet Nam; T**he Pacific** = American Samoa, Australia, Cook Islands, Fiji, French Polynesia, Guam, Kiribati, Marshall Islands, Micronesia, Nauru, New Caledonia, New Zealand, Niue, Palau, Papua New Guinea, Samoa, Tokelau, Tonga, Tuvalu, Vanuatu, Wallis and Futuna; **North Africa** = Algeria, Egypt, Libya, Morocco, Tunisia; **West Africa** = Benin, Burkina Faso, Cape Verde, Ghana, Guinea, Guinea-Bissau, Ivory Coast, Liberia, Mali, Mauritania, Senegal, Sierra Leone, The Gambia, Togo; **East Africa** = Burundi, Djibouti, Eritrea, Ethiopia, Kenya, Rwanda, Seychelles, Somalia, Sudan, Tanzania, Uganda; **Central Africa** = Cameron, Central African Republic, Chad, Congo Brazzaville, Equatorial Guinea, Gabon, Niger, Nigeria, Republic of Congo, São Tomé et Principe; **Southern Africa** = Angola; Botswana, Lesotho, Madagascar, Malawi, Mauritius, Mozambique, Namibia, South Africa, Zambia, Zimbabwe; **North America** = Canada, USA; **Central America** = Belize, Costa Rica, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama; **Caribbean** = Antigua and Barbuda, Bahamas, Barbados, Bermuda, Cayman Islands, Cuba, Dominica, Dominican Republic, Grenada, Guadeloupe, Jamaica, Haiti, Martinique, Netherlands Antilles, Puerto Rico, S:t Christopher and Nevis, S:t Lucia / S:t Vincent, Saint Kitts-Nevis, The Grenadines, Trinidad and Tobago, Virgin islands; **South America** = Bolivia, Brazil, Colombia, Ecuador, French Guiana, Guyana, Honduras, Paraguay, Peru, Suriname, Uruguay, Venezuela.

Model selection

Table 4.2 shows the comparison of six fitted models. Adding spatial random effects (Model II) to the baseline model (Model I) significantly improved the fit of the model (a difference in DIC of 4898.5). Furthermore, expanding Model III to Model IV by incorporating temporal random effects reduced the DIC from 443.6 to 424.2 (a difference of 19.4). Adding a structured temporal random effect (Model V) and a space-time interaction term (Model VI) did not improve the model. Based on a combination of DIC, pD and CPO criteria, there is no significant difference between Models IV, V and VI (see Table 4.2). Therefore, the model containing spatial and temporal random effects

(Model IV) has the best combination of the model comparison parameters and parsimony, suggesting that it fits the data best. This model contained a structured and non-structured spatial random effects, in addition to a non-structured temporal term.

Table 4.2. The goodness-of-fit assessment of spatio-temporal models for estimating travel associated risk of campylobacteriosis in New Zealand.

Model	Parameters	DIC[1]	pD^2	M_CPO[3]
I	*alpha*	5418.5	1.2	12.43
II	$alpha + u + v$	520.5	19.3	1.17
III	$alpha + u + v + \beta t$	443.6	20.3	0.97
IV[4]	$\boldsymbol{alpha + u + v + \delta}$	**424.2**	**28.4**	**0.90**
V	$alpha + u + v + \delta + \gamma$	422.0	26.3	0.91
VI	$alpha + u + v + \delta + \gamma + \xi_i$	424.7	28.6	0.91

alpha: intercept; *u*: spatially structured random effect; *v*: spatially unstructured random effect; β: global temporal trend term; γ: structured temporal random effect (rw1); δ: unstructured temporal random effect; ξ_i: space-time interaction term.
[1]Deviance Information Criteria; [2]Effective number of parameters
[3]Mean of the log of Conditional Predictive Ordinate: M_CPO = -mean (log (CPO))
[4]**Bold** highlighted: selected model

We also computed the fractional variance for the unstructured and structured spatial components (i.e., the proportion of variance explained by each component). Dividing the variance of each random component by the total variance gives the proportion explained by each part. The structured random effect captured much of the variance attributed to unobserved confounders as compared to the unstructured and the temporal components (see Table 4.3).

Table 4.3. The fraction of standard deviation (SD) explained by each random effect component from the final model.

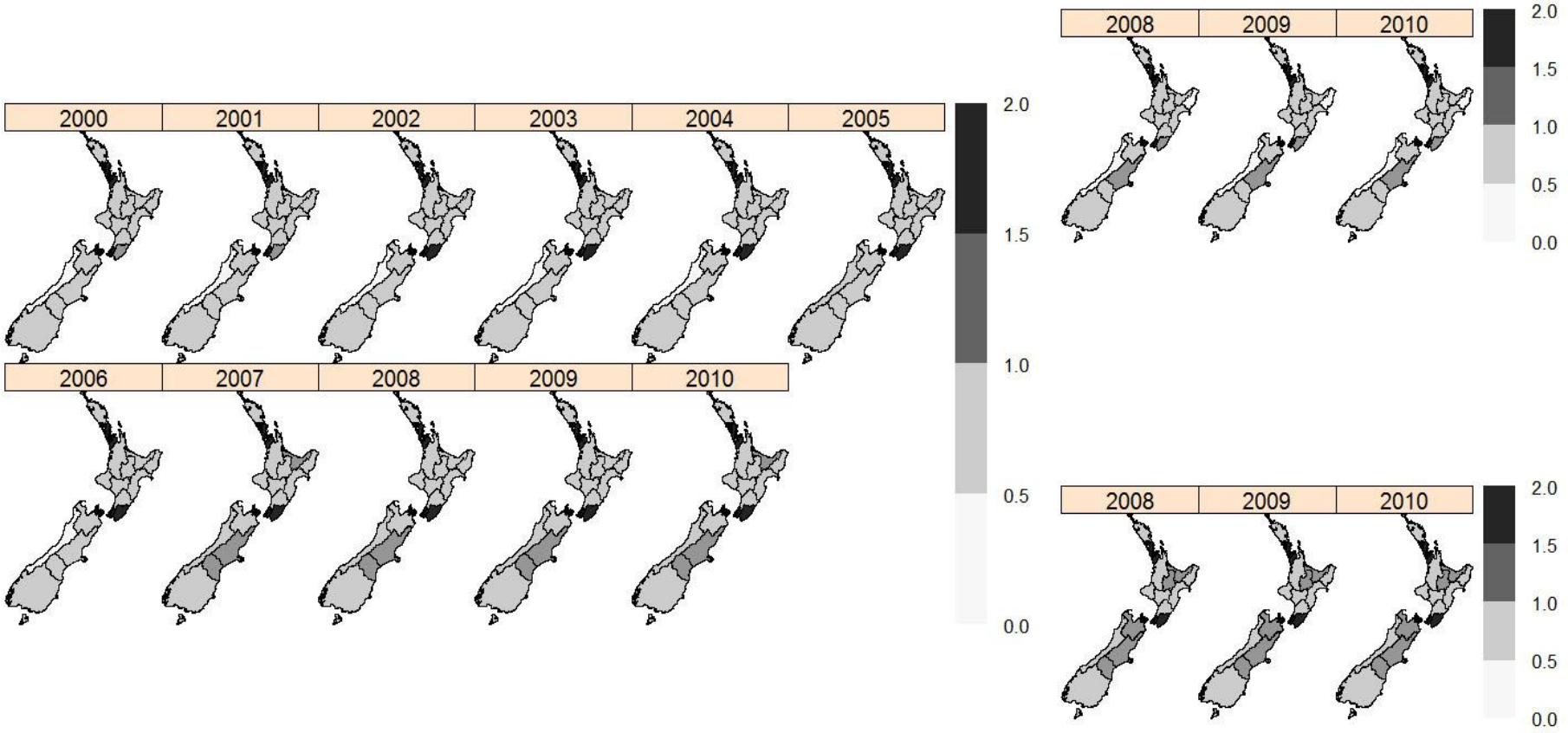
Random effect Parameters	Mean	95% Credibility Interval	
		Lower	Upper
u	0.73	0.52	0.99
v	0.01	0.003	0.02
δ	0.09	0.05	0.14

Notes: Variance=SD^2 , variance explained by ***u*** = *98%*; *u:* spatially structured random effect; ***v***: unstructured random effect; ***δ***: temporally unstructured random effect.

Spatial distribution of risk

The risk map of travel associated campylobacteriosis in New Zealand between 2000 and 2010 is shown in Figure 4.3. The mean relative risk indicated in the map is estimated using the model that takes into account both temporal and spatial random effects. The region with highest relative risk is characterized by darker colors and low risk of travel associated campylobacteriosis is shown in lighter colors. The Capital and Coast, Hutt, Wairarapa, Auckland, Counties Manukau and Waitmata DHBs have comparatively higher risk of travel associated campylobacteriosis than other DHBs. The progression of the risk over the period between 2000 and 2010 is evident from the risk map. Particularly from 2007 onwards there is an increasing risk of travel related campylobacteriosis in the Canterbury region (Fig 4.3).

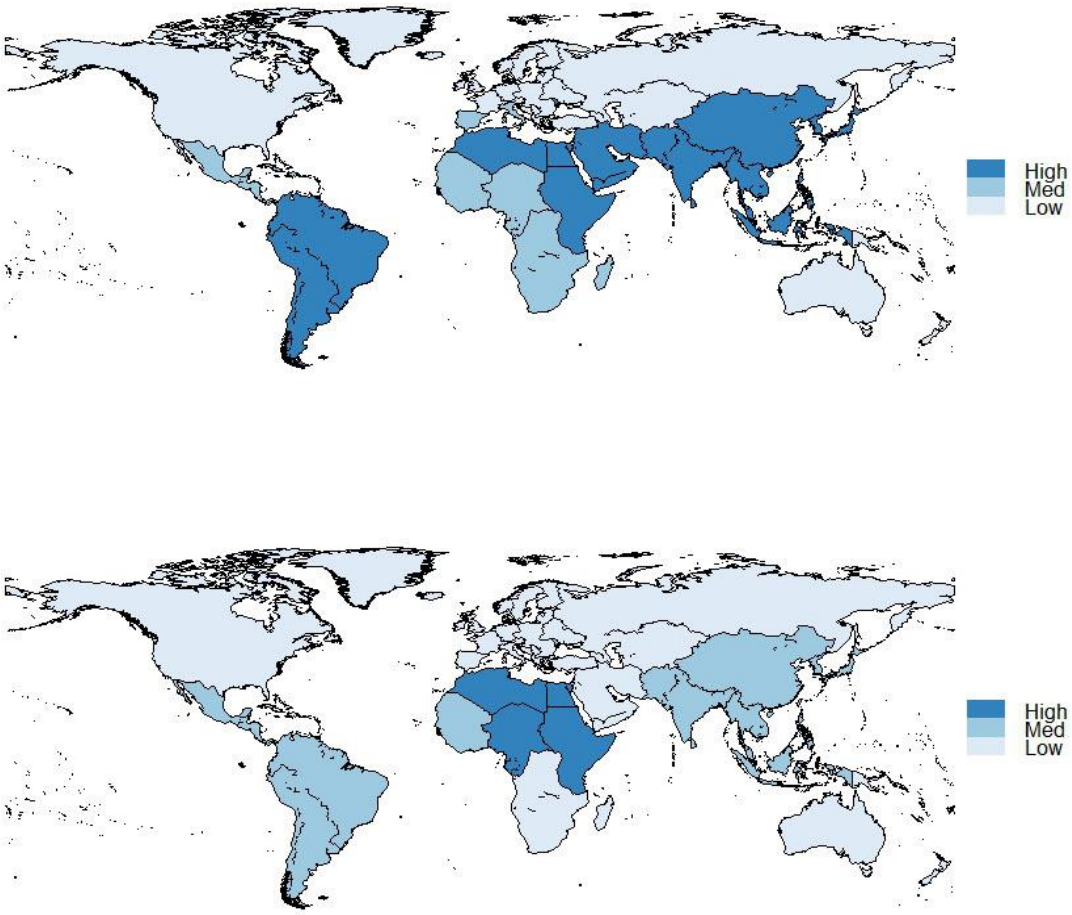
Figure 4.3. The spatial and temporal distribution of travel associated campylobacteriosis risk in New Zealand using observed and model predicted travel associated cases



Notes: Left panel: posterior mean Relative Risk (2000-2010); Right top: Lower 95% CI; right bottom: Upper 95% CI (both 2008-2010); the regions in the map are New Zealand District Health Boards (DHBs).

Additionally, the risk of travel related cases with respect to 19 main travel destinations (world regions) has been estimated (see Figure 4.4).

Figure 4.4. The distribution of campylobacteriosis risk with regard to main travel destinations (upper panel: using Ekdahl's campylobacteriosis risk[1], bottom panel: using a model of observed travel associated cases in New Zealand, 2000-2010).



[1]Ekdahl qualitative campylobacteriosis risk (top panel): the original Ekdhal's risk map shows a qualitative campylobacteriosis risk of returning travelers to Sweden (high, medium and low risk regions). The risk map at the bottom panel is developed using confirmed case notifications in New Zealand with known short term travel destinations (n=3109). The three risk groups were created by classifying the relative risk into three classes (>0.009, 0.006-0.009, <0.006). Since only 3109 (<2%) of the total campylobacteriosis notifications in New Zealand was used for constructing this map, the true risk may have been underestimated and therefore the map should only be used for general comparison if there are overlaps from Ekdhal's risk map.

Notes: Countries of main destination are grouped into nineteen world regions (excluding Antarctica): Nordic countries, Western Europe, Eastern Europe, Eastern Mediterranean, Southern Europe, Russia and former USSR, Arab countries and Iran, Indian Subcontinent, East Asia, The Pacific, North Africa, West Africa, East Africa, Central Africa, Southern Africa, North America, Central America, Caribbean and South America.

4.4 Discussion

There is a scarcity of information regarding the risk of campylobacteriosis attributable to overseas travel in New Zealand. One of the biggest challenges for identifying travel associated cases from those acquired within the country has been lack of complete data. This study attempted to estimate the spatial and temporal distribution of the risk of travel associated campylobacteriosis in New Zealand using observed and model predicted travel related cases. The dataset used for this analysis does not identify whether or not individuals have made multiple trips during the study period. Therefore, it should be noted that some individuals may have traveled multiple times during the study period, and may have contributed more than once to the dataset at different times.

Spatial estimation of disease risk in the Bayesian context offers a mechanism to "borrow strength" from neighboring areas to improve local estimates, resulting in the smoothing of extreme rates caused by small local sample sizes [20]. Some DHBs have very low rates of travel associated

campylobacteriosis (e.g., West Coast and Northland) while others have much higher rates (e.g., Auckland and Counties Manukau). Spatial analysis of such data using the traditional approach may overestimate the risk because the results may be based on a few cases from regions with smaller populations [36]. In the latter case, the Bayesian methods provide a flexible platform to incorporate various forms of random effects and spatial correlation for developing a best model to describe spatial and temporal risk. The Bayesian approach takes into account the uncertainties in the modeling process about shrinking extreme disease risk values towards the local rates [37]. A significant improvement of the model fit through incorporation of spatially structured and unstructured random components in our analysis indicated that there is a need to consider spatial and temporal dimensions while trying to determine the risk of travel associated campylobacteriosis in New Zealand. It has been previously indicated that, in general, marked regional differences can be found regarding campylobacteriosis risk that may be associated with travel, because the tendency and destinations of individuals travelling from various regions may vary [13].

A consistently higher relative risk of travel associated campylobacteriosis was observed in the major urban areas of New Zealand throughout the study period. This could be attributed to a higher rate of short term international trips from these DHBs to certain destinations that have a higher incidence of campylobacteriosis infections [38]. The diverse ethnic structures of the urban areas in New Zealand could contribute to the majority of international travels from these areas to high risk regions. For example, about 88% of all New Zealand-based Pacific population reside in urban areas of New Zealand, with 67% living in the Auckland region alone [39]. Furthermore, higher risk may also be related to individual attributes such as duration of stay upon overseas travel, personal precaution and other unknown factors that require further investigation [40].

Although the similarity between Ekdhal's qualitative risk zones and our risk map regarding travel destinations is evident, there are some differences. This discrepancy might be due to lack of sufficient information on travel destinations for the notified cases in New Zealand. The proportion of campylobacteriosis cases that was used to build a model for producing the risk map may have underestimated the true risk with regard to travel destinations. Variation in the availability of information concerning travel destinations for the New Zealand campylobacteriosis cases may also have contributed to the apparent differences between the two risk maps.

The application of Bayesian modelling in the current study conferred advantages over traditional, frequentist modelling approaches, because the spatial dependence structure of the observed data is incorporated in the modeling process. The uncertainties surrounding the predicted risk estimates can be corrected for, enabling objective decisions about priorities for future data collection.

Through comparison of models with different structures, the current study suggested that the incorporation of temporal and spatial dimensions in risk mapping makes the best usage of the available data. It can be inferred that the spatial and temporal components of the disease risk play a large role in explaining sources of variation without direct measurement of the variation.

4.5 Conclusions

Disease mapping of spatio-temporal data about travel associated campylobacteriosis using Bayesian frameworks identifies high disease risk areas that require further attention from health

policy makers. Local area specific estimates of campylobacteriosis risk inform strategic allocation of public health resources. Mapping disease risks in different areas subject to surveillance strengthens the surveillance of diseases. The result provides a better understanding of the residual spatial risk and the variability of risk due to spatial and non-spatial components. Further investigation of areas with high risk of travel associated campylobacteriosis yields clues towards the understanding of disease causation.

Conflicts of interest

None

Acknowledgements

This work was supported by the NIH Ruth L. Kirschstein National Research Service Award Institutional Training Grant T32 RR023916 and T32 OD010423. Parts of this material are based on data and information provided by the Institute of Environmental Science & Research Ltd on behalf of the Ministry of Health. However, the analyses, conclusions, opinions and statements expressed herein are those of the authors, and not necessarily those of the Institute of Environmental Science & Research Ltd or the Ministry of Health.

4.6 References

1. Hearnden M, Skelly C, Eyles R, Weinstein P: The regionality of campylobacteriosis seasonality in New Zealand. Int J Environ Health Res 2003, 13:337–348.

2. Blaser MJ: Epidemiologic and Clinical Features of Campylobacter jejuni Infections. J Infect Dis 1997, 176(Supplement 2):S103–S105.

3. Zenner D, Gillespie I: Travel-Associated *Salmonella* and *Campylobacter* Gastroenteritis in England: Estimation of Under-Ascertainment Through National Laboratory Surveillance. J Travel Med 2011, 18:414–417.

4. Altekruse SF, Stern NJ, Fields PI, Swerdlow DL: Campylobacter jejuni--an emerging foodborne pathogen. Emerg Infect Dis 1999, 5:28–35.

5. Nichols GL, Richardson JF, Sheppard SK, Lane C, Sarran C: Campylobacter epidemiology: a descriptive study reviewing 1 million cases in England and Wales between 1989 and 2011. BMJ Open 2012, 2.

6. Sears A, Baker MG, Wilson N, Marshall J, Muellner P, Campbell DM, Lake RJ, French NP: Marked Campylobacteriosis Decline after Interventions Aimed at Poultry, New Zealand. Emerg Infect Dis 2011, 17:1007–1015.

7. Institute of Environmental Science and Research Ltd (ESR). Notifiable and Other Diseases in New Zealand: Annual Report 2011. Porirua: ESR; Apr 2012, Updated Oct 2012. [https://surv.esr.cri.nz/PDF_surveillance/AnnualRpt/AnnualSurv/2011/2011AnnualSurvRpt.pdf]

8. Silva J, Leite D, Fernandes M, Mena C, Gibbs PA, Teixeira P: Campylobacter spp. as a Foodborne Pathogen: A Review. Front Microbiol 2011, 2.

9. Friedman CR, Hoekstra RM, Samuel M, Marcus R, Bender J, Shiferaw B, Reddy S, Desai Ahuja S, Helfrick DL, Hardnett F, Carter M, Anderson B, Tauxe RV, the Emerging Infections Program FoodNet Working Group: Risk Factors for Sporadic Campylobacter Infection in the United States: A Case-Control Study in FoodNet Sites. Clin Infect Dis 2004, 38:S285–S296.

10. Kapperud G, Espeland G, Wahl E, Walde A, Herikstad H, Gustavsen S, Tveit I, Natås O, Bevanger L, Digranes A: Factors Associated with Increased and Decreased Risk of Campylobacter Infection: A Prospective Case-Control Study in Norway. Am J Epidemiol 2003, 158:234–242.

11. Stafford RJ, Schluter P, Kirk M, Wilson A, Unicomb L, Ashbolt R, Gregory J: A multi-centre prospective case-control study of campylobacter infection in persons aged 5 years and older in Australia. Epidemiol Infect 2007, 135:978–988.

12. Studahl A, Andersson Y: Risk factors for indigenous campylobacter infection: a Swedish case-control study. Epidemiol Infect 2000, 125:269–275.

13. Ekdahl K, Andersson Y: Regional risks and seasonality in travel-associated campylobacteriosis. BMC Infect Dis 2004, 4:54.

14. Hongoh V, Hoen AG, Aenishaenslin C, Waaub J-P, Bélanger D, Michel P, $author.lastName $author firstName: Spatially explicit multi-criteria decision analysis for managing vector-borne diseases. Int J Health Geogr 2011, 10:70.

15. Kanevski M, Timonin V, Pozdnukhov A: Machine Learning for Spatial Environmental Data: Theory, Applications, and Software. CRC Press; 2009.

16. Lawson AB, Browne WJ, Rodeiro CLV: Disease Mapping with WinBUGS and MLwiN. John Wiley & Sons; 2003.

17. Bernardinelli L, Clayton D, Montomoli C: Bayesian estimates of disease maps: How important are priors? Stat Med 1995, 14:2411–2431.

18. Waller LA, Carlin BP: Disease mapping. Chapman Hall CRC Handb Mod Stat Methods 2010, 2010:217–243.

19. Lesaffre E, Lawson AB: Bayesian Biostatistics. John Wiley & Sons; 2012.

20. Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M: Bayesian analysis of space—time variation in disease risk. *Stat Med* 1995, 14:2433–2443.

21. Knorr-Held L, Besag J: Modelling risk from a disease in time and space. *Stat Med* 1998, 17:2045–2060.

22. Knorr-Held L: Bayesian modelling of inseparable space-time variation in disease risk. *Stat Med* 2000, 19:2555–2567.

23. Bernardo JM: Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting. Clarendon Press; 1999.

24. Breslow NE, Clayton DG: Approximate Inference in Generalized Linear Mixed Models. *J Am Stat Assoc* 1993, 88:9.

25. Gelman A, Hill J: Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press; 2006.

26. Besag J, York J, Mollié A: Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 1991, 43:1–20.

27. Blangiardo M, Cameletti M, Baio G, Rue H: Spatial and spatio-temporal models with R-INLA. *Spat Spatio-Temporal Epidemiol* 2013, 4:33–49.

28. De Oliveira V: Bayesian analysis of conditional autoregressive models. *Ann Inst Stat Math* 2012, 64:107–133.

29. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A: Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol* 2002, 64:583–639.

30. Pettit LI: The Conditional Predictive Ordinate for the Normal Distribution. J R Stat Soc Ser B Methodol 1990, 52:175–184.

31. Spiegelhalter DJ, Best NG, Carlin BP: Bayesian Deviance, the Effective Number of Parameters, and the Comparison of Arbitrarily Complex Models. 1998. Research Report 98-009, Division of Biostatistics, University of Minnesota. www.biostat.umn.edu/~brad.

32. Lindgren F, Rue H: Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software 2015, 63.

33. R Core Team (2015). R: A language and environment for statistical computing. R Foundatio n for Statistical Computing, Vienna, Austria. [URL http://www.R-project.org/]

34. Jones ME, Swerdlow AJ: Bias in the Standardized Mortality Ratio when Using General Population Rates to Estimate Expected Number of Deaths. *Am J Epidemiol* 1998, 148:1012–1017.

35. Best N, Richardson S, Thomson A: A comparison of Bayesian spatial models for disease mapping. *Stat Methods Med Res* 2005, 14:35–59.

36. DuPont HL: Systematic review: prevention of travellers' diarrhoea. *Aliment Pharmacol Ther* 2008, 27:741–751.

37. About Pacific peoples in New Zealand. Ministry of Pacific Island Affairs, New Zealand [http://www.mpia.govt.nz/pacific-peoples-in-new-zealand/]

38. Ravel A, Nesbitt A, Marshall B, Sittler N, Pollari F: Description and Burden of Travel-Related Cases Caused by Enteropathogens Reported in a Canadian Community. *J Travel Med* 2011, 18:8–19.

Chapter 5: Mapping of *Plasmodium falciparum* exposure in rural homesteads of the Victoria Lake Crescent and Busia region, Western Kenya

Amene, E.[1], N.C. Wamae[5], W.A. de Glanville[2,3], E. M. Fèvre[3,4], M. Tremblay[1], D. Döpfer[1]*

[1]*Department of Medical Sciences, University of Wisconsin-Madison, 2015 Linden Drive 53706 Madison WI*

[2]*Centre for Immunity, Infection and Evolution, Institute for Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Ashworth Laboratories, West Mains Road, Edinburgh EH9 3JT, UK*

[3]*International Livestock Research Institute, Old Naivasha Road, PO Box 30709, 00100 Nairobi, Kenya*

[4]*Institute of Infection and Global Health, University of Liverpool, Leahurst Campus, Chester High Road, Neston, CH64 7TE, United Kingdom, UK.*

[5]*Centre for Microbiology Research, Kenya Medical Research Institute (KEMRI), Mbagathi Road, Nairobi, Kenya*

*Corresponding author
Ermias Amene
2015 Linden Drive
53706 Madison WI
amene@wisc.edu

Abstract

Background: Disease mapping is frequently applied to assess the pattern of disease and identify areas characterized by abnormally high or low risk of disease. Bayesian frameworks provide a convenient platform to generate a map that reflects the inherent heterogeneity of disease risk in small areas. This study applies a Bayesian hierarchical disease mapping approach to estimate the risk of *Plasmodium falciparum* exposure in the rural homesteads of Western Kenya.

Methods: Important risk factors of *Plasmodium falciparum* exposure were identified by applying a Bayesian Model Averaging approach to a dataset obtained from the People's Animals and their Zoonoses project. A Bayesian hierarchical model including spatial and non-spatial random effects was fitted to the data. Various models were compared using the Deviance Information Criteria and the preferred model was employed to predict the risk of *Plasmodium falciparum* exposure in 143 Sub-locations in the Busia region, Western Kenya.

Results: Five and seven predictors of malaria were selected from 22 and 25 candidate variables from two separate datasets, respectively. The disease risk and 95% Credibility Intervals adjusted for the covariates and spatial random effects ranges from 0.19 (0.07, 0.35) to 2 (1.01, 3.38) in Kokeyo and Umla Sub-locations (SL); and 0.16 (0.10, 0.26) to 1.90 (0.79, 3.82) in Yenga and Siranga SLs, for the two datasets, respectively. The variability of the risk in the region is mainly attributed to spatial factors.

Conclusion: The subset of variables identified in this study help to prioritize predictors for *Plasmodium falciparum* exposure in the region. The health authorities engaged in control and prevention of malaria in the Busia region can adopt the risk maps as a useful tool for identifying

priority regions for intervention efforts given the respective individual risk profiles per SL. The methods in this report can guide data collection, study designs and they can be applied to other disease studies that include spatial representation and prioritization of risk. Data covering a wider geographical area and including environmental and climatic factors can improve the prediction of malaria risk.

Key words: Busia, *Plasmodium falciparum* , Bayesian, mapping

5.1 Background

Plasmodium falciparum is a protozoal parasite predominantly responsible for causing malaria in many parts of Sub-Saharan Africa. Malaria is the leading cause of human illnesses and deaths in this region including Kenya [1]. Due to high mortality and morbidity associated with the disease, it has been recognized as a major health and socioeconomic burden in the region [2]. Although substantial reduction in the transmission of malaria has been achieved in the past decade in endemic areas of the Sub-Saharan Africa, the disease still remains a major health problem locally and worldwide [3].

It has been reported that the Western part of Kenya has the highest burden of malaria in the country [4–6]. A number of risk factors are associated with the high occurrence of the disease in the region such as the agro-ecological zone, poverty, ignorance about the disease and lack of health resources [7]. Moreover, there is a high density of humans and animals living in close proximity in the region carrying a heavy burden of endemic and epidemic diseases including malaria. Prevalence and transmission intensity varies with geographical regions and therefore identifying important

predictors and consideration of the regional distribution of *Plasmodium falciparum* exposure helps to prioritize strategic health resources allocation.

With the advent of routine health data indexed at a fine geographical resolution, small area risk mapping has become an established technique in studying the epidemiology of diseases [8]. The mapping of disease risks or rates, a technique that is frequently applied to assess the pattern of disease and to identify areas characterized by abnormally high or low risk, is important for the design and validation of epidemiological studies [9, 10]. Since disease risk is often not evenly distributed over a geographic area, detailed understanding of the geographical distribution of disease rates has important public health and epidemiological implications while studying diseases in both humans and animals.

Conventional analysis of count data in a small geographic grid usually produces maps that rely on the raw counts of cases. As a result, the estimated spatial distribution of risk does not reflect the underlying variation of risk that depends on risk "hot spots", because the analysis relies on the raw data alone (see Section 1.2.4 of Chapter 1) [11]. Consequently, the hierarchical Bayesian disease mapping approach has been suggested as one way to generate a risk map that reflects the inherent heterogeneity of risk in a geographical area [10, 12, 13]. In recent years, Bayesian approaches have been increasingly used for modeling spatial distributions of a number of diseases including malaria. Bayesian Disease Mapping (BDM) provides a way to stabilize or smooth regional risk estimates [14–16]. In the Sub-Saharan Africa where malaria incidence is the highest, the Mapping Malaria Risk in Africa (MARA) project has been working towards developing a complete malaria risk atlas for targeted control of the disease [17]. MARA uses a biological model that sets decision

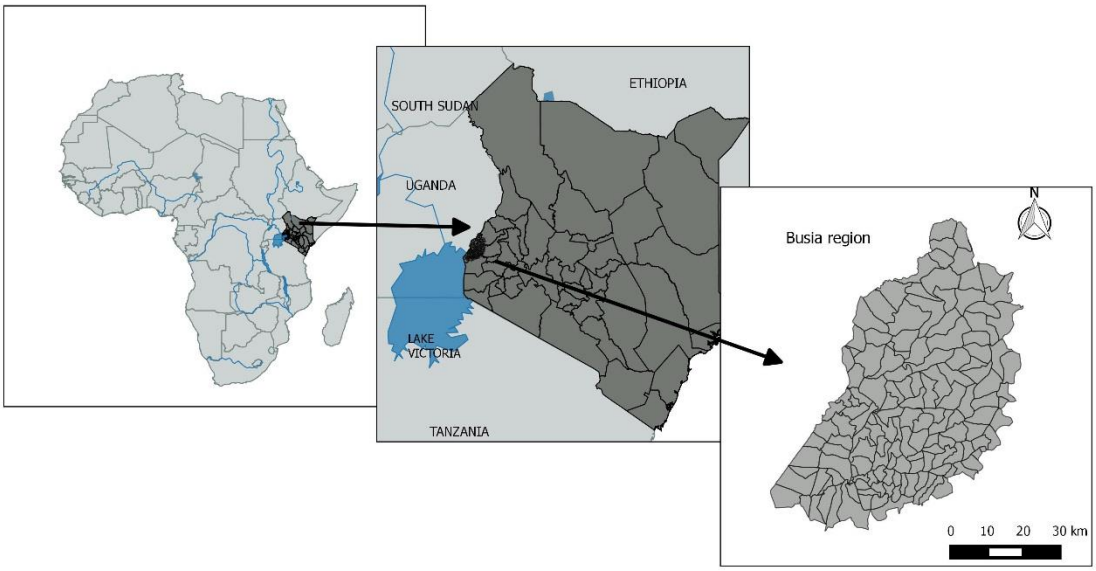
rules on how precipitation and temperature affect the development of *Plasmodium falciparum* (the main etiology of malaria in the region) and survival and breeding of the mosquito vector host [18]. In the current study, we applied Bayesian model averaging (BMA) and BDM approaches to select a subset of meaningful predictors and to estimate the smoothed relative risk, respectively, for mapping of *Plasmodium falciparum* exposure in the Victoria Lake Crescent and Busia region, Western Kenya.

5.2 Methods

Study area

The Busia County covers an area of 1694.5 km^2 that is located in the Western Province of Kenya, bordered by Lake Victoria in the South and by Uganda in the West (Fig 5.1).

Figure 5.1. Map of the study area (Busia region, Western Kenya).



Note: The Busia County is located in the Western part of Kenya, bordered by Lake Victoria in the South and Uganda in the West.

The population of Busia County was about 816,452 people in 2012 distributed into 10 Divisions, 60 Locations and 181 Sub-locations (SL) where the latter are 1st administrative units of the country[19]. The SLs consist of homesteads, which are the smallest geographical units typically consisting of several huts in close proximity and occupied by many generations of the same family. The spatial unit for the current disease mapping effort is the SL.

Data source

The dataset for this analysis was obtained from the People's Animals and their Zoonoses project (PAZ). Since 2010, the PAZ project has investigated endemic and neglected zoonotic diseases in the Busia region, Western Kenya with the aim of identifying the link between risk factors, socio-economic status, and incidence of infectious diseases [20]. The original dataset was collected in 2010 and observations were recorded from 416 randomly selected rural homesteads [21]. For the purpose of our study, all individual level data and homestead level data were aggregated to the SL level and linked by a unique SL identifier. The number of individuals living in a homestead were used as a weighting variable during the data analysis.

Plasmodium falciparum exposure

A positive case was defined as a subject being positive for *Plasmodium falciparum* in thick or thin blood smears [22]. Therefore, a case in our study represents more an exposure to malaria at some point in time rather than an acute infection with clinical signs. Counts of positive cases per homestead and other relevant information (such as the number of people residing in a homestead

that are at risk of exposure, and the potential predictors of the disease) were aggregated at the SL level and used to estimate the spatial distribution of risk in the region.

Predictor variables

The original raw dataset was divided into cattle, human and homestead datasets collected using separate questionnaires, each containing a number of socioeconomic, environmental and health-related information [21]. We start our analysis from two subsets of predictors that were previously selected from several hundred variables during a stepwise model selection process that consisted of a combination of Elastic net regularization (see Section 1.2.3 of Chapter 1) and stepwise regression analysis [23]. The first subset (hereafter called Subset A) consists of 224 homesteads with 22 variables describing the homesteads, humans and their cattle, whereas the second subset (hereafter called Subset B) was comprised of 415 homesteads with 25 variables describing information about homesteads and humans only. The reason for splitting the raw dataset into two is that, for half of the homesteads the 'cattle variables' were missing, because no cattle were being kept at those homesteads.

Statistical analysis

Selection of variables

Shrinking the large data set to a subset of most meaningful variables is a pre-requisite for BDM to avoid overparameterization, overfitting and instability of the model outcomes [24]. Since all the variables in both subsets have been selected from a much larger pool of predictors, further reduction of these variables requires careful consideration of both biological and statistical

significance. In order to select a meaningful subset of variables, we used a Bayesian Model Averaging (BMA) approach that compares models using Bayes' rule [25]. The application of BMA for variable selection has been extensively described in literature [26–28]. BMA is a Bayesian framework approach for choosing among competing models while accounting for model uncertainty [25]. For example, in a dataset with 25 predictors, there are up to 2^{25} possible models that can be fitted for all combinations of the predictors. However, only a small proportion of these models are usually considered to have an acceptable goodness-of-fit (usually the first 5000 models are retained and the results are reported based on them) [29, 30]. In BMA, the potential explanatory variables are quantified by their Posterior Model Probabilities (PMP) that indicate the importance of the variables for explaining the data [30]. In other words, the PMPs are the probabilities of a variable being in a model. We considered a strong association between the risk factor (X_i) and the malaria risk (Y_i) given the data (D), whenever the Posterior Inclusion Probability (PIP) (i.e., the sum of PMPs for all models wherein the variable was incorporated) was greater than 75% [26], i.e., $Pr(\beta \neq 0/ D) > 0.75$, where β's are regression parameters. This cut-off was chosen, because the models showed a jump in PIP below and above the value of 75%.

Model Development

After the shrinkage step described above, the selected subsets of predictors were used for a Poisson regression model and the outcomes were mapped. Poisson regression in the Bayesian hierarchical framework has been used for disease mapping whenever the counts of cases were low and located on a relatively small geographical area or whenever the disease was a rare event [10, 31]. While mapping *P. falciparum* exposure, accounting for the spatial dependence of the neighboring SLs

aids in "borrowing information" from neighboring regions to strengthen the robustness of risk estimates. Bayesian formulation of this problem has been proposed to overcome the substantial extra variability of small area disease risk estimates [32, 33].

Counts of cases of are considered to follow a Poisson process. The generalized form of the equation for the Poisson model used is shown in equation *(5.1)*

$$O_i \sim Poisson\ (\lambda_i E_i) \qquad (5.1)$$

where O_i denotes the observed counts of positive cases in SL_i , λ_i is the relative risk of the disease in SL_i and E_i is the expected number of cases in SL_i. The expected number of cases, used as an offset in our model, is defined as the product of the number of individuals at risk of the disease in a given geographical area and a constant "baseline" risk [10, 34]. The baseline risk for malaria was computed by dividing the previously reported malaria prevalence in the Western lake endemic zone (i.e., 38%) [2] to the total number of individuals that were evaluated in each SL during the PAZ project as shown in equation *(5.2)*. For example, the expected count of malaria cases in a homestead in each SL was defined as:

$$E_i = (Cases_i / NumbHS)*prev, \qquad (5.2)$$

where $\boldsymbol{Cases_i}$ is the total number of malaria cases in homesteads per SL, ***NumbHS*** refers to the total number of residents in the homesteads per SL and ***prev*** is the known prevalence of malaria in the region.

To adjust for the impact of the observed predictors and random effects on the incidence of exposure, we fitted a Poisson random effect regression model specified as a linear combination of

a common intercept, two independent random terms and relevant predictors in equation *(5.3)*. The mean log-relative risk was therefore modeled as follows:

$$log\ (\lambda_i) = log(E_i) + (\alpha + \beta_i X_{ij}) + \theta_i + \psi_i \qquad (5.3)$$

$i=1,2,\ldots I$ (number of Sub-locations), $j=1,2\ldots J$ (number of predictors)

where α is the intercept, θ_i and ψ_i are structured and unstructured random effects, X_i is a vector of predictors, and β_i represents the regression coefficients. The structured random effect (also called correlated heterogeneity) denotes local spatial structure by taking into account the neighborhood structure of the SLs. The neighborhood structure reflects that areas in close geographical proximity to one another may have more similar risk compared to those areas that are far apart [35]. The unstructured random component (uncorrelated heterogeneity) represents the variation in risk that does not depend on geographic location. The random effect for each area is the sum of the spatially structured and unstructured components [35]. This formulation allows the model to compensate for how much of the residual risk is attributable to spatially structured variation and how much is unstructured over-dispersion after accounting for the effect of the predictor variables. The relative contribution of the variance explained by the spatial component was computed for each model.

Bayesian analysis requires that all unknown parameters in the model have a prior information that describes the uncertainty of the parameters [36]. For all the regression coefficients, standard non-informative priors were assigned, i.e., those are normal distributions with mean 0 and precision 10^{-3} (this precision is equivalent to a variance of 1000). In WinBUGS a normal distribution is defined by the mean and precision where precision is inverse of the variance [37]. The spatial relationship between the SLs (θ) is assigned a so-called CAR (Conditional Autoregressive) prior,

making sure that neighboring SLs sharing a same border are more alike than arbitrary locations [35]. The CAR prior defines such structure through an adjacency weights matrix, i.e., weight =1 for any two SLs that shared a border, and weight =0 otherwise. The value of each spatially structured random effect depends on the mean and variance of disease risk of adjacent SLs. In other words, the CAR prior specifies that the distribution of each area specific structured effect, given all other spatial effects, is a normal distribution with mean equal to the average of its bordering regions and variance inversely proportional to the number of regions sharing the same border [35, 38]. A detailed technical description of the CAR models can be found in literature [14, 35, 36]. In addition, we assigned a normal prior distribution with mean 0 and precision 10^{-3} for the non-spatial random effect (ψ) (see Appendix 5.1).

The goodness-of-fit and complexity in terms of number of parameters estimated were compared for the fitted models using the Deviance Information Criterion (DIC) [39]. The DIC is a generalization of Akaike Information Criteria (AIC) that indicates the sum of the posterior mean deviance (to reflect the model fit), and the effective number of parameters, pD (to reflect model complexity) [37]. A difference in the DIC values of 7 or more is considered to be significant in favor of the model with the smallest DIC [40].

All the statistical analysis was performed in R version 3.2.1 [41]. Bayesian analysis was implemented in WinBUGS 1.4 (Windows for Bayesian Inference Using Gibbs Sampling) program that can be called within the R environment through the *r2winbugs* package [42]. We used the *maptools* library to import shape files to R, and the *spdep* library to create the adjacency matrices

of the regions (i.e., to define the neighborhood structure of the SLs) for the CAR model (see Appendix 1.1).

Each model was fitted in WinBUGS using 50,000 iterations while discarding the first 5,000 iterations (so-called "burn-ins"). Convergence of the models was assessed by running two chains of initial values and visually inspecting a subset of trace plots and density plots (see WinBUGS code provided in Appendix 5.1). Both R and WinBUGS are freely available on the web using the following links: [http://www.r-project.org/] and [http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/], respectively. The *BMS* (Bayesian Model Sampling) package in R was used to perform Bayesian model averaging [30].

Ethical clearance

Human data and samples collected in this study were collected following approval by the KEMRI Ethical Review Committee, SC#1701. Animal samples were collected following approval from the Roslin Institute Animal Welfare and Ethical Review Committee AWA004. In addition, the Institutional Review Board (IRB) at UW-Madison approved this study (IRB no. 2013-0072).

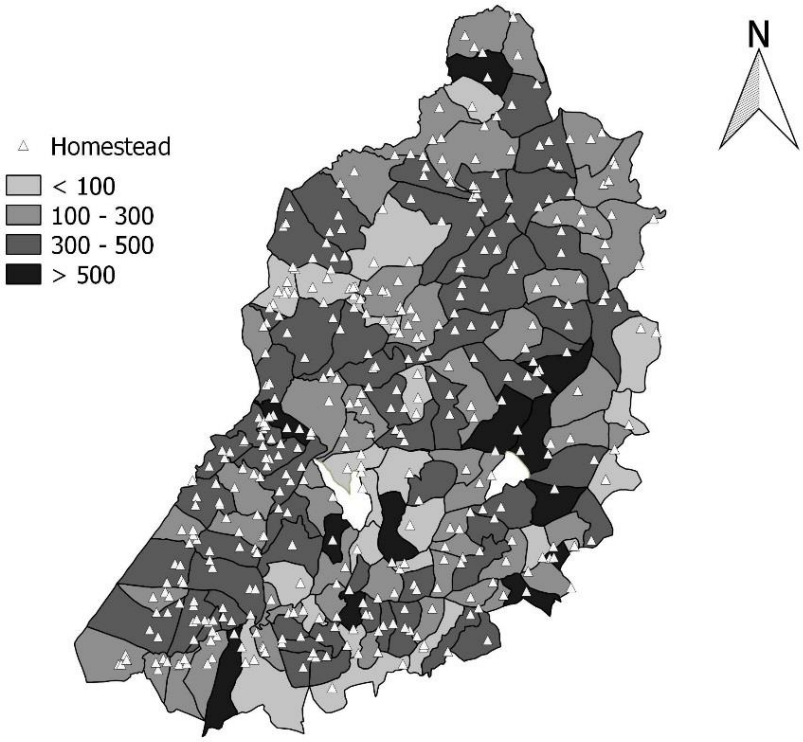
5.3 Results

Descriptive analysis

The 415 homesteads assessed in this study are distributed across 143 SLs of the Busia region ranging from 0 to 6 homesteads per SL (see the distribution of homesteads across SLs in Figure

5.2). A total of 624 cases was reported from a population of 2098 living in the homesteads of the Busia region, representing a crude incidence of 297 cases per 1000 individuals per year for the region (Fig. 5.2).

Figure 5.2. The distribution of homesteads and the crude incidence of malaria cases per 1000 individuals in 2010 in the Busia region, Western Kenya.



Notes: (1) Crude incidence was computed by dividing the total number of observed malaria cases by the total number of individuals living in the homesteads evaluated within the Sub-locations (2). No data was available from the two unshaded SLs in the middle of the map (Kagonia and Murumba SLs)

Bayesian Variable selection

As indicated before, variable selection was performed using a Bayesian approach. Tables 5.1 and 5.2 indicate the selected variables and the posterior summaries of their parameters from Subset A and Subset B, respectively. The BMA Analysis of Subset A resulted in 5 meaningful predictors to be retained for BDM. These variables mainly reflect the influence of cattle ownership on malaria risk. Keeping livestock may distract the mosquito and hence may result in reduced biting rate. However, there are mixed reports regarding livestock keeping and malaria incidence. The 5 variables include livestock feeding practices and activities involving the control of internal and external parasites. While prophylactic treatment when ticks are seen was positively associated with malaria risk, drenching to control internal parasites (with unknown medicine) was negatively associated (Table 5.1). Furthermore individuals that do not obtain treatment despite fever and children under the age of 9 are at a higher risk of malaria in this region (see Table 5. 1).

Table 5.1. The variables selected from Subset A using Bayesian Model Averaging.

Predictors	PIP†	Posterior Mean‡	Posterior SD*
a5[1]	1.000	0.365	0.077
cattle_dewormer_drench[2]	0.993	-0.104	0.032
prophylactic_timing_ticksseen[3]	0.992	0.157	0.046
had_fever_treatment_dontseektx[4]	0.904	0.178	0.084
feeding_livestock_onceaweek[5]	0.834	0.153	0.088

Note: The top five predictors have posterior inclusion probabilities greater than 75%. Model inclusion was based on best 5000 models.
†Posterior Inclusion Probabilities.
‡ The mean of regression coefficients averaged over 5000 models.
*Standard Deviation of the mean.
[1]proportion of individuals in the age group 5-9; [2]Control worms in cattle with drench (unknown drug); [3]Prophylactic treatment of cattle when ticks seen; [4]Had fever but didn't seek treatment; [5]Feeding livestock once a week.

Separate analysis of Subset B (i.e., 25 variables containing information about homesteads, cattle and humans from Tremblay et al 2015 [22]) resulted in 7 predictors for malaria risk in the region. These variables are mainly demographic descriptors such as age, and indictors of socioeconomic status including occupation, quality of housing and access to health treatment (see Table 5.2).

Table 5.2. The variables selected from Subset B using Bayesian Model Averaging.

Predictors	PIP†	Posterior Mean‡	Posterior SD*
age5[1]	1.000	0.306	0.068
age10[2]	1.000	0.308	0.075
occupation_none[3]	0.952	0.489	0.191
had_fever_treatment_chemist[4]	0.951	-0.205	0.080
count_floorcement[5]	0.857	-0.048	0.026
recent_illness_abdominalpain[6]	0.838	0.151	0.086
drought_last6months[7]	0.779	0.076	0.050

The seven predictors have PIP of greater than 75%. Model inclusion was based on best 5000 models.

†Posterior Inclusion Probabilities.
‡ The mean of regression coefficients averaged over 5000 models.
*Standard Deviation of the mean.
[1]Number of individuals in the age group 5-9; [2]Number of individuals in the age group 10-15; [3]Occupation- none; [4]Had fever and treated by chemist; [5]Number houses with cement floors; [6]Recent illness- abdominal pain; [7]Experienced drought in the last 6 months.

Model comparison

The goodness-of-fit of the models with regard to adding spatial and non-spatial random effects, and covariate adjustments is shown in Table 5.3. According to the DIC criteria, the model including both spatial random effects and the covariates has the smallest DIC and therefore it fits the data best. We used this model to produce the risk maps for the region. The spatially structured component accounted for 84.2% of the total unexplained variance for Subset A, and 58.7% for Subset B (see $\theta ratio$ in Table 5.3).

Table 5.3. The goodness-of-fit assessment of Bayesian models for estimating the spatial distribution of the risk of *Plasmodium falciparum exposure* in the Busia region, Western Kenya

	Subset A		Subset B	
Model	DIC[1]	pD[2]	DIC	pD
intercept	581.1	1.0	611.6	1.01
intercept + X_{ij}	504.1	6.0	563.1	8.9
intercept + X_{ij} + θ_i + ψ_i	481.7	36.1	562.1	17.4
$\theta ratio$[3]	84.2%		58.7%	

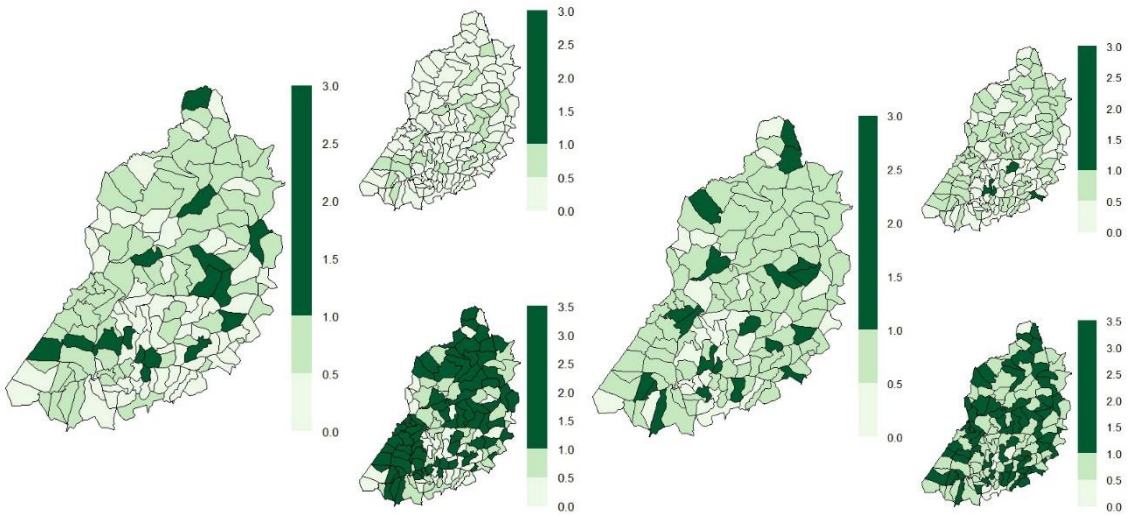
Intercept: baseline; θ: spatially structured random effect; ψ: spatially unstructured random effect;
[1]Deviance Information Criteria;
[2]Effective number of parameters;
[3] The proportion of variance explained by the spatial component (θ): is the variance explained by the spatially structured component divided by the sum of the variances contributed by both spatially structured and unstructured components.
X: covariates; i=1,2,..I (I= number of Sub-locations); j=1,2..J (J=number of covariates)

Mapping risk

The spatial distribution of the risk of *Plasmodium falciparum* exposure in the Busia region is shown in Figure 5.3. We produced separate maps for the two datasets. Sub-locations with darker hues have a relatively higher risk of malaria than those shown in lighter hues compared to the "baseline risk" of 38%. In both maps, many of the Northern and South Western SLs were found to have higher risk of malaria.

The relative risk and 95% Credibility Intervals (CI) adjusted for the covariates and spatial random effects range from 0.19 (0.07, 0.35) to 2 (0.96, 3.38) in Kokeyo and Umla SLs, for Subset A, and 0.16 (0.10, 0.26) to 1.90 (0.79, 3.82) in Yenga and Siranga SLs, for Subset B, respectively. In addition, maps for the lower and upper 95% CI estimates are also displayed to show the uncertainty of the predictions (Figure 5.3). The SLs are shaded in the same way for the ease of comparing maps. As noticed in Figure 5.3, some SLs in the Central, Northern and South Western regions show a relatively higher risk for malaria. However, there is not much variation in Bayesian relative risk estimates between most of the SLs.

Figure 5.3. The predicted risk map of *Plasmodium falciparum* exposure in the Busia region, Western Kenya, estimated by a Bayesian model. Left panel: Subset A, Right panel: Subset B. The risk estimates in the map are based on the median (left), 2.5% (top) and 97.5% (bottom) percentiles of the posterior distribution.



Notes: (1)Subset A has five predictors (*proportion of individuals in the age group 5-9, Control worms in cattle with drench (unknown drug); Prophylactic treatment of cattle when ticks seen, Had fever but didn't seek treatment* and *Feeding livestock once a week*); wheras Subset B has seven predictors (*Number of individuals in the age group 5-9,Number of individuals in the age group 10-15, Occupation- none, Had fever and treated by chemist, Number houses with cement floors, Recent illness- abdominal pain and Experienced drought in the last 6 months*).
(2) The expected rate is computed from the reported prevalence of malaria in the region (38%)

In both Subsets A and B, variances explained by the spatially structured component was higher than non-spatial factors.

5.4 Discussion

Advances in the control and prevention of endemic diseases such as malaria require the identification of most meaningful predictors and stratification of regions based on disease risk of malaria for the Busia region, Western Kenya. The resulting risk maps help to prioritize high risk areas and aid in the efficient implementation of prevention and control measures by health

authorities. The current study presents a Bayesian framework for geostatistical analysis of the PAZ dataset.

The variable selection process using the BMA approach in this report can be viewed either as an independent general approach for risk factor analysis, or as a strategy for selection of a meaningful subset of predictors for specific mapping of malaria risk. The subsets of variables selected from both datasets in this study represent a wide range of risk factors including socioeconomic indicators (housing floor type, unemployment, access to treatment), demographic factors (age), cattle management practices (feeding frequency, internal and external parasite control practices) and history of drought. The direction of association found between socioeconomic indicators (having a cement floor housing – negative, not seeking treatment and unemployment - positive) confirmed earlier reports [42–45]. Good housing condition prevents the exposure to mosquitos and reduces the incidence of malaria while difficulty in access to health care increases the risk of malaria prevalence in a region. A strong positive association between the risk of malaria and history of drought in a region and the presence of high proportion of younger age groups has also been reported before [46]. Other selected predictors of malaria risk in this report such as recent abdominal illness (positive association) and the history of current medication (negative association) are also in agreement with other studies [47, 48]. A negative association of malaria risk and individuals currently taking medication in this study can be a proxy to health treatment access and therefore a low risk of malaria. However, other studies have found that treatment seeking behavior can be associated with both accessibility and the severity of malarial disease [49].

We chose the Bayesian disease mapping approach over the other spatial analysis methods because the BDM accounts better for spatial correlation. This approach assigns pertinent prior information to the variability of disease rates in the overall study region in addition to the observed rates in each SL. This will smooth disease rates towards the overall rate whenever few or no malaria case are reported in selected regions.

In both datasets, inclusion of the selected variables into the Bayesian models significantly reduced the DIC (a reduction by 78 and 49, for Subsets A and B, respectively) as compared to the model containing only the intercept (see Table 5.3). This indicates that the variables indeed explain a significant amount of variance in the data. In addition, incorporating spatial random effect terms into the models further improved the fit, showing that there is a need for 'borrowing strength' from other regions and without the spatial effects the risk is modeled less efficiently. Particularly, the DIC for the model with spatially structured random effect was larger compared to the non-spatial effect models for both datasets. This indicates that local rather than global shrinkage is more important to achieve good model fit. Several reports have shown that the incorporation of spatial and non-spatial terms into a disease mapping regression model improved prediction of risk [15, 50]. Larger variance attributed to the spatial components in Subset A compared to Subset B can be due to differences of the two datasets in terms of the number of homesteads and location for each dataset. Subset A included variables from about half the homesteads of Subset B although Subset A has more diverse variables. Since Subset A has fewer homesteads and hence less spatial coverage, a significant amount of unmeasured factors attributable to variation in risk may be more spatially related compared to Subset B. These spatial effects can be considered as proxy measures for unobserved indicators of risk such as climatic and environmental factors. Updating maps

should be carried out regularly as new information arrives and this is well suited to the Bayesian Disease Mapping approach described in this study.

5.5 Conclusions

The results of this study confirms reported subsets of predictors for malaria and identifies high risk areas for decision makers to prioritize resources for strategic control and prevention of malaria in the Busia region. The risk maps provide a visual representation of estimates of risk at Sub-locations of the Busia region. Although the risk maps are informative regarding variation in risk of *Plasmodium falciparum* exposure in the region, interpretation of the maps should be done with care. Depending on the type and umber of predictors used for constructing maps, variation in risk distribution may be observed, as seen by the differences noticed between the two risk maps. Although a representative subset of homesteads are evaluated in the SLs, there is always a risk that information has been omitted. In addition, the estimated risk is a function of the reported prevalence of malaria in the region and therefore relies on the accurateness of those baseline reports. More data covering a wider geographic area and including environmental and climatic factors can improve the prediction of *Plasmodium falciparum* exposure when structured per SLs.

List of abbreviations

AIC: Akaike Information Criteria; BDM: Bayesian Disease Mapping; BMA: Bayesian Model Averaging; BMS : Bayesian Model Sampling; BMS: Bayesian Model Sampling and Averaging; CAR: Conditional Autoregressive; DIC: Deviance Information Criteria; IRB: Institutional Review Board; KEMRI: Kenya Medical Research Institute; MARA: Mapping Malaria Risk in Africa;

NIH: National Institute of Health; PAZ: Peoples Animals and their Zoonoses; PIP: Posterior Inclusion Probabilities; SD: Standard Deviation; SL : Sub-locations; WinBUGS: Windows for Bayesian Inference Using Gibbs Sampling.

Acknowledgments

This work was supported by the NIH Ruth L. Kirschstein National Research Service Award Institutional Training Grant T32 RR023916 and T32 OD010423. The PAZ project was supported by the Wellcome Trust (E.M.F., grant number 085 308).

Conflict of interest

None

5.6 References

1. WHO | Malaria [http://www.who.int/mediacentre/factsheets/fs094/en/]

2. Kenya Malaria Operational Plan-2014 President's Malaria Initiative.

3. Snow RW, Omumbo JA: Malaria. In Disease and Mortality in Sub-Saharan Africa. 2nd edition. Edited by Jamison DT, Feachem RG, Makgoba MW, Bos ER, Baingana FK, Hofman KJ, Rogo KO. Washington (DC): World Bank; 2006.

4. Ernst KC, Adoka SO, Kowuor DO, Wilson ML, John CC: Malaria hotspot areas in a highland Kenya site are consistent in epidemic and non-epidemic years and are associated with ecological factors. Malar J 2006, 5:78.

5. Kenya Malaria Indicator Survey : Kenya National Bureau of Statistics. 2011.

6. Okach DO, Ayisi JG, Onyango R: Severe malaria in western Kenya: Analysis of hospital records to determine the influence of transmission level on clinical presentation. Sky Journal of Medicine and Medical Sciences 2014, 29:73–78.

7. O'Meara WP, Mangeni JN, Steketee R, Greenwood B: Changes in the burden of malaria in sub-Saharan Africa. Lancet Infect Dis 2010, 10:545–555.

8. Best N, Richardson S, Thomson A: A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 2005, 14:35–59.

9. Gilks WR, Richardson S, Spiegelhalter D: Markov Chain Monte Carlo in Practice. CRC Press; 1995.

10. Lawson AB: Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Second Edition. 2 edition. Boca Raton: Chapman and Hall/CRC; 2013.

11. Bernardinelli L, Clayton D, Montomoli C: Bayesian estimates of disease maps: How important are priors? Stat Med 1995, 14:2411–2431.

12. Bernardinelli L, Montomoli C: Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. Stat Med 1992, 11:983–1007.

13. Richardson S, Thomson A, Best N, Elliott P: Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies. Environ Health Perspect 2004, 112:1016–1025.

14. Waller LA, Carlin BP, Xia H, Gelfand AE: Hierarchical Spatio-Temporal Mapping of Disease Rates. J Am Stat Assoc 1997, 92:607–617.

15. Clements ACA, Lwambo NJS, Blair L, Nyandindi U, Kaatano G, Kinung'hi S, Webster JP, Fenwick A, Brooker S: Bayesian spatial analysis and disease mapping: tools to enhance planning and implementation of a schistosomiasis control programme in Tanzania. Trop Med Int Health TM IH 2006, 11:490–503.

16. Zacarias OP, Andersson M, others: Mapping malaria incidence distribution that accounts for environmental factors in Maputo Province-Mozambique. Malar J 2010, 9:79.

17. Snow RW, Marsh K, Le Sueur D: The need for maps of transmission intensity to guide malaria control in Africa. Parasitol Today 1996, 12:455–456.

18. Towards an Atlas of Malaria Risk in Africa: First Technical Report of the MARA /ARMA Collaboration. 1998.

19. Busia County Integrated Development Plan (2013 -2017) –CIDP
[http://www.busiacounty.go.ke/wp-content/uploads/2015/05/CIDP-Final-Draft.pdf]

20. People, Animals and their Zoonoses (PAZ) in Kenya - Zoonotic and Emerging Diseases [http://www.zoonotic-diseases.org/home/research/paz]

21. WHO | Basic laboratory methods in medical parasitology (archived) [http://www.who.int/malaria/publications/atoz/9241544104_part1/en/]

22. Tremblay M, Dahm JS, Wamae CN, DE Glanville WA, Fèvre EM, Döpfer D: Shrinking a large dataset to identify variables associated with increased risk of Plasmodium falciparum infection in Western Kenya. Epidemiol Infect 2015:1–8.

23. Liu G: Alternative Methods for Variable Selection in Generalized Linear Models with Binary Outcomes and Incomplete Covariates. ProQuest; 2007.

24. Hoeting JA, Madigan D, Raftery AE, Volinsky CT: Bayesian model averaging: a tutorial. Stat Sci 1999:382–401.

25. Viallefont V, Raftery AE, Richardson S: Variable selection and Bayesian model averaging in case-control studies. Stat Med 2001, 20:3215–3230.

26. O'Hara RB, Sillanpää MJ: A review of Bayesian variable selection methods: what, how and which. Bayesian Anal 2009, 4:85–117.

27. Brown PJ, Vannucci M, Fearn T: Bayes model averaging with selection of regressors. J R Stat Soc Ser B Stat Methodol 2002, 64:519–536.

28. Fernández C, Ley E, Steel MFJ: Model uncertainty in cross-country growth regressions. J Appl Econom 2001, 16:563–576.

29. Zeugner S: Bayesian Model Averaging with BMS. mimeo, Available at http://cran. rproject. org/web/packages/BMS/vignettes/bms. pdf; 2011.

30. Richardson S, Abellan JJ, Best N: Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). Stat Methods Med Res 2006, 15:385–407.

31. Clayton D, Kaldor J: Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987, 43:671–681.

32. Marshall RJ: Mapping disease and mortality rates using empirical Bayes estimators. J R Stat Soc Ser C Appl Stat 1991, 40:283–294.

33. Handbook of Statistics: Bayesian Thinking, Modeling and Computation. Gulf Professional Publishing; 2005.

34. Besag J, York J, Mollié A: Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math 1991, 43:1–20.

35. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB: Bayesian Data Analysis, Third Edition. CRC Press; 2013.

36. Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D: The BUGS Book: A Practical Introduction to Bayesian Analysis. CRC Press; 2012.

37. Downing A, Forman D, Gilthorpe MS, Edwards KL, Manda SO: Joint disease mapping using six cancers in the Yorkshire region of England. Int J Health Geogr 2008, 7:41.

38. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A: Bayesian measures of model complexity and fit. J R Stat Soc Ser B Stat Methodol 2002, 64:583–639.

39. Best N, Richardson S, Thomson A: A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 2005, 14:35–59.

40. R: The R Project for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria [http://www.r-project.org/]

41. Sturtz S, Ligges U, Gelman AE: R2WinBUGS: a package for running WinBUGS from R. J Stat Softw 2005, 12:1–16.

42. Gamage-Mendis AC, Carter R, Mendis C, De Zoysa AP, Herath PR, Mendis KN: Clustering of malaria infections within an endemic population: risk of malaria associated with the type of housing construction. Am J Trop Med Hyg 1991, 45:77–85.

43. Konradsen F, Amerasinghe P, van der Hoek W, Amerasinghe F, Perera D, Piyaratne M: Strong association between house characteristics and malaria vectors in Sri Lanka. Am J Trop Med Hyg 2003, 68:177–181.

44. UNDP (Roll Back Malaria Partnership): Multisectoral Action Framework for Malaria. 2013.

45. Ayele DG, Zewotir TT, Mwambi HG: Prevalence and risk factors of malaria in Ethiopia. Malar J 2012, 11:10–1186.

46. Mawili-Mboumba DP, Bouyou Akotet MK, Kendjo E, Nzamba J, Medang MO, Mbina JR, Kombila M: Increase in malaria prevalence and age of at risk population in different areas of Gabon. Malar J 2013, 12:3.

47. Sarma PS, Kumar RS: Abdominal pain in a patient with falciparum malaria. Postgrad Med J 1998, 74:425–427.

48. World Health Organization: Management of Severe Malaria: A Practical Handbook. Geneva: World Health Organization; 2012.

49. Müller O, Traoré C, Becher H, Kouyaté B: Malaria morbidity, treatment-seeking behavior, and mortality in a cohort of young children in rural Burkina Faso. Trop Med Int Health TM IH 2003, 8:290–296.

50. Waller LA, Carlin BP: Disease mapping. Chapman Hall CRC Handb Mod Stat Methods 2010, 2010:217–243.

Chapter 6: Conclusions, Implications and Future Direction

6.1 Introduction

The three objectives of this thesis were to explore (1) the common problem of missing data, (2) variable selection (multidimensionality problem or data shrinkage), and (3) spatial and temporal analysis and representation of disease risk using real word datasets about global health topics. The general aim of the study was to explore quantitative methods for deriving best inferences from imperfect datasets. The thesis represents the collection and application of statistical tools to be applied to imperfect data sets. In the first section of Chapter 1 (1.1), the background and context of the problem as well as the scope of the study is described. The second section of Chapter 1 (1.2) represents a literature review of the statistical analysis of imperfect data and outlined selected statistical tools for remediating the problems encountered when analyzing imperfect data along with their implications and shortcomings. Chapters 2, 3, 4 and 5 are self-contained chapters consisting of their own introduction, methods, results and relevant conclusions. These latter chapters present methods to apply to real world datasets that are imperfect and they result in recommendations for improved statistical inferences. This final Chapter 6 of the thesis summarizes the findings, discusses limitations of the study and suggests future directions for research.

6.2 Specific Findings

This study used three real world global health imperfect datasets from developed and developing countries and each one of the three datasets had at least one of the problems to be addressed by the three aims of the thesis. The following section describes the specific findings of the study.

1). Chapter 2 concentrated on working with missing data and variable selection. For this project, a number of predictors obtained from publicly available databases and mortality rates from the WHO vital registration database were used. The objective was to identify major predictors of mortality associated with foodborne disease and to develop a model for predicting missing mortality rates of countries lacking the data. The study identified eight major predictors of mortality rates and developed four data driven clusters of WHO countries based on the selected predictors (see Table 2.1 and Figure 2.2). A Bayesian hierarchical model was fitted to the data and produced predictions for mortality rates for countries lacking the data. The mortality rates predicted in this study can be used to conduct global burden of foodborne disease studies. This study also demonstrated how mortality rates from a subset of countries (units) with missing values can be estimated from well-defined clusters of countries using Bayesian hierarchical models.

2). In Chapter 3, the general aim of the study was to find ways to fill in national disease notification data gaps for New Zealand travelers. For this specific project, 11 years of the campylobacteriosis notification database from New Zealand were used. The main challenge of the dataset was the lack of information regarding the international travel status of notified campylobacteriosis cases so as to identify these cases as acquired domestically or from overseas travel. Covariate driven Bayesian models were developed and the distribution of the probability of overseas travel was predicted for the cases missing the data. The analysis prioritizes areas in New Zealand that require further attention regarding *Campylobacter* risk from global travel. Estimating the proportion of travel associated *Campylobacter* cases is particularly important for regions with a high rate of incomplete data. In addition to helping generate new hypothesis, the results of this study assist in designing

strategic interventions for the control and prevention of travel associated campylobacteriosis in New Zealand.

3). Chapter 4 presents a Bayesian disease mapping approach to generate spatial and temporal maps displaying the travel associated campylobacteriosis risk in New Zealand. The District Health Boards (DHB), which are geographically defined regions of New Zealand used as units for displaying the risk. Risk maps of DHBs of New Zealand identifying high and low risk areas have been developed. Major urban areas of New Zealand were found to have higher rates of travel associated cases compared to other regions in the country. In addition, a *Campylobacter* risk map for the main global travel destinations was generated. The disease mapping study for New Zealand demonstrates how the hierarchical structure of the Bayesian framework is used to estimate the disease risk for areas with limited data by 'borrowing strength' from neighboring areas. Visual representation of disease risk through maps visualizes the underlying geographical patterns of disease, identifies hot spots and such visualizations are hypotheses generating with regards to causes of disease.

4). Finally, Chapter 5 shows an approach towards the selection of a subset of predictors for malaria prevalence and mapping of malaria risk in the Busia region of Western Kenya. Using a Bayesian Model Averaging approach, 5 and 7 meaningful predictors were identified from two groups of variables (consisting of 22 and 25 variables each as shown in Table 5.1 and Table 5.2). Malaria risk maps for the Busia region were developed. The results of this study provide a visual representation of the variation in the malaria risk in the region. This aids in prioritizing health resources aimed at the strategic control and prevention of malaria exposure in the region. However,

any visualization of the maps must consider the uncertainty around the risk estimates to avoid over-interpretation and bias.

6.3 Implications

In addition to the specific results, the thesis more generally represents the collection and application of statistical tools applied to imperfect data analysis. It focuses on Bayesian frameworks for statistical inference and spatial analysis resulting in tools and resources for future studies. The implications of the study are general in that the approaches can be applied to other diseases and health topics to improve the allocation of health resources for the control and prevention of diseases. The main implications of the findings of the current study are:

- The study assembled a unique collection of methods and tools that can be applied to imperfect datasets for producing valid inference.
- Special emphasis is on Bayesian approaches, which are very applicable for the statistical analysis of imperfect datasets. No single method accommodates all three data challenges such as the imputation of missing values, shrinkage of number of variables and risk mapping, but the collection of tools presented can be used in parallel or a series of combination of methods for generating statistical inferences.
- The approaches toward identifying meaningful small subsets of predictors for disease risk guide the design of future studies. Therefore the approaches help to save time and resources.
- The systematic approaches presented can be applied to other global and public health topics of interest.

- The predicted mortality rates give important clues regarding the burden of mortality associated with foodborne disease around the world.
- Campylobacteriosis control and prevention efforts in New Zealand benefit from the findings of this study because of the stratification of regions based on the magnitude of travel related campylobacteriosis. Such stratification can guide intervention efforts. In addition, the high variation in the proportion of travel associated cases among the DHBs requires further study, for example by collecting more demographic data.
- Malaria control and prevention efforts in the Busia region benefit from the results of this study for intervention efforts in high risk areas.
- The results have implications for policy makers. The control and prevention of endemic and foodborne disease relies on a good and up-to-date knowledge about the epidemiology of disease. This requires the proper identification of the spatial and temporal distribution of the diseases and their major risk factors. Strategic allocation of resources to the control and prevention efforts are the consequence. For example, more public health resources should be directed to high risk areas shown in the risk maps of this study for prioritizing preventive measures.

6.4 Limitation and Future research direction

As indicated in the literature review section of the thesis (Chapter 1 Section 1.2), missing data analysis requires assumptions such as the “missing at random” assumption (MAR). In this study, the MAR assumption was checked through statistical testing alone. However, the exact reason for the mechanism of missingness is unknown and requires further investigation. The results of data

analysis and inferences on the other hand can be significantly flawed if the MAR assumption is violated.

Validation of the models was performed by internal validation only, that is by leave-one-out cross validation in most cases. However, internal validation may not be sufficient and the results and conclusions of the data analysis are strengthened if external and independent data were available to fully evaluate the predictive performance of the models.

Some of the geo-referenced datasets (e.g., number of homesteads in Chapter 5) represent only a fraction of the total number of homesteads in the Sub-locations of the Busia region. Therefore the results may not represent the full picture of the malaria risk in the region. More data covering a wider geographic area and including environmental, demographic and climatic factors can improve the prediction of malaria risk. In addition, the estimated disease risk is a function of the expected background risk (e.g., prevalence of malaria in Chapter 5 and proportion of observed travel associated campylobacteriosis cases in Chapter 3). The results, therefore, rely on the accurateness of the background risks. Finally, analysis of geo-referenced data frequently depends on aggregate measures per units of observation (e.g., demographic, socioeconomic factors etc. summarized for the geographic unit). Individual level inference from such analysis is prone to the so-called 'ecological fallacy'[1]. This type of bias occurs when the degree and magnitude of association between the risk factors and the outcome differ between the levels of the units of observations.

While acknowledging the above limitations, the methods employed to predict missing data, to reduce excess variables, and to depict geographic representation of spatial and temporal risk are quite robust. Future research direction in the area of imperfect data analysis should improve

measures to test validity and reproducibility of the research. This can be achieved by obtaining external data from other regions with comparable disease risks, or through simulation studies. Missing data are difficult to avoid, however, efforts have to be made to minimize their occurrence and non-randomness during the data collection phase of a research.

Systematic data collection is advisable for the areas of New Zealand with unusually high rate of missing travel information. Such efforts will result in new hypothesis as to why particularly the urban areas are at a higher risk from travel acquired campylobacteriosis. Missing data analysis should identify type of missingness in the data, particularly when information exists about the mechanism by which missing values was generated. In the Bayesian context, this includes formulating informative priors obtained from expert opinions or from pilot studies for the models.

6.5 Conclusions

Despite careful planning, study design, and extra precaution while collecting data, it is often impossible to completely avoid missing values and lack of statistical power due to the imbalance between multiple predictors and limited numbers of observations. Data quality challenges are always to be reckoned with when statistical inferences are drawn from datasets. Many data augmentation techniques to remedy missing values have been suggested in literature, each of which with their own advantages and disadvantages. This thesis research assembled and demonstrated several methods for handling imperfect data. The application of Bayesian modelling in the current study with regard to missing data estimation, variable selection and disease mapping has conferred a considerable advantage over traditional, frequentist modelling approaches. Given

the growing availability of computational tools in recent years, this directly recommends to use Bayesian frameworks for analyzing imperfect data. This thesis research showed that the Bayesian methods along with other quantitative tools aid in deriving valid inferences from imperfect data. Since these methods are based on probabilistic sampling from a distribution, uncertainties around parameter estimates (i.e., 95% Credible Intervals) should always be reported together with model specification. While Bayesian frameworks seem superior for analyzing imperfect data, it should be noted that specifying the prior information should always be taken with great care as the method can be substantially affected by the choice of priors.

In conclusion, incorporating Bayesian statistical tools and other quantitative methods, in light of the stepwise applications suggested in this thesis research, into a decision making process will enable policy makers to make efficient use of available resources. The communication between development and application of Bayesian models for imperfect data and decision makers has to be undertaken with great care to optimize inference for health policy and improve health outcomes.

Appendices

Appendix 1.1 Useful statistical programs and resources.

A number of statistical programs were used either in parallel or independently to perform data cleaning, preparation and analysis. Below are short descriptions, versions and resources for some of them. The following software was used:

Preparation of shape files

- **R** statistical program (version 3.1.3): R is a freely available software program for performing a number of statistical computing and graphics [3]. R is highly flexible programming language through the use of user-submitted "packages" for executing specific functions or specific areas of study. There are over 5800 "packages" available in various repositories (such as the Comprehensive R Archive Network (CRAN), Biconductor, GitHub etc.) and it is increasing over time. Packages are typically installed to the computer and accessed through "library". R can be freely downloaded from: http://www.r-project.org/

- **Maps2WinBUGS**: available as a standalone program or a QGIS plugin to facilitate data processing for Bayesian spatial modelling [2]. It converts shape files to appropriate formats export to GeoBUGS. The map format used by GeoBUGS differs from the standard formats used in geographical information systems (GIS). maps2WinBUGS, helps the user prepare maps data for use in GeoBUGS. With this tool, one can obtain adjacency lists, convert maps, and merge back the results of model runs with a source map in QGIS. It can be freely downloaded from:

  http://sourceforge.net/projects/maps2winbugs/

Data Analysis

- **R** statistical program (version 3.1.3): R is a freely available software program for performing a number of statistical computing and graphics [3]. R is highly flexible and extensible program particularly through the use of user-submitted "packages" for executing specific functions or specific areas of study. There are over 5800 "packages" available in various repositories (such as the Comprehensive R Archive Network (CRAN), Biconductor, GitHub etc.) and it is increasing over time. Packages are typically installed to the computer and accessed through "library". R can be freely downloaded from: http://www.r-project.org/

- **JAGS**: stands for **J**ust **A**nother **G**ibbs **S**ampler, is a tool for analysis of Bayesian models using Marcov Chain Monte Carlo simulation [4]. Unlike BUGS, JAGS runs in all commonly used platforms (MacOS, Linux, and Windows). It can be called within the R environment through *rjags* package [5]. Currently JAGS does not have a functionality to perform spatial analysis. JAGS can be freely downloaded from: http://sourceforge.net/projects/mcmc-jags/files/

- **INLA** (Integrated **N**ested **L**aplace **A**pproximation): INLA is a less computationally expensive alternative to the Marcov Chain Monte Carlo method to estimate the posterior distribution of a Bayesian model [6]. It returns a similar result with commonly used software like BUGS and JAGS with significantly less amount of time. For analysis of spatial data using the INLA package, additional tools are required. These include setting up a neighborhood adjacency matrix for use with conditional autoregressive (CAR) models and to map results. There are several ways to import and process shape files in R. For

example, The R package "***maptools***" is essential in this case. The "***readShapePoly()***" function from this package will read in shapefiles into the R environment, and the ***spdep*** package provides "***poly2nb()***" followed by ***nb2INLA()*** functions to create the adjacency matrix neighbor structures for use with a CAR model in the INLA format. Mapping the results is performed by either the ***spplot*** (from ***sp*** package) or ***ggplot*** (from ***ggplot2*** package) functions. INLA can be freely downloaded from: http://www.r-inla.org/download

- **WinBUGS** 1.4: WinBUGS (**W**indows for **B**ayesian Inference **U**sing **G**ibbs **S**ampling) is a freely available program which is used to fit a complex statistical models using Markov chain Monte Carlo (MCMC) methods [7]. The program produces posterior distributions from which estimates, standard deviations, Credible Intervals as well as monitoring and convergence diagnostics plots are produced. WinBUGS can be freely downloaded from: http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/

1. QGIS Geographic Information System. Open Source Geospatial Foundation Project [qgis.osgeo.org]
2. Solymosi N, Wagner SE, Maróti-Agóts Á, Allepuz A: maps2WinBUGS: a QGIS plugin to facilitate data processing for Bayesian spatial modeling. *Ecography* 2010, 33:1093–1096.
3. R: The R Project for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria [http://www.r-project.org/]
4. Plummer M, others: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*. *Volume 124*. Technische Universit at Wien; 2003:125.
5. Plummer M, Stukalov A, Denwood M, Plummer MM: Package "rjags." *update* 2015, 16:1.
6. Lindgren F, Rue H: Bayesian Spatial Modelling with R-INLA. .
7. Lunn DJ, Thomas A, Best N, Spiegelhalter D: WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput* 2000, 10:325–337.

Appendix 2.1. A JAGS code for the Bayesian Hierarchical Model for predicting missing mortality rates associated with foodborne diseases.

The JAGS model for performing regression analysis of mortality associated with foodborne diseases is provided below

```
modelstring = "
model{
for (i in 1:n){
y[i] ~ dnorm(mu[i], sigma.tot)
     mu[i] <- b.0[Cluster[i]]+ b[1]*x1[i]+ b[2]*x2[i]+ b[3]*x3[i]+
            b[4]*x4[i]+ b[5]*x5[i] + b[6]*x6[i]+ b[7]*x7[i]+ b[8]*x8[i]
}

# Random effects for each Cluster (J=4)
for (j in 1:J){
b.0[j] ~ dnorm(mu.j, tau.u)
}

# Prior for regression coefficients
for (a in 1:8){
b[a] ~ dnorm(0, 0.01)
}

# Prior for the mean of the random effects
    mu.j~dnorm(0, 0.01)

# Prior for precisions
    tau.u <- pow(sigma.u,-2)
    sigma.u ~ dunif(0,10)
    tau.tot <- pow(sigma.tot,-2)
    sigma.tot ~ dunif(0,10)

# Compute and extract predicted values
for (i in 1:n) {
Predictions[i] <- mu[i]                # Predicted values

}
"
writeLines(modelstring,con="model.txt")
modelCheck( "model.txt" )          # Check the model

# Bundle data:
data.list <- list(
  Cluster =data$Cluster,
  J=length(unique(data$Cluster)),
```

```
  n=nrow(data),
  y=(data$y),
  x1=(data$lifeexpectancy),
  x2=log(data$animalcalpercap),
  x3=log(data$birthperadolescent),
  x4=( data$pctareableland),
  x5=log(data$fertilityrate),
  x6=log(data$maternaldeathrisk),
  x7=( data$laborfemmale),
  x8=log(data$kcalperday))

# Parameters to monitor:
params= c("b.0", "b", "sigma.tot","predictions" )

# Gibbs sampling
jags.model <- jags(data=data.list,
                   inits=NULL,
                   parameters.to.save=params,
                   model.file="model.txt",
                   n.thin=5,
                   n.chains=2,
                   n.burnin=5000,
                   n.iter=50000)
```

Appendix 2.2. The predictions of log-total mortality rates and associated 95% Credible Intervals for all WHO countries using Bayesian hierarchical model.

	Country	Log-total mortality rates			
		Observed	Predicted		
			Median	95% CI	
				lower	upper
Cluster 1	Afghanistan	-	5.00	2.65	7.36
	Albania	-	0.53	0.08	0.98
	Algeria	-	2.11	0.76	3.48
	Andorra	-	1.35	-0.25	2.96
	Antigua and Barbuda	-	3.65	2.14	5.13
	Argentina	0.29*	1.38	0.80	1.95
	Armenia	-	0.61	0.25	0.97
	Australia	-1.31*	-0.15	-0.59	0.29
	Austria	-0.89	-0.97	-1.49	-0.45
	Azerbaijan	-	1.31	0.46	2.15
	Bahamas	-	0.83	0.17	1.48
	Bahrain	-	1.42	0.60	2.22
	Barbados	-	0.22	-0.45	0.88
	Belarus	-	-0.54	-1.14	0.05
	Belgium	-	-0.36	-0.82	0.09
	Belize	-	2.24	1.74	2.75
	Bhutan	-	2.54	1.65	3.44
	Bolivia	-	3.26	2.42	4.08
	Bosnia and Herzegovina	-	-0.76	-1.51	-0.01
	Botswana	-	2.83	1.60	4.05
	Brazil	1.44	1.19	0.73	1.65

Cluster 1	Brunei Darussalam	-	0.97	0.54	1.39
	Bulgaria	-0.65	-0.46	-1.10	0.17
	Canada	0.17	-0.34	-0.88	0.24
	Cape Verde	-	2.25	1.78	2.71
	Chile	0.61	0.70	0.09	1.31
	China	-	0.61	-0.14	1.38
	Colombia	1.36	1.94	1.40	2.36
	Cook Islands	-	3.53	1.91	5.14
	Costa Rica	1.09	1.22	0.74	1.72
	Croatia	-	-0.33	-0.68	0.02
	Cuba	1.03	0.67	-0.14	1.49
	Cyprus	-	-0.05	-0.53	0.43
	Czech Republic	-	-0.93	-1.40	-0.45
	Denmark	0.29	-0.41	-1.23	0.39
	Djibouti	-	3.51	2.27	4.77
	Dominica	-	3.52	2.14	4.91
	Dominican Republic	-	2.38	1.80	2.94
	Ecuador	2.02	2.22	1.69	2.75
	Egypt	3.05	2.27	1.23	3.29
	El Salvador	2.03	1.86	1.27	2.45
	Equatorial Guinea	-	3.74	2.22	5.26
	Estonia	-	-0.05	-0.50	0.40
	Fiji	-	1.55	0.70	2.38
	Finland	-0.12	-0.30	-0.80	0.21
	France	0.43	-0.13	-0.73	0.47
	Gabon	-	2.99	2.06	3.92
	Georgia	-0.19*	0.59	0.03	1.16
	Germany	-0.37	-0.93	-1.35	-0.40

Cluster 1	Greece	-	-1.25	-2.03	-0.45
	Grenada	-	1.22	0.31	2.12
	Guyana	2.63	2.34	1.69	2.99
	Honduras	-	2.83	2.18	3.46
	Hungary	-1.05	-0.82	-1.49	-0.13
	Iceland	-	-0.13	-0.77	0.53
	Iran	-	0.83	-0.24	1.90
	Iraq	-	3.28	1.66	4.87
	Ireland	-	-0.77	-1.73	0.18
	Israel	0.51	0.69	-0.18	1.56
	Italy	-1.90*	-1.02	-1.61	-0.41
	Jamaica	-	1.74	1.28	2.21
	Japan	0.07	-0.65	-1.28	-0.03
	Jordan	-	2.61	1.37	3.86
	Kazakhstan	-	1.12	0.45	1.80
	Kiribati	-	3.30	2.21	4.39
	Kuwait	-	0.38	-0.52	1.27
	Kyrgyzstan	1.94	1.75	1.26	2.25
	Latvia	-	-0.38	-0.93	0.17
	Lebanon	-	0.85	-0.19	1.89
	Libyan Arab Jamahiriya	-	2.09	0.81	3.36
	Lithuania	-	-0.85	-1.42	-0.27
	Luxembourg	-	-0.12	-0.84	0.59
	Macedonia	0.65	-0.04	-0.76	0.69
	Malaysia	-	1.69	1.03	2.35
	Maldives	-	1.59	0.79	2.40
	Malta	-	-0.59	-1.51	0.34
	Marshall Islands	-	3.83	2.25	5.40

Cluster 1	Mauritania	-	4.03	2.58	5.48
	Mauritius	-	0.62	-0.05	1.31
	Mexico	1.62	1.53	1.00	2.07
	Micronesia (Federated States of)	-	2.79	1.69	3.89
	Monaco	-	3.74	2.39	5.08
	Mongolia	-	1.67	0.82	2.51
	Montenegro	-	-0.10	-0.98	0.78
	Morocco	-	2.34	1.34	3.33
	Namibia	-	3.16	2.21	4.10
	Nauru	-	3.60	2.38	4.81
	Netherlands	-0.71	-0.25	-0.80	0.30
	New Zealand	-	0.48	-0.06	1.04
	Niue	-	3.79	2.39	5.18
	Norway	0.76	0.20	-0.48	0.86
	Oman	-	1.87	0.88	2.87
	Pakistan	-	3.39	1.85	4.90
	Palau	-	2.68	1.60	3.77
	Panama	1.45	2.08	1.55	2.53
	Papua New Guinea	-	3.27	2.10	4.47
	Paraguay	1.83	2.57	1.91	3.11
	Philippines	-	2.52	1.75	3.27
	Poland	-2.21*	-1.22	-1.79	-0.63
	Portugal	-1.83*	-0.64	-1.06	-0.20
	Qatar	-	1.07	0.11	1.99
	Republic of Korea	-0.76	-0.43	-1.46	0.63
	Republic of Moldova	-0.03	-0.14	-0.96	0.68
	Romania	-0.19	-0.25	-0.83	0.35
	Russian Federation	-	0.15	-0.79	1.11

Cluster 1	Saint Kitts and Nevis	-	3.58	2.15	4.98
	Saint Lucia	-	0.94	0.40	1.49
	Saint Vincent and the Grenadines	-	1.34	1.03	1.65
	Samoa	-	3.03	1.98	4.07
	San Marino	-	4.01	2.50	5.51
	Saudi Arabia	-	2.07	0.73	3.41
	Serbia	-0.42	-0.30	-0.79	0.19
	Seychelles	-	2.21	1.25	3.17
	Singapore	-	-0.63	-1.37	0.13
	Slovakia	-	-1.02	-1.67	-0.37
	Slovenia	-	-0.40	-1.10	0.29
	Somalia	-	4.76	3.17	6.41
	South Africa	3.25	2.41	1.34	3.48
	South Sudan	-	3.35	2.22	4.50
	Spain	-0.06	-0.66	-1.26	-0.06
	Sudan	-	4.19	2.81	5.59
	Suriname	-	2.17	1.63	2.72
	Swaziland	-	3.15	1.72	4.60
	Sweden	-0.05	-0.57	-1.18	0.06
	Switzerland	-0.63	-0.68	-1.20	-0.15
	Syrian Arab Republic	-	2.63	1.27	3.94
	Thailand	0.44	0.74	-0.07	1.51
	Tonga	-	2.76	1.64	3.85
	Trinidad and Tobago	1.62*	0.67	0.09	1.26
	Tunisia	-	1.33	0.19	2.47
	Turkey	-	1.09	0.12	2.05
	Turkmenistan	-	1.92	1.13	2.74
	Tuvalu	-	3.53	2.21	4.84

Cluster 1	Ukraine	-	-0.54	-1.25	0.18
	United Arab Emirates	-	0.61	-0.28	1.51
	United Kingdom	0.64	0.20	-0.47	0.85
	United States of America	-0.43*	0.46	0.02	0.91
	Uruguay	3.39*	0.96	0.52	1.39
	Uzbekistan	0.09	1.32	0.69	1.94
	Vanuatu	-	3.06	2.10	4.00
	Venezuela	1.92	2.11	1.63	2.58
	Viet Nam	-	1.36	0.34	2.33
Cluster 2	Angola	-	4.36	-9.80	12.14
	Bangladesh	-	2.37	-11.96	10.15
	Benin	-	3.90	-10.34	11.75
	Burkina Faso	-	3.94	-10.29	11.74
	Cameroon	-	3.84	-10.30	11.64
	Chad	-	4.68	-9.45	12.42
	Congo	-	3.79	-10.41	11.54
	CongoDR	-	4.25	-9.90	12.07
	Cote dIvoire	-	3.70	-10.52	11.43
	Eritrea	-	3.53	-10.75	11.32
	Ethiopia	-	3.95	-10.20	11.80
	Gambia	-	3.75	-10.54	11.55
	Ghana	-	3.42	-11.00	11.21
	Guinea	-	4.23	-9.94	12.04
	Guinea-Bissau	-	4.15	-9.98	11.89
	Haiti	-	3.20	-10.94	10.99
	Lesotho	-	3.07	-10.98	10.81
	Liberia	-	4.43	-9.73	12.30
	Madagascar	-	3.69	-10.66	11.45

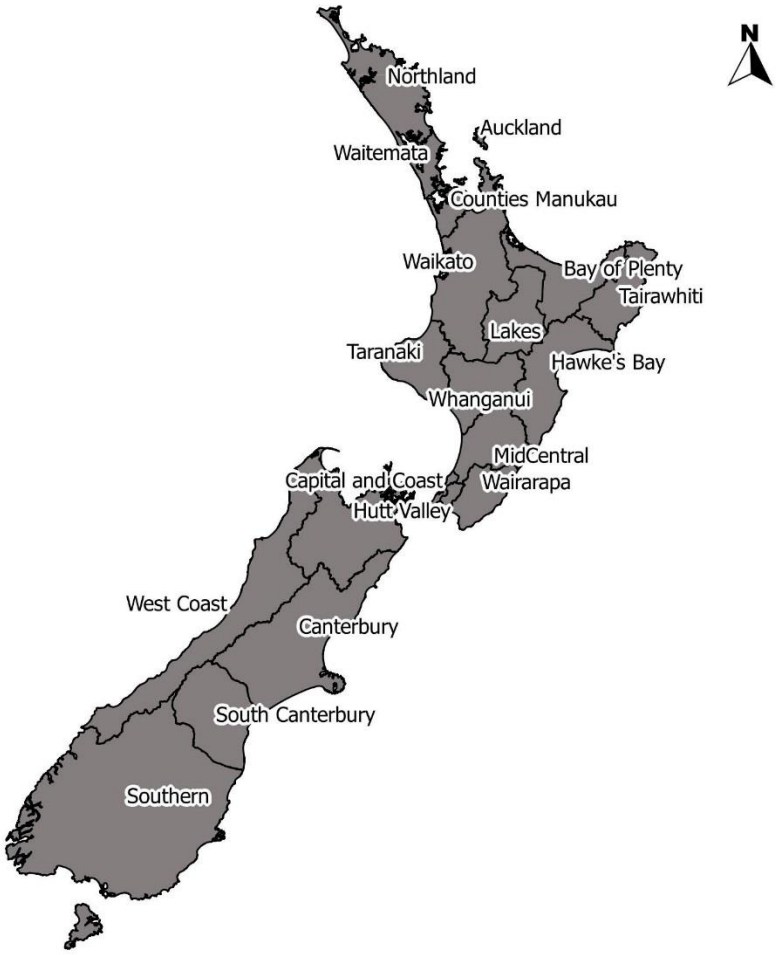
Cluster 2	Malawi	-	4.22	-9.88	12.24
	Mali	-	4.33	-9.92	12.08
	Mozambique	-	3.94	-10.16	11.81
	Nepal	-	2.52	-11.98	10.19
	Niger	-	4.58	-9.82	12.32
	Nigeria	-	3.98	-10.32	11.79
	Sierra Leone	-	4.03	-10.15	11.94
	Uganda	-	3.90	-10.46	11.74
	United Republic of Tanzania	-	3.74	-10.66	11.54
	Zambia	-	4.14	-9.87	12.01
Cluster 3	Burundi	-	3.90	-10.17	11.92
	Rwanda	-	3.85	-10.33	11.74
	Togo	-	3.25	-11.02	11.12
Cluster 4	Cambodia	-	3.20	1.55	5.02
	Central African Republic	-	4.71	2.84	6.66
	Comoros	-	4.37	2.77	6.07
	Guatemala	3.73	3.56	2.27	4.98
	India	-	3.05	1.55	4.69
	Indonesia	-	2.70	1.13	4.44
	Kenya	-	4.30	2.62	6.07
	Korea	-	2.04	-0.30	4.48
	Lao	-	3.77	1.75	5.92
	Myanmar	-	2.36	0.56	4.36
	Nicaragua	-	2.64	1.30	4.13
	Peru	-	2.42	0.94	4.07
	Sao Tome and Principe	-	3.23	1.74	4.81
	Senegal	-	4.42	2.91	6.05
	Solomon Islands	-	3.80	2.30	5.38

Cluster 4	Sri Lanka	-	1.91	0.30	3.62
	Tajikistan	-	2.95	1.31	4.66
	Timor-Leste	-	4.92	3.43	6.49
	Yemen	-	4.52	3.04	6.07
	Zimbabwe	-	3.69	1.73	5.77

Notes: (1) Cluster 1: 142 countries; Cluster 2: 20 countries; Cluster 3: 3 countries; Cluster 4: 29 countries; (2) Eight predictors of mortality associated with food brome disease were used to form the four clusters. These predictors are: *Life expectancy at birth, Average Calorie Supply from Animal Products - per Capita, Adolescent fertility rate(the number of births per 1,000 women ages 15-19), Percent arable land, Total fertility rate, Maternal mortality ratio, Labor force participation rate for females ages 15-24 and Calorie supply per capita per day*

*observed log-total mortality rate outside of the predicted 95% Credible Intervals.

Appendix 3.1. Map of New Zealand showing the District Health Boards



Notes: There are 20 District Health Boards (DHB) in New Zealand which were created after the New Zealand Public Health and Disability Act 2000. They are organizations responsible for ensuring the provision of health and disability services to populations within a defined geographical boundary [1].

References

1. New Zealand Ministry of Health. My DHB [http://www.health.govt.nz/new-zealand-health-system/my-dhb]

Appendix 3.2. JAGS code used for the Bayesian Hierarchical model to estimate travel related campylobacteriosis cases in New Zealand

The JAGS code used for predicting the overseas travel status of campylobacteriosis notifications in New Zealand is presented below. All the unknown parameters in the model were given non-informative prior distributions (i.e., assigned dnorm(0, 0.0001) priors, which implies a normal distribution with mean = 0 and precision =10^{-4}). Two chains of dispersed initial values were ran each with 30,000 iterations and a burn-in of 3,000. Model convergence was assessed by observing the mixing of the two chains over time and visually checking the density plots and autocorrelation.

```
modelstring = "
model {
for(i in 1:N) {
OvseasCat [i] ~ dbern(p[i])     # Likelihood: observed  travel status of individual cases is
                                                    Bernoulli distributed
p[i] <- 1/(1+exp(-(alpha  +
          beta.ur*urban[i] +beta.dep*DepIndex[i]+beta.tr*TrvlRate[i]+beta.age[age[i]]+
          beta.seas[season[i]]+ beta.sex[sex[i]])))   # Predicted values
  }

alpha ~ dnorm(0, 0.0001)        # Uninformative priors on regression coefficients
beta.dep~dnorm(0, 0.0001)
beta.ur~dnorm(0, 0.0001)
beta.tr~dnorm(0, 0.0001)

beta.age[1]<-0                  # Set first categories of factors to 0 (reference categories)
beta.seas[1]<-0
beta.sex[1]<-0

for (b in 2:4)                  # Uninformative exchangeable priors for categorical predictors
   { beta.age[b]~dnorm(0, 0.0001)  }
for (a in 2:4)
  { beta.seas[a]~dnorm(0, 0.0001) }
beta.sex[2]~dnorm(0, 0.0001)
 }
"
writeLines(modelstring,con="model.txt")
```

```
# Check the model
modelCheck( "model.txt" )

# Define data
data <- list(
  N=length(data$OvseasCat),   OvseasCat = data$OvseasCat,    ProU = data$urban,
  DepIn = data$DepInex, TrvlRate = data$TrvR,  age = data$age,
  season = data$season,  sex = data.$sex
  )

# Parameters to monitor
parameters = c("alpha","beta.dep","beta.ur","beta.tr", "beta.age",
                "beta.seas","beta.sex", "p")

set.seed (123)

# MCMC settings
nc <- 2                          #Number of Chains
ni <-  30000                     #Number of draws from posterior
nb <- 3000                       #Number of draws to discard as burn-in
# Gibbs sampling
 jags.fit <-  jags(data=data,  inits=NULL,
                   parameters.to.save=parameters,
                   model.file="model.txt",
                   n.chains=nc,
                   n.burnin=nb,
                   n.iter=ni)
```

Notes: OvseasCat: Overseas travel status (YES, NO, UNKNOWN).

JAGS (Just Another Gibbs Sampler): is a standalone program for simulating from a Bayesian Hierarchical models that takes a model string written in an R-like syntax and that compiles and generated a Monte Carlo Marcov Chain (MCMC) samples from this model using the Gibbs sampling algorithm [1]. The main advantage of *JAGS* over the classical *BUGS* (Bayesian Inference Using Gibbs Sampling e.g., *WinBUGS*) is its platform independence that it can operate in all main operating systems, while *BUGS* is broadly Windows specific. *JAGS* is called and controlled within the R environment through *rjags* package [2]. It parameterizes distributions using precision instead of standard deviation (σ), where the precision $\tau = 1/\sigma^2$. Therefore in the above JAGS code, the standard deviation of the prior distribution becomes:

$0.0001 = 1/100^2$

References

1. Plummer M, others: **JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling**. In *Proceedings of the 3rd international workshop on distributed statistical computing*. *Volume 124*. Technische Universit at Wien; 2003:125.

2. Plummer M, Stukalov A, Denwood M, Plummer MM: **Package "rjags."** *update* 2015, **16**:1.

Appendix 4.1. An R-INLA code for spatio-temporal analysis of travel related campylobacteriosis in New Zealand

The R-INLA code for conducting spatio-temporal analysis of travel associated campylobacteriosis risk in New Zealand is presented below.

```
# Load required libraries

library(spdep)
library(sp)
library(maptools)
library(INLA)


# 1. Prepare spatial data

# 1.1 Load New Zealand DHB shapefile

DHB <- readShapePoly("DHB.shp")

# 1.2 Create adjacency matrix for the DHBs (define neighborhood structure)

adjacency <- poly2nb(DHB)

# 1.3 Convert the adjacency matrix into a file in the INLA format

nb2INLA("DHB.graph", adjacency)

# 1.4 Create region identifier ID

DHB$ID<-1:20

# 2. Prepare long data format (220 rows= 20 regions by 11 time points)

case<-case                          # case counts for each region and time point
exp<-exp                            # expected value for each region and time point
year<-rep(1:11,each=20,len=220)     # "year" is time identifier variable
region<-ID                          # each DHB region is identified by a unique ID
region1<-ID                         # make a duplicate column for region
data <-data.frame(case=case, exp=exp, year=year,    # Inla requires a dataframe
                  region=region,region1=region1)


# 3. Analysis

formula <-case~1+f(region, model="bym",
            adjust.for.con.comp = FALSE, graph="DHB.graph",  #"bym" for specifying both
                                                               spatially  structured and
                                                               unstructured random effects

 hyper = list(prec.unstruct = list(prior="loggamma",
                                   param=c(1,0.0005)),        # Non-informative prior for
              prec.spatial = list(prior="loggamma",             precisions(default)
```

```
                                    param=c(1,0.0005))))+
        f(year, model="iid")                          # Unstructured temporal
                                                      component


# NB: Alternatively it is possible to specify the two "bym" components separately using
# f(ID,model="besag",graph="DHB.graph") for the spatial structured (CAR) [1] and
# f(ID2,model="iid",graph="DHB.graph") for the unstructured component. In this case
# the region ID needs to be duplicated (region=region1) as it is not allowed to define two
# functions on the same variable [2]


inla.output<-inla(formula,
        family="poisson",                  # define family of distribution
        data=data,E=exp,                   # provide dataframe and expected values
        control.compute=list(dic=TRUE,cpo=TRUE), # compute DIC and CPO
        control.inla=list(int.strategy = "grid", diff.logdens = 4))

inla.output$dic$dic                # Extract DIC
inla.output$dic$p.eff              # Extract pD
-mean(log(inla.output$cpo$cpo )) # "leave-one-out" measures of fit (negative mean log of
                                   Conditional Predictive Ordinate)

# Computing the variance explained by each random component on an SD scale
Out <- inla.contrib.sd(inla.output, nsamples=10000)
SD<-Out$hyper                      # Extract the SDs


# Variance explained (variance=SD2)
var.U<-sd[1,1]^2  # structured random effect (spatial)
var.V<-sd[2,1]^2  # unstructured random effect (non-spatial)
var.T<-sd[3,1]^2  # unstructured temporal random effect

var.U.prop=var.U/(var.U+var.V+var.T) # Proportion of variance explained by the spatial component

# Extract mean and 95% Credibility Intervals of predicted values
RR<-as.data.frame(inla.output$summary.fitted.values[c("mean", "0.025quant","0.975quant")])
```

References

1. Besag J, York J, Mollié A: Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 1991, 43:1–20.

2. Blangiardo M, Cameletti M, Baio G, Rue H: Spatial and spatio-temporal models with R-INLA. *Spat Spatio-Temporal Epidemiol* 2013, 4:33–49.

Appendix 5. 1. A WinBUGS code for doing spatial analysis of malaria prevalence in the Busia region, Western Kenya

A model for conducting spatial analysis of malaria prevalence in the Busia region, Western Kenya is provided below.

```
Model
{
# Likelihood

   for (i in 1:N)                                  # N= number of Sub-locations in the study

{
    Cases[i]  ~ dpois(mu[i])                       # Cases are observed counts and assumed Poisson
                                                     distributed with mean mu[i], which is a function of
                                                     the RR[i] and the expected count E[i]

log(mu[i]) <-log(E[i])+ beta0 + beta.x[i]*X[i]+ U[region[i]]+ V[region[i]]

      RR[i]<- exp(beta0 + beta0 + beta.x[i]*X[i]+ U[region[i]]+ V[region[i]])

                                                  # the log of the RR is a linear function of both fixed effects
                                                     (beta0 and X's) and random effects (U and V). Here,
                                                     beta0 denotes the overall log of the RR (i.e., for the
                                                     entire Busia region), U denotes unknown random
                                                     effects that are not spatially correlated, and U denotes
                                                     random effects that are spatially correlated. X's are
                                                     covaraites and beta.x's are the associated regression
                                                     coefficients

   }

# Prior distributions

         # prior on the intercept

   beta0 ~ dflat()                                 # Uninformative prior for the intercept, assumes that the
                                                     risk is same for all Sub-locations

         # prior for the CAR distribution

U[1:N]~car.normal(adj[],weights[],num[],tau.u)     # Conditionally autoregressive (CAR)[2] prior
                                                     distribution for the spatially correlated random
                                                     effect; this specification assumes that the spatial
                                                     random effects for each region[i ] are conditional on
                                                     adjacency for each region[i], and weighted
                                                     according to the strength of the adjacency and with a
                                                     variance of tau.u
```

```
                    # Weights   for the CAR prior distribution

 for(i in 1:sumNumNeigh)

     { weights[i] <- 1                    # Weights for the CAR distribution where the regions are  adjacent
                                            (share same border), the assigned weight is 1, otherwise 0.


            # prior for the unstructured random effect

   for (i in 1:N)
   {
    V[i]~dnorm(0,tau.v)                   # exchangeable normal prior distribution for the uncorrelated
                                            random effect component; it assumes that the non-spatial
                                            random effects for each region[i] are normally distributed
                                            with mean of zero and a precision of tau.v.
                                            Precison =1/variance
   }


          # prior for regression coefficients

beta.x~dnorm(0,0.0001)                    # for all regression coefficients in the model, a normal
                                            distribution with mean 0 and precision 10-3 is assumed.

   # Hyper priors

   tau.u~dgamma(0.5,0.005)

   tau.v~dgamma(0.5,0.005)                # tau.u and tau.v  are precisions (inverse of the variance)  of the  random
                                            effect priors, and assumed to follow the gamma distribution, which
                                            are defined by the scale (a=0.5) and the shape (b =0.0005)
                                            parameters. This is a skewed distribution with mean=a/b and
                                            variance= κ / b2

 # Variance explained by the spatial component

   sigma2.u<-1/tau.u
   sigma2.v<-1/tau.v
   uratio <-  sigma2.u/( sigma2.u+sigma2.v)   # uratio computes the fractional variance  explained by    the
                                                spatial component.
}
```

References

1. Lunn DJ, Thomas A, Best N, Spiegelhalter D: WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput* 2000, 10:325–337.

2. Besag J, York J, Mollié A: Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 1991, 43:1–20.