MODELLING DISTRIBUTION OF UNDER-FIVE CHILD DIARRHOEA ACROSS MALAWI

MSc. (BIOSTATISTICS) THESIS

TSIRIZANI MUMDERANJI MWALIMU KAOMBE

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

NOVEMBER, 2012

MODELLING DISTRIBUTION OF UNDER-FIVE CHILD DIARRHOEA ACROSS MALAWI

MSc. (BIOSTATISTICS)

By

TSIRIZANI MUMDERANJI MWALIMU KAOMBE

BSc. (Mathematical Sciences Education: Maths and Statistics)-University of Malawi

Thesis submitted to the Department of Mathematical Sciences, Faculty of Science, in partial fulfillment of the requirements for the degree of Master of Science (Biostatistics)

UNIVERSITY OF MALAWI

CHANCELLOR COLLEGE

NOVEMBER, 2012

DECLARATION

I, the undersigned, hereby declare that this thesis is my own original work which has not been submitted to any other institution for similar purposes. Where other people's work has been used acknowledgements have been made.

TSIRIZANI MUMDERANJI MWALIMU KAOMBE

Name

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and effort and has been submitted with our approval.

Signature: _____ Date: _____

J. J. NAMANGALE, PhD (Associate Professor)

Supervisor

Signature: _____ Date: _____

J. SIMBEYE, MSc. (Lecturer)

Programme Coordinator

ACKNOWLEDGEMENTS

I am sincerely grateful to my supervisor, Dr. J. J. Namangale, for his untiring and generous advice as well as his accessibility at all times. I must admit that I have benefited and learnt a lot from his wisdom and intelligible guidance at every stage of this work. It has helped me in widening my research abilities as well as my academic writing and presentation skills. Despite his stretched schedule, he made himself available to me whenever I needed his help. That has assisted me also to easily meet deadlines in finalizing this work.

My special thanks should also go to the former and current Head of Mathematical Sciences Department at Chancellor College, Mr. G. Kunyenje and Dr. L. K. Eneya, alongside the former and current Coordinator of Master of Science in Biostatistics Programme, Dr. L. N. Kazembe and Mr. J. Simbeye, for their commitment to ensuring that every candidate works towards set deadlines in the Programme. Their regular and fatherly reminders have assisted me, again, in finishing this work in time.

Through the Coordinator of the Programme, I am also indebted to the Commissioner of Statistics, Mr. C. Machinjili, and the entire staff at Malawi National Statistical Office (NSO) for allowing me access to the 2006 Malawi Multiple Indicator Cluster Survey data which I have used in this study. It is worth noting that the process of collecting nationwide data is not an easy enterprise. It demands a lot of ingenuity and resources for the data to be of high quality; hence, without the assistance of the NSO the data collection stage in this study would have been very challenging to me.

Further, I do not take for granted the contribution made by my lecturers in this work, including those academics who took part in my presentations at different stages of this work for their lucid criticisms and expert comments. It is my lecturers who introduced me to the more fascinating theories in statistics such as the Bayesian data analysis, Generalized Linear Models, and others. My exposure to such materials inspired me to provide a more practical way of looking at statistical theories in the under-five children's health in Malawi than I would have otherwise. Needless to say, it is the comments made by panelists in my presentations that have helped me in shaping and moulding this work.

Likewise, I owe many thanks to the University of Livingstonia Librarian, Dr. Augustine W. C. Msiska for providing me with feedback on the draft of this write-up. I have to commend him for sparing his precious time to offer professional scrutiny to the structure and flow of language used in this work. I concede that his contribution has greatly helped in minimizing the typos, structural and grammatical errors in this work.

Similarly, I appreciate the editing services done by the Dean of Faculty of Education at Laws Campus of the University of Livingstonia, Mrs. Joyce Mlenga. Her contribution has assisted in tracing and reducing the punctuation and other grammatical errors in this write-up.

Finally, I would like to express my appreciation to my family; my wife, Rachel, for her unreserved support and understanding throughout my study, and my children Pythagoras, Vanessa, and Leticia, whose patience I treasure.

ABSTRACT

Analysis of diarrhoea data in Malawi has been commonly done using classical methods. However, of late new approaches, such as Bayesian methods, have been introduced in literature. This study aimed at trying out new statistical techniques in comparison with the classical ways as well as finding out how each isolates dominant factors for a child's risk to diarrhoea.

To isolate dominant factors, Logit, Poisson, and Bayesian models were fitted to 2006 Malawi Multiple Indicator Cluster Survey data, collected with an aim of estimating key indicators of women and child health per district. The comparison between Logit and Poisson models was done via chi-square's goodness-of-fit test. Confidence and Credible Intervals were used to compare Bayesian and Logit/Poisson model estimates. Modelling and inference in Bayesian method was done through MCMC techniques.

The results showed agreement in directions of estimates from Bayesian and Poisson/ Logit models, but Poisson provided better fit than Logit model. Further, all models identified child's age, breastfeeding status, region of stay and toilet-sharing status as significant factors for determining the child's risk. The models ruled out effects of mother's education, area of residence (rural or urban), and source of drinking water on the risk. But, Bayesian model proved significant closeness to lake/river factor, which was not the case with Poisson/ Logit model.

The findings imply that classical and semiparametric models are equally helpful, while Poisson is better than Logit model when estimating the child's risk to diarrhoea.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AIC	Akaike Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ANOVA	Analysis of Variance
BIC	Bayesian Information Criterion
CI	Confidence Interval
CrI	Credible Interval
DHS	Demographic and Health Survey
DIC	Deviance Information Criterion
GLM	Generalized Linear Model
HIV	Human Immunodeficiency Virus
LSE	Least Squares Estimation
MDG	Millennium Development Goal
MCMC	Markov Chain Monte Carlo
MH OR	Mantel Haenszel Odds Ratio
MICS	Multiple Indicator Cluster Survey

MLE	Maximum Likelihood Estimation
MoH	Ministry of Health
OR	Odds Ratio
RR	Relative Risk
UNICEF	United Nations Children's Fund
WHO	World Health Organization

CHAPTER 1: INTRODUCTION

1.1 Background to the problem

Infants and young children are more vulnerable to many kinds of diseases than adults due to general weakness of their still growing bodies. Approximately 3.5 million deaths each year are attributable to diarrhoea worldwide, 80% of which occur in children under the age of 5 years (WHO, 2010). In Malawi, the disease accounts for 11% of deaths in children aged below 5 years (2004 Demographic and Health Survey, DHS). Improving a child's health is one of the eight Millennium Development Goals (MDGs) adopted by governments at the United Nations Millennium Summit in 2000. In MDG 4, countries have committed to reducing child mortality rates by three quarters from their national baseline rates between 1990 and 2015 (WHO, 2007).

Study reports have indicated that Malawi is on the right track to achieving MDG 4. For instance, the 2006 Malawi Multiple Indicator Cluster Survey (2006 MICS) showed that infant and under-five children mortality rates declined from 104 and 189 per 1,000 live births, respectively in 2000 to 72 and 122 in 2006 (UNICEF-Malawi, 2010). The decline was attributed to high coverage of immunization and Vitamin A supplementation, elimination of neonatal tetanus, malaria control activities, and increased rates of

exclusive breastfeeding and access to safe drinking water in the country (UNICEF-Malawi, 2010).

1.2 Early childhood development and diarrhoea control policies in Malawi

In October, 2003 the Malawi government through Ministry of Gender, Youth and Community Services adopted a policy that provides guidelines on what has to be done to promote early childhood development. The policy goal is to promote a comprehensive approach to early childhood development programmes and practices for children aged 0 - 8 years, to ensure fulfillment of the rights to fully develop their physical, emotional, social, and cognitive potential. One of the policy objectives is to provide the best start for the children's life, in which strategies such as sensitizing caregivers and communities on the Rights of the Child, encouraging exclusive breast-feeding from 0 to 6 months and encouraging timely introduction of complementary foods are prioritized (Malawi's National Policy on Early Childhood Development, 2003). These strategies provide a right direction toward achieving MDGs 4/6.

In particular, Malawi government adopted UNICEF/WHO's seven-point plan for diarrhoea control, which include fluid replacement to prevent dehydration; zinc treatment; rotavirus and measles vaccinations; promotion of early and exclusive breastfeeding and vitamin a supplementation; promotion of hand washing with soap; improved water supply quantity and quality, including treatment and safe storage of household water; and community-wide sanitation promotion (PATH, 2011).

Based on these guidelines, the diarrhoea control policies were formulated in between 2010 and 2011 by the steering committee having representatives from Ministries of Health and Irrigation, University of Malawi-The Polytechnic and other organisations. The committee recommended achieving political support for raising the profile of diarrhoeal disease; ensuring that policies are effectively coordinated and implemented; increasing collaboration and integration through a Technical Working Group (TWG); developing national programs; and information, education and communication to allow one clear message to be disseminated at a national level (PATH, 2011). Having such a policy document is a positive development for the fight against the disease.

1.3 Childhood diarrhoea situation and risk factors in sub-Saharan Africa

Diarrhoea is an increase in volume of stool or frequency of defecation. It is one of the most common clinical signs of gastrointestinal diseases, but also can reflect primary disorders outside of the digestive system (Mwambete and Joseph, 2010). The disease can be manifested in different levels of clinical intensity, ranging from acute to chronic or severe stages. Acute diarrhoea, which is a common cause of death in developing countries, appears rapidly and may last from five to ten days. Chronic diarrhoea lasts much longer and is the second cause of childhood death in the developing world (www.medicalnewstoday.com).

Diarrhoeal disease remains a leading cause of mortality and morbidity of children in sub-Saharan Africa, a region where unique geographic, economic, political, sociocultural, and

personal factors interact to create distinctive continuing challenges to its prevention and control (Hamer, Simon, Thea and Keusch, 1998).

Many studies have attempted to identify risk factors for childhood diarrhoea in Africa. Hamer, et al (1998) observe that a number of different social, political, and economic factors are present in sub-Saharan Africa which contribute to the constant morbidity from acute and persistent diarrhoea, as well as intermittent epidemics of cholera and dysentery common to this region of the world. This was found through a meta-analysing on studies done in Gambia (1960-87), Guinea-Bissau (1987-90), Kenya (1975-78), Malawi (1983-88), Nigeria (1977-78), Tanzania (1984-85), DRC (1989-90) and Sudan. The data was obtained through Medline database, with an aim of highlighting key areas for future research. They assert that morbidity and mortality from childhood diarrhoea are further compounded by inappropriate household case management and the frequent misuse of antibiotics. They observe that limited knowledge among many health care providers of the proper treatment of diarrhoea also contributes to poor outcomes. Such findings could inform the factors useful for inclusion in a model if statistical models are to be applied on diarrhoea data.

A cross-sectional descriptive survey was conducted in Temeke Municipality, Dar es Salaam, Tanzania over a 4-month period to investigate on knowledge and perception of mothers/caregivers of under-five children on childhood diarrhoea, with focus on frequency of diarrhoea episodes and their risk factors as well as effectiveness of traditional remedies used for its management prior to seeking medical attention. The results from 161 mothers indicated that frequency of diarrhoea episodes was high among

the under-fives and was comparable between females and males. In addition, Mothers' knowledge on predisposing factors of childhood diarrhoea was poor, which was directly correlated with education level. It was also found out that only about one-third of the respondents were aware of risk factors for childhood diarrhoea that cited poor sanitation and water as the main factors. Also, diarrhoea episodes were perceived wrongly as normal growth stage and that they were caused by several other ''illnesses'' (Mwambete and Joseph, 2010).

While the results on ignorance of mothers on potential risk factors to diarrhoea in Tanzania agree with those of Munthali (2005) in Malawi, who found out that mothers and caregivers in Rumphi wrongly associated child diarrhoea to child's teeth development and breastfeeding by a pregnant mother, use of correlation coefficient as a tool to judge usefulness of the factors was statistically weak procedure. Correlation coefficient does provide little information about the relationship between variables. Although it gives a measure of association, the coefficient is not indicative of a regression relationship between the variables (Kleinbaum and Kupper, 1978).

In Accra, Ghana, Boadi and Kuitunen (2005) examined two weeks incidence of diarrhoea among children less than 6 years, with an aim of identifying the risk factors for morbidity due to diarrhoea, using multivariate analysis. The results showed that household economic status, mother education, access to water and sanitation facilities, hygiene practices, flies infestation and regular consumption of street food were significantly associated with the risk. Although the study has a number of risk factors that could be tested if they hold true in Malawian setting, the model used missed to evaluate some key

factors such as child's age and breastfeeding status which have been reported useful in many studies. Thus, a typical statistical model involving several key factors could be much helpful.

1.4 Facts about childhood and diarrhoea

One of the important roles of the healthy bowel is to reabsorb water from the faeces. With diarrhoea, the bowel is unable to do this; hence, the watery bowel movements. This fluid loss can cause the body to become dehydrated (run short of water). This can happen quickly and is a serious problem if not attended to, particularly in the elderly and the very young. The younger the child, the easier it is for it to become dehydrated (UBM Media, 2009). Children are more susceptible to the complications of diarrhoea because a smaller amount of fluid loss leads to dehydration, compared to adults (www.medicalnewstoday.com).

The signs of serious dehydration in children include dry mouth, lips and tongue, or no tears, sunken eyes or fontanelle (the soft spot on top of a baby's head), cold hands and feet or mottled bluish skin, unusual lack of energy, sleepiness or difficult to wake, and fewer wet nappies than usual or unable to drink (UBM Media, 2009).

'Gastro' spreads very easily to others. It is spread when a person touches something that has been in contact with diarrhoea, and then puts his/her hand to the mouth. Some viruses can live on items (including children's toys) for up to 14 days. The spread can be prevented by washing one's hands thoroughly with warm soapy water, especially after using the toilet, before preparing food and after nappy changes (if a child is unwell),

washing and rinsing soiled clothing separately, and not sharing food and drinks (UBM Media, 2009).

1.5 Under-five child diarrhoea situation in Malawi

The UNICEF-Malawi (2010) report cited neonatal conditions, pneumonia, diarrhoea, malaria, AIDS and malnutrition as immediate and most common causes of infant and child mortality in the country, and recommended that the prevailing efforts needed to be sustained and scaled up in some areas in order to maintain the established trend. The report and its recommendation respectively implied that diarrhoea was one of the illnesses that were troubling lives of young children in Malawi, and that the size of the problem was not the same in all parts of the country.

These findings by UNICEF-Malawi were in agreement with those from 2004 Malawi Demographic and Health Survey (2004 Malawi DHS) which found out that dehydration caused by severe diarrhoea was a major cause of morbidity and mortality among young children in Malawi.

The 2004 Malawi DHS indicate that 22% of children had diarrhoea at some time in two weeks preceding the survey. Further, it was found out that under-five child diarrhoea problem was not the same across age groups (that is, highest in age 6-11 months, 41%), across districts (that is, most prevalent in Salima, Kasungu, and Thyolo, $\geq$ 27%), across areas of residence (that is, most prevalent in rural areas, 23%), among other background characteristics of the child. The variations are expected as the effects of place on health and health behaviours are far from uniform across population groups (Burton et al, 2011).

Thus, different children or parts of the country may influence the disease outcome differently, which may be attributed to various factors as well.

As it has been indicated, one may appreciate that diarrhoea becomes a serious problem in infants and young children compared to adults, and hence the former need greater attention from all stakeholders in order to save lives which are under great risk. However, a meaningful consideration will require knowledge of the distribution of the disease throughout the country and across various groups of the children. Thus, there was need for an investigation to find a way of modelling the said distribution.

1.6 Statement of the problem

Statistical models are rarely used in quantifying under-five child diarrhoea distribution in Malawi, despite their potential. It should be appreciated that statistical models have the ability of estimating amount of a child's risk to a disease, such as diarrhoea. Thus, although nationwide surveys such as DHS and MICS have over the years adequately highlighted the high prevalence and incidence rates of diarrhoea in under-five children, very little efforts have been made in such studies to include use of statistical models in identifying the possible risk factors. It should be emphasized that high prevalence or incidence rate of a disease in children with a certain characteristic is not a guarantee that the studied characteristic is a risk factor for the disease, unless proven so by a valid statistical procedure. So, it is high time nationwide studies could go beyond reporting disease incidence or prevalence only, but factors behind such observed rates. However, ignoring statistical models may reflect lack of technical expertise in applying such

models; hence this study creates an opening to proper application of such methods in the studied disease.

1.7 Purpose of the study

This study aimed at investigating variations in risk of diarrhoea in under-five children in Malawi.

1.8 Study objectives

1.8.1 General objective

This study uses statistical models to explain incidence of under-five child diarrhoea in Malawi.

1.8.2 Specific objectives

1.8.2.1 To compare estimates found using classical models and modern semiparametric models.

1.8.2.2 To identify appropriate classical model to use when explaining a child's risk to diarrhoea.

1.8.2.3 To evaluate influence of socio-economic and bio-demographic factors on the child's risk to diarrhoea.

1.9 Significance of the study

It was assumed that stakeholders in the health sector would appreciate the value of using statistical models in explaining under-five child's risk to diarrhoea, based on results from

this study. While most Malawians are aware of the burden of diarrhoea in the life of under-five child in the country, there is limited work done on estimating the possible risk factors of the disease. Thus, use of statistical models will provide another dimension from which proper interventions can be decided.

Various stakeholders in health might have produced point estimates on proportion of young children who risk catching diarrhoea in Malawi, based on some operational observations. But, such raw estimates usually fall short of predictive power, expected validity and reliability, and hence do not command the required believability from some audience in the population. It is for this reason that statistical model estimates may often stand supreme to raw estimates.

Estimates from statistical models usually have some degree of confidence attached to them which, depending on sample size and sampling procedure adopted, guarantees the level of believability to finding the same estimates more number of times if the study was carefully replicated in the population being studied. In addition, statistical techniques are sensitive to insignificant factors in ascertaining relationships between variables. Thus, as opposed to raw estimates, statistical modelling helps to identify helpful as well as worthless factors in measuring relationships of variables.

1.10 Summary

For Malawi to attain MDG 4/6 as well as to monitor progress towards achievement of the same, it needs various initiatives. One approach to do this is to employ modelling techniques to understand the dynamics of various diseases, such as diarrhoea. The aim of

this study is therefore to analyse the variations in the risk of diarrhoea in under-five children in Malawi using statistical models.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter reviews the theory of statistical modelling. Two distinct approaches to modelling, frequentist and Bayesian approaches have been reviewed. Finally, studies that have applied statistical modelling techniques in child diarrhoea are reviewed.

2.2 Statistical modelling

2.2.1 Definition of a statistical model

A statistical model is a set of mathematical formulae and assumptions that describe a real-world situation (Aczel and Sounderpandian, 2002). Thus, the concept of statistical model goes hand in hand with that of probability distribution and the population parameters. A parameterized statistical model is a parameter set, Θ, together with a function, $P: \Theta \rightarrow \Phi(S)$, which assigns to each parameter point $\theta \in \Theta$ a probability distribution P_θ on a sample space, S (McCullagh, 2002). Further, a model is referred to as parametric when its functional form is completely specified, except for the values of the unknown parameters; it is non-parametric when it is a set of probability distributions with infinite dimensional parameters; and it is referred to as semi-parametric model when it

also has infinite dimensional parameters, but is not dense in the space of distributions (Kleinbaum & Klein, 2005).

Essentially, a statistical model is a formalization of relationships between variables in the form of mathematical equations. It describes how random variables are related to one another. The model is statistical as the variables are not deterministically but stochastically related. It is frequently thought of as a pair (*Y*, *P*), where *Y* is the set of possible observations and *P* the set of possible probability distributions on *Y*. It is assumed that there is a distinct element of *P* which generates the observed data. Statistical inference enables one to make statements about which element(s) of this set are likely to be the true one (Spanos, 2003; Long, 1977; Dobson, 2002; Agrest, 1996).

Briefly, it suffices to use statistical modelling in this study to explain the distribution of under-five diarrhoea in Malawi, as this involve seeking relationships between the diarrhoea variable and other variables.

2.2.2. Classes of statistical models

Statistical models can be classified according to number of endogenous variables (that is, variables whose values are determined directly within the system of equations) and the number of equations. In this regard, models in which number of equations equals number of endogenous variables are referred to as complete models, and incomplete models are those in which there is imbalance between number of equations and the number of endogenous variables. Other popular classifications are the general linear model (that is, a model restricted to continuous dependent variables, for instance, linear regression

model), the generalized linear model (that is, a model that can allow discrete dependent variables, for instance, logistic regression model), the multilevel model (that is, a model with systematic term reflecting different levels of data, for instance, empirical Bayes model), and the structural equation model (that is, a model that takes into account both linear and nonlinear effects of systematic term, for instance, Bayesian structured additive model) (Ader, 2008).

In summary, it is important to learn various classes of statistical models so that the value of the ones proposed in this study can be judged accordingly.

2.2.3 Stages of building a statistical model

It is often the wish of every researcher that the assumed model should explain as much as possible about the process underlying the data at hand. But, due to uncertainty inherent in all real-world situations, the proposed model may not explain everything; there would always be some remaining errors. The errors may be due to unknown outside factors that can affect the process generating the data at hand (Aczel and Sounderpandian, 2002). Thus, dealing with errors forms part of the model building process.

The process of model building generally involves four steps that almost follow each other logically. These include model specification, which is basically laying down the formula and stating its assumptions. The next stage is to estimate the parameters of the model from the data set. Thereafter, examination of the residuals and testing for appropriateness of the model is done. Finally, the model is used for its intended purpose, if appropriate;

otherwise, it is aborted (Aczel and Sounderpandian, 2002; Dobson, 2002; Kleinbaum and Kupper, 1978).

Concisely, at the first stage a particular model is proposed that describes a given situation, for instance a simple linear regression model that describes the relationship between two variables may be proposed. A model is specified in two parts: an equation linking the response and explanatory variables and the probability distribution of the response variable (Dobson, 2002).

The second stage, which involves estimation of model parameters, is achieved through many techniques, such as Maximum Likelihood Estimation (MLE) procedure, Least Squares Estimation (LSE) procedure, among other possible ways. The MLE method finds estimate of the parameter that maximizes the likelihood function (joint probability distribution function) of the parameter given data set at hand. Thus, making assumptions about the probability distribution of response variable in the model and getting a random sample are both necessary with the MLE procedure (Dobson, 2002). The LSE method, commonly used in linear regression models, finds estimates of the parameters that are best linear unbiased estimators (blue) of regression parameters and that have lowest or minimum variance of all possible unbiased estimators of the regression parameters as specified by the *Gauss-Markov theorem* (Aczel and Sounderpandian, 2002; Dobson, 2002; Kleinbaum and Kupper, 1978).

The third stage considers the observed errors that result from fitting the model to data. The observed errors, called residuals, represent the information in the data not explained

by the model. In other words, they reflect the difference between the observed and fitted values of the response variable (Aczel and Sounderpandian, 2002; Dobson, 2002). Thus, this stage checks the adequacy of the model - that is, how well the model fits or summarizes the data (Dobson, 2002). For instance, in Analysis of Variance (ANOVA) model the within-group variation is due to the residuals. If the residuals are found to contain some non-randomness, systematic component, the proposed model is re-evaluated, and if possible, adjusted to incorporate the systematic component found in the residuals; or, the model may be discarded and a different one can be tried (Aczel and Sounderpandian, 2002).

Finally, at fourth stage it is where the model is used for its intended purpose; that is, prediction of variable, control of variable, or the explanation of the relationships among variables. This happens when it is believed that the model residuals contain nothing more than pure randomness (Aczel and Sounderpandian, 2002). This stage is a statistical inference stage where calculation of confidence intervals and testing of hypotheses about the parameters in the model and interpretation of results are made (Dobson, 2002).

Thus, reviewing stages of building a model informs procedures that this study has to abide by when applying chosen models to the diarrhoea dataset.

2.2.4 General and Generalized Linear Regression Models

As earlier alluded to, General Linear Models are those which are restricted to formalizing relationships between one or more explanatory variables and a continuous dependent

variable. Generalized Linear Models (GLMs), on the other hand, allow any form of dependent variables.

2.2.4.1 Historical perspective of regression models

The term 'regression', as it is used today, refers to the statistical technique of modelling the relationship between variables. Its history dates back to 1889 when an Englishman by the name Sir Francis Galton published a paper on heredity, "Natural Inheritance" (1889) in which he reported his discovery that sizes of seeds of sweet pea plants appeared to "revert," or "regress," to the mean size in successive generations (Aczel and Sounderpandian, 2002, p. 435). He also reported results of a study of the relationship between heights of fathers and the heights of their sons. A straight line was fit to the data pairs: height of father versus height of son. Here, too, he found a "regression to mediocrity": the heights of the sons represented a movement away from their fathers toward the average height (Aczel and Sounderpandian, 2002, p. 435).

2.2.4.2 Generalized Linear Modelling

The term Generalised Linear Model is due to Nelder and Wedderburn (1972) who showed how the linearity could be exploited to unify apparently the diverse statistical techniques (McCullagh & Nelder, 1989; Dobson, 2002). Later Wedderburn (1974) used quasi-likelihood method that allows less strict error assumptions to estimate the regression parameterss (McCullagh & Nelder, 1989). But earlier contributions were made by Gauss (1823) who introduced the least squares method; Fisher (1922) who used model techniques in Agriculture experiments; Bliss (1935) who introduced the probit regression; Dyke & Patterson (1952) who used Logit for proportions (McCullagh & Nelder, 1989).

The main concepts about GLMs are that the response variables have distributions other than Normal, they may even be categorical rather than continuous, and that the relationship between the response and explanatory variables need not be of the simple linear form. In short, a GLM is linear model for transformed mean of response variable with distribution in exponential family. It extends classical regression models to encompass non-normal response distributions and modeling functions of mean (Agrest, 2002; McCullagh & Nelder, 1989).

GLMs expound ideas of "regression to the mean" as they relate a function of mean of a random response variable to the explanatory variables through a prediction equation having linear form. Briefly, a GLM contains three components: the "random component" that identifies the response variable, for example, *Y* and assumes a probability distribution for it; the "systematic component" that specifies the explanatory variables used as predictors in the model; and the "link" that describes the functional relationship between the systematic component and the expected value (mean) of the random component (Agrest, 1996). Good examples of GLMs include logistic and Poisson models, these are discussed in detail in Chapter 3. Thus, the random component consists of the response variable *Y* with independent observations *y1*, . . . , *yN* from a distribution in natural exponential family, that is a distribution whose probability density or mass function can be written in the form;

$$f(\text{yi};\ \theta\text{i}) = a(\theta\text{i})b(\text{yi})\exp[\text{yi}Q(\theta\text{i})] \text{ or } f(y;\ \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)],$$

where $Q(\theta i)$ or $b(\theta)$ is called a *natural parameter*, while any other parameters, in addition θ, are called *nuisance parameters* forming parts of *a*, *b*, *c* and *d*, and treated as though they are known. Further, if $a(y) = y$, that is the link function that transforms the mean to the natural parameter is called the *canonical link*, and the distribution is said to be in *canonical* (i.e. standard) form. The link function $g(\mu i) = \mu i$ is called *identity link*, and it has $\eta i = \mu i$. The systematic component relates a vector ($\eta i, \ldots, \eta N$) to explanatory variables *xi* through a linear model, that is, $\eta i = \Sigma \beta j x i j$, which is called a *linear predictor*. Then the link function connects the random and systematic components. For example, if $\mu i = E(Yi)$, the model links μi with ηi by $g(\mu i)$, where *g* is a monotonic, differentiable function. Thus, *g* links $E(Yi)$ to *xi* through $g(\mu i) = \Sigma \beta j x i j$ (McCullagh & Nelder, 1989; Dobson, 2002).

Therefore, it is beneficial to learn that GLMs allows any form of dependent variable in a relationship, since the response variable proposed in this study is the discrete diarrhoea variable.

2.2.5 Bayesian modelling

2.2.5.1 Historical perspective of Bayesian modelling

As earlier alluded to, a statistical model tries to estimate the values of unknown population parameters using observed data. Apart from GLMs which fall under classical models, the estimation can also be done through modern Bayesian techniques which assert that population parameters are random (but unknown) quantities rather than being regarded as fixed quantities, as in classical approaches. Thus, Bayesian techniques

consider population parameters as having their own probability distributions in the population, called prior distributions.

The approach is called 'Bayesian' because the mathematical link between the probabilities associated with data results and the probabilities associated with the prior information is Bayes' theorem discovered in 1761 by the English clergyman Thomas Bayes. His work was presented to the Royal Society of London by a friend in 1763-after Bayes' death (Aczel and Sounderpandian, 2002). The theorem allows one to combine the prior information with the results of sampling to obtain posterior (post sampling) information.

2.2.5.2 Bayesian statistical inference

Classical/Frequentist statistical approaches to inference assume that population parameter θ is a constant (but unknown) quantity in the population and can be easily estimated from sample data by just knowing the probability distribution of the random dependent variable Y in the population (that is, knowing $P(Y)$, $E(Y)$ and $Var(Y)$), which is usually done via point estimates, confidence interval estimates, or hypothesis testing. On the other hand, Bayesian approaches assert that θ is a random (but unknown) quantity and it has its own probability distribution in the population (called prior distribution), that is, $P(\theta)$, $E(\theta)$, and $Var(\theta)$. Hence, estimation of θ which is based on its posterior distribution has to combine/mix the distribution of Y (sampling information) and that of θ (prior information) (Ridall, 2007; Carlin and Louis, 2000; Aczel and Sounderpandian, 2002). The mathematical formula that accomplishes this is Bayes' theorem. The Bayes' theorem for a discrete random variable is given as

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{\sum_{i=1}^{n}[P(y \mid \theta_i)P(\theta_i)]},$$

where θ is an unknown population parameter to be estimated from the data *y*. The summation in the denominator is over all possible values of the parameter of interest θ, and *y* stands for the particular data set.

Thus, at a minimum, a Bayesian model requires an additional component in the form of a prior distribution on Θ. A Bayesian model in the sense of Berger (1985), Smith (1984) or Bernardo and Smith (1994) requires an extra component in the form of a judgment of infinite exchangeability or partial exchangeability in which parameters are *defined* by limits of certain statistics.

In addition, Bayesian statistical conclusions about the parameter θ, or unobserved data, are made in terms of probability statements which are conditional on the observed data values y (Gelman et al, 2004). An attractive feature of Bayesian inference is that inferences are made conditional on the observed data which is not the case in classical statistics where one must think about the possibilities of data sets distinct from the one actually observed. The only data set relevant for drawing conclusions in Bayesian inference is the data set that one sees (Albert, 1995).

In a nutshell, the under-five child diarrhoea prevalence or incidence being estimated in this study may not be static in the population but changing with time due to other factors. Thus, engaging Bayesian as well as classical approaches was a necessary mix.

2.2.5.3 Parameter estimation in Bayesian inference

It should be appreciated that the difficulties of Bayesian modelling are how one gets to the posterior. One needs to specify a prior (sometimes admitting lack of knowledge about the model initially: a so-called *uninformative* or *flat prior*) and then estimation of the shape of the posterior distribution follows. The analytical problem in this is that all the posterior probabilities must add up to 1 (the prior and posterior distributions are probability density functions), and thus the product of the likelihood function and the prior distribution must be standardized by a sum of probabilities of all possible outcomes. In most cases this cannot be figured out analytically and so estimation of the shape of the posterior must be done using a numerical random sampling technique, for instance, Markov chain Monte Carlo simulation (MCMC). In many ways this is similar to the optimization routines needed to calculate Maximum Likelihood Estimates (MLEs): like the simulated annealing algorithm, MCMC works sequentially to find new parameter values using random jumps through parameter space.

Of the several common MCMC variants used to 'solve' (actually sample from) the posterior distribution, perhaps the most popular is called *Gibbs sampling*. This algorithm (often called the Gibbs sampler), available in most statistical software for Bayesian data analysis, is particularly appropriate for estimating the value of one parameter conditional on values of a host of other parameters, which is especially effective for fitting hierarchical Bayesian models. Given a set of parameters in the specified model, the Gibbs sampler starts a Markov chain with a set of initial values and then performs the $\boldsymbol{i}^{\text{th}}$ iteration, for example, for i = 1, 2, …, m, by updating successively from the full

conditional distributions. The completion of such a loop results in a single iterate of the Gibbs sampler with an update of first parameter estimate. The process is repeated *m* times to obtain a Gibbs sample of *m* vectors. From Markov chain theory it is obvious that such a chain will eventually converge to a *stationary* or *equilibrium* distribution which is *precisely* the posterior distribution upon which the Bayesian data analysis is based (Banerjee, 2011).

It suffices to apply MCMC estimation techniques in this study, considering the fact that some variables assumed in the relationship were continuous which were believed to have nonlinear or varying effects that could not be captured using MLE methods.

2.2.6 Strengths and limitations of Bayesian modelling over classical modelling

As one may appreciate, Bayesian modelling is a generalization of linear and generalized linear modelling in which regression coefficients are themselves given a model whose parameters are also estimated from data. Like regression methods, multilevel Bayesian models can be used for a variety of purposes, including prediction, data reduction, and causal inference from experiments, and observational studies (Gelman, 2005). Compared to classical regression, multilevel modelling is almost always an improvement, but to different degrees: "for prediction, multilevel modelling can be essential, for data reduction it can be useful, and for causal inference it can be helpful" (Gelman, 2005, p.1).

In terms of data reduction, the inferences from the Bayesian models are more reasonable compared to classical estimates. It is common to have identical estimates for all levels when using classical regression. For example, an estimate may be obtained that predicts

amount of change in number of children to suffer from diarrhoea corresponding to a one unit change in region of stay which may be particularly inappropriate for multilevel application whose goal is to identify the locations in which residents are at high risk of diarrhoea. In addition, the classical regression model may over fit the data, for example giving an implausibly high estimate of the average number of children to suffer from diarrhoea. This can happen in areas where only few diarrhoea observations were available. Multilevel modelling avoids this by taking into account variations in the data at both individual and group levels (Gelman, 2005).

Another advantage of multilevel modelling for this application is that it allows one to study the relation of, for example, household parameters to household-level predictors. It would be possible to estimate this second-level relation using classical regression, but this would mean fitting two separate models: one for unpooled, and the other for completely pooled data. The multilevel model has the appeal of fitting the two levels together, and can actually be implemented using a Gibbs sampler alternating between the data-level and household-level regression steps (Gelman, 2005).

In terms of prediction, Gelman (2005) acknowledges that perhaps the clearest advantage of multilevel models comes in prediction. He demonstrated using cross-validation analysis that root-mean-squared cross-validation errors for multilevel model estimates were always the smallest compared with complete pooling and no-pooling classical regression model estimates for radon levels in U.S. homes. Thus, the multilevel model gives more accurate predictions than the no-pooling and complete-pooling regressions, especially when estimating group averages. Thus, multilevel models have the ability to

separately estimate the predictive effects of an individual predictor and its group-level mean which are sometimes interpreted as “direct” and “contextual” effects of the predictor (Gelman, 2005).

In addition, most parametric models often lack the capability of identifying non-linear relationships between dependent and independent variables. The use of Bayesian semiparametric approaches avoids these shortcomings (Jerak and Wagner, 2003).

In terms of causal inference, multilevel models can easily be misinterpreted. With identical data of a social nature, it would be easy to leap to a misleading conclusion and find contextual effects. However, strong predictive effects of model predictors cannot necessarily be interpreted causally for observational data even if these data are a random sample from the population of interest (Gelman, 2005). Complication arises if one considers possibility of correlation between individual-level predictor, *x*, and for example, cluster-level error. By simply multiplying likelihood and prior densities, the posterior density implicitly assumes the cluster errors are independent of *x*.

Nevertheless, classical models are essential in a wealth of applications where one needs to compensate for the paucity of the data (McCullagh, 2002). The various approaches to data analysis (Frequentist, Bayesian, machine learning, exploratory or whatever) should be seen as complementary to one another rather than as competitors for outright domination (Julian Besag in McCullagh, 2002). Unfortunately, parametric formulations become easy targets for criticism when, as occurs rather often, they are constructed with too little thought (Julian Besag in McCullagh, 2002). The lack of demands on the user

made by most statistical packages does not help matters and, despite enthusiasm one may have for Markov chain Monte Carlo (MCMC) methods, their ability to fit very complicated parametric formulations can be a mixed blessing (Julian Besag in McCullagh, 2002).

In summary, the researcher has estimated child's risk to diarrhoea using both classical and multilevel models with the aim of assessing consistency of findings from both groups of models.

2.3 Model strategies on child diarrhoea distribution in sub-Saharan Africa

Much as various organizations working in the health sector may provide estimates on child diarrhoea prevalence in various locations of Malawi, there will always be need for properly conducted scientific research on the same in order to complement their efforts as well as to come up with scientifically relational, valid and reliable findings.

One such study was conducted in Nigeria by Kandala et al (2008) on diarrhoea mapping, where data from 1999 and 2003 Nigerian Demographic and Health Surveys (Nigerian DHS) were compared in their analysis. The aim of the study was to reveal and explore inequalities in the health of Nigerian children by mapping the spatial distribution of childhood morbidity associated with diarrhoea, cough, and fever, and accounting for important risk factors, using Bayesian geo-additive model based on Markov-Chain-Monte-Carlo techniques.

This was done against the background that diarrhoea, cough, and fever were the leading causes of childhood morbidity and mortality in sub-Saharan Africa, and that geographical location was seldom considered as an explanatory factor for the large regional variations in childhood morbidity attributed to the three causes in the region.

The results showed that overall prevalences of diarrhoea, cough, and fever recorded in 1999 (among children aged 3 years) were similar to those seen in 2003 (among children aged 5 years). However, the results revealed that morbidity attributable to each of the three causes varied differently at state level. In addition, place of birth (hospital v. other), type of feeding (breastfed only v. other), parental education, maternal visits to antenatal clinics, household economic status, marital status of the mother, and place of residence (urban v. rural) were each significantly associated with the childhood morbidity investigated.

Further, both surveys revealed that children from urban areas were found to have a significantly lower risk of fever than their rural counterparts. It was also found out that most other factors affecting diarrhoea, cough, and fever differed in the two surveys. Besides, the risk of developing each of the three conditions increased in the first 6-8 months after birth, but then gradually declined.

Similar methods could expose same variations in child diarrhoea in Malawi at individual, household or regional level if applied to Malawian national data. Other catchy results from the study were the agreement and disagreement of some risk factors to diarrhoea, cough, or fever in the 1999 and the 2003 surveys conducted in the same country. Thus,

there was need to investigate the under-five child diarrhoeal situation in Malawi in order to investigate if strength of some risk factors found in previous studies remain the same.

A similar study to that conducted in Nigeria was done in Malawi by Kazembe et al (2009) where they investigated joint and disease-specific spatial clusters of fever and diarrhoea at a highly disaggregated level while estimating the influence of other covariates. They fitted to the 2000 Malawi DHS data a logistic model with spatial random effects that were partitioned into shared and specific effects. The results showed that shared area-specific effects were persistently high in central and southern regions of the country. On the other hand, fever-specific effects were high along the Lakeshore areas, and diarrhoea-specific effects were excessive in central and south-eastern zones of the country.

While the results from the study should be appreciated, there was need to verify if similar approaches would yield same results on most current national data. This is the case since different stakeholders in health have used various interventions to respond to the 2000 under-five diarrhoeal situation in the country. In addition, a number of factors in Malawian localities have changed between 2000 and 2011.

Interestingly, diarrhoea-specific effects were excessive in central and south-eastern regions, while fever-specific effects were high along Lakeshore areas. One would expect diarrhoea-specific effects to be prominent in lakeshore areas due to expected high use of non-treated and unsafe water for domestic purposes by inhabitants of these areas. This invited further questions for study.

2.4 Summary

This Chapter has explored the notion of statistical modelling, that is, a tool for establishing relationships among random variables. Cases of models in child diarrhoea in sub-Saharan Africa have been explored. A search for literature has revealed that very few studies have recently used statistical models to estimate distribution of child diarrhoea in Malawi.

CHAPTER 3: METHODOLOGY

3.1 Research design

This study was an applied quantitative research on secondary data. It employed statistical procedures of Bayesian semiparametric additive, Logistic, and Poisson regression modelling on 2006 Malawi Multiple Indicator Cluster Survey (MICS) data.

3.2 Appropriateness of design

The national survey data used was large enough to allow implementation of intended statistical analyses, since estimates of a random variable from a large random sample are believed to possess all optimal properties of an estimator. Perhaps due to the rigorous process of random sampling employed in most surveys it happens that surveys usually give accurate estimates of population parameters (Hansen et al, 1996), a property that is desirable in statistical inference. Further, the national survey data had cross-sectional information from all districts which would make it possible for the researcher to estimate the distributions of child diarrhoea and amount of risk posed by various parts of the country.

3.2.1 Logistic regression model

The logistic regression model was used due to the fact that the outcome variable, two-week total number of diarrhoea cases, was believed to follow binomial distribution. Introduced in the 1940s, Logistic regression is an example of a GLM where the random

component is a Bernoulli random variable whose distribution is specified by probabilities $P(Y = 1) = \pi$ of success and $P(Y = 0) = 1 - \pi$ of failure. If the outcome of a trial can only be either a success or a failure, then the trial is called a Bernoulli trial. The total number of successes $\Sigma(Y=1)$ in one Bernoulli trial, which can be 1 or 0, is called a Bernoulli random variable ($\Sigma(Y=1) \sim Ber(\pi)$. When many independent and identical Bernoulli trials n have been carried out, the resulting sequence of identically and independently distributed Bernoulli variables is called a Bernoulli process (Aczel and Sounderpandian, 2002). For n independent observations on a binary response with parameter π, the total number of successes, $\Sigma(Y=1)$ has the Binomial distribution specified by the indices n and π (*that is,* $\Sigma(Y=1) \sim Bin(n, \pi)$), and belonging to the exponential family of distributions, that is, the probability mass function has the form $f(y; \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$.

Each diarrhoea observation in the MICS data had two possible outcomes; either a child suffered from diarrhoea or did not. Thus, each outcome was a Bernoulli process. Further, it is known that the total number of children that were observed is fixed, with n = 15, 018, and from the 2004 DHS prevalence of under-five child diarrhoea in Malawi was 22%, hence π = 0.22, which was believed constant from one observation to another in the children population. In addition, each observed child was an individual and therefore the outcome in an observed child could not influence the outcome of the next child. Hence, the outcomes were independently distributed in the children's population. Therefore, the total number of children who could suffer from diarrhoea in the country at any time of observation was a binomial random variable. Its probability mass function is specified as

$$f(y;(n,\pi)) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y} = \binom{n}{y}(1-\pi)^{n}\left[\frac{\pi}{1-\pi}\right]^{y} = \exp\left[\log\binom{n}{y} + n\log(1-\pi) + y\log\frac{\pi}{1-\pi}\right],$$

which is an exponential form with $c(\pi) = (1- \pi)^n$, $d(y) = \binom{n}{y}$, and $b(\pi)$ or $Q(\pi)= \log(\pi/(1-\pi))$ and $a(y) = y$.

The natural parameter is therefore $\log(\pi/(1- \pi))$, log of odds of response 1, the *logit of* π, it's *canonical link*. Because of this link function, the binomial or logistic model is also called logit model.

The actual value of π in the population can vary as the value, x of X varies; hence the notation π may be replaced by $\pi(x)$ to reflect its dependence on that value (Agrest, 1996).

The relationship between x and $\pi(x)$ is a nonlinear *S*-shaped curve, called logistic function, given below:

$$\pi(x) = E(\Sigma(Y=1)|x_1, \ldots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1+\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)},$$

where $\beta_0, \ldots, \beta_p$ are parameters to be estimated from the data y.

In the situation where the explanatory variable x_1 is a binary exposure of interest, $\exp(\beta 1)$ is the adjusted ratio of the odds of the outcome occurring in the exposed group versus the non-exposed group, adjusting for effects of the other explanatory variables $x_2, \ldots, x_p$ (Aczel and Sounderpandian, 2002; Agrest, 1996). As x gets large, $\pi(x)$ approaches 0 if $\beta < 0$ and it approaches 1 if $\beta > 0$.

The transformation given below, logarithm of odds of success, called Logit transform, linearizes the logistic function;

$$\hat{\pi}(x) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p .$$

The estimation of parameters is usually done through Maximum Likelihood technique, explained before. The Maximum Likelihood estimates of the parameters β, and consequently of the probabilities $\pi i = g(\boldsymbol{xi}^T\beta)$, are obtained by maximizing the log-likelihood function;

$$l(\pi; y) = \sum_{i=1}^{N}\left[y_i \log \pi_i + (n_i - y_i)\log(1-\pi_i) + \log\binom{n_i}{y_i}\right]$$

using iterative weighted least squares procedure (see Dobson, 2002).

3.2.2 Poisson regression model

Since for *n* Bernoulli iid observations, the total number of diarrhoea cases, Σ(Y=1), at any time was a positive integer, then Poisson distribution was assumed for the total number of cases throughout the four-month period. When the response is a count, one can use a count regression model to explain this response in terms of the given predictors. Sometimes, the total count is bounded, in which case a binomial response regression should probably be used. In other cases, the counts might be sufficiently large that a normal approximation is justified so that a normal linear model may be used (Faraway, 2006). One of the common distributions for counts is Poisson. If $\Sigma(Y=1)$ is a Poisson random variable with mean $\mu = E[\Sigma(Y=1)] = Var[\Sigma(Y=1)] > 0$, then:

$$P(\Sigma(Y=1) = y) = \frac{\exp(\mu)\mu^{y}}{y!} = \exp\left[\mu - \log(y!) + y\log\mu\right],\ y = 0, 1, 2, \ldots$$

From the exponential form, it is clear that the link function relating μ with predictors is log link given by $\log\boldsymbol{\mu} = \alpha + \Sigma\beta x$, where the parameters are estimated using usual procedure of MLE.

According to Faraway (2006), the Poisson distribution arises naturally in several ways. For instance, if the count is some number out of some possible total, then the response would be more appropriately modelled as a binomial. However, for small success probabilities and large totals, the Poisson is a good approximation and can be applied. For example, in modelling the incidence of rare forms of cancer, the number of people affected is a small proportion of the population in a given geographical area. A Poisson regression model can be used in preference to a binomial. If $\mu = n\pi$ while $n \to \infty$, then $B(n, \pi)$ is well approximated by $Pois(\mu)$. Also, for small π, $\text{logit}(\pi) \approx \log\pi$, so that the use of the Poisson with a log link is comparable to the binomial with a logit link.

It is important to mention that to allow for correlation within households, robust standard error was to be calculated using residuals at the cluster level. An important result concerning Poisson random variables is that their sum is also Poisson. Specifically, suppose that $Y_i \sim Pois(\mu_i)$ for $i = 1,2,\ldots$ and are independent, then $\Sigma_i Y_i \sim Pois(\Sigma_i \mu_i)$. This is useful because sometimes one has access only to the aggregated data. If we assume the individual-level data is Poisson, then so is the summed data and Poisson regression can still be applied (Faraway, 2006; Dobson, 2002).

From the 2004 Malawi DHS, $\pi = 0.22$ and $n \approx 10{,}000$ for under-five child diarrhoea. Therefore, π was considerably small while n being a large number. Hence, approximation of binomial regression with logit link by a Poisson regression with a log link was valid.

3.2.3 Bayesian semiparametric structured additive model

The term Bayesian semiparametric model comes from the fact that both parametric and nonparametric forms of relationship are assumed in one model. In this respect, continuous covariates are treated non-parametrically with the help of smoothing functions whereas categorical variables are related parametrically to the response variable (see Kandala, 2001; Jerak and Wagner, 2003). In general, a Bayesian model is considered to be a regression (a linear or generalized linear model) in which the parameters—the regression coefficients—are given a probability model (Gelman and Hill, 2011).

The use of semiparametric model was therefore thought of in order to capture both linear effects of discrete covariates and nonlinear effects of continuous covariates on the child's risk to diarrhoea. Further, the data had some categorical explanatory variables with more than two levels; hence the model was employed in order to show the results in reduced form of the covariates.

Assuming that total number of observed cases at any time in the four months MICS study, $\Sigma(Y=1)$ is a random variable belonging to an exponential family with parameters n and π, then $\Sigma(Y=1)$ satisfies the logistic model

$$\log it(\pi) = \alpha + \sum \beta x,$$

where α and β stand for parameter components, and X for a vector of factors or covariates.

Further, it was assumed that α and β were distributed as gamma random variables with fixed scale and location parameters, u = v = 0.001, except for a continuous variable child age whose parameters were assumed to have normal prior distributions with 0 means and inverse gamma distributed variances.

It was assumed that regression parameters in this model are not static, but vary at: (1) child's individual-level; with focus on child's age and breast feeding status; (2) child's family-level; focusing on mother's education, family's source of drinking water, and whether or not the family toilet is shared and (3) child's residential location; with focus on region of stay (north, centre, or south), closeness to the lake or river (lake/river shore or highland), and area of residence (rural or urban).

Briefly, the GLMs assume that, given covariates u and unknown parameters, the distribution of the response variable y belongs to an exponential family, with mean $\mu = E(y \mid u, \gamma)$ linked to a linear predictor η by

$$\mu = h(\eta), \qquad \eta = u'\gamma .$$

Here, h denote a known response function, and γ are unknown regression parameters. The following structured additive predictor was used in this study to estimate a flexible Bayesian semiparametric model that was fitted to the MICS data (see Brezger, Kneib and Lang, 2005):

$$\eta_r = f_1(x_{r1}) + \ldots + f_p(x_{rp}) + u'_r \gamma ,$$

where $\boldsymbol{r}$ is a generic observation index, x_{rj} denote generic covariates of different type and dimension, and f_j ($j = 1, 2, \ldots, p$) are (not necessarily smooth) functions of the covariates. The functions f_j may comprise the usual nonlinear effects of continuous covariates, time trends and seasonal effects, two-dimensional surfaces, varying coefficient terms, *i.i.d.* random intercepts and slopes, spatially correlated effects, and geographically weighted regression (Brezger, Kneib and Lang, 2005).

Once a model of this type is specified, inferences can be drawn from available data for the population means at any level of the data. These estimators, which can be regarded from a Bayesian perspective as posterior means or from a Frequentist perspective as "Best Linear Unbiased Predictors" (BLUPs), often have better properties than simple sample-based estimators using only data from the unit in question. This makes them useful in the problem of "small-area estimation," that is, making estimates for units or domains for which there is a very limited amount of information (Skinner et al, 1989).

3.3 Geographic location and population distribution

As earlier alluded to, MICS was conducted in all districts in Malawi, a country that is located in south-east Africa, landlocked between Mozambique to its eastern and southern sides, Zambia to its western side, and Tanzania to its northern side. It covers a total earth surface area of 118,484 km^2, of which 94,276 km^2 (79.6%) is made of land and 24,208 km^2 (20.4%) is made of water. By 2008, the country had a population of 13,077,160 people and its land was divided into three major regions: the central, 35,592 km^2 had 5,510,195 people (42.14% of national population); the northern, 26,931 km^2 had

1,708,930 people (13.08% of national population); and the southern region, 31,753 km^2 with 5,858,035 people (44.80% of national population). About 90% of the country's population lives in rural areas where, among other things, access to health services and poverty are major hardships (National Statistical Office of Malawi, 2008 Population & Housing Census). The country's population had 7,157,985 (45.1%) people in the age group of 0-14 years (3,586,696 males; 3,571,298 females), as of October, 2011 (2011 Index Mundi).

3.4 Study Population

The MICS study sampled 31,200 occupied households and interviewed 30,553 of them, indicating a 97.9% household response rate. In addition, 23,238 under-five children were listed from the interviewed households, of which questionnaires for 22,994 were completed, corresponding to a 98.9% response rate. Further, 27,073 women (age 15-49 years) were identified from the interviewed households, of which 26,259 were interviewed, yielding a response rate of 97.0%. Also, 8,556 men (age 15-49 years) were identified in every third household and 7,636 of them were interviewed, giving a response rate of 89.2% (2006 MICS Report).

The targeted population in this study was children aged at most 5 years. The outcome variable of interest was cases/non-cases of diarrhoea in the diarrhoea as in 2006 MICS. The explanatory variables included child's age, child's breastfeeding status (weaned or still breastfeeding), a child's area of residence (rural or urban), region of stay (northern,

central or southern), toilet facility (shared between families or not), mother's education, source of drinking water, and closeness to the lake/river.

3.5 Sampling design of 2006 MICS

3.5.1 Sample size

With an aim to obtain estimates, at district level, on key indicators related to the wellbeing of children and women, the MICS study targeted a sample of size 1,200 households (HHs) per district to obtain statistically valid estimates at 95% CI for the majority of indicators. By then, there were 28 districts in Malawi, two of which (Likoma and Neno) were too small to draw 1,200 HHs out of the total available HHs. As a result, Likoma was merged with Nkhata Bay and Neno with Mwanza, thereby reducing the number of study districts to 26. Weighted estimates for the three regions and Malawi as a whole were obtained based on the data from the 26 districts (2006 MICS Report).

3.5.2 Sampling technique

A two-stage cluster sampling design was used to select the HHs, where within each district 40 census enumeration areas (identified as clusters) were selected, and within each cluster a systematic sample of 30 households was drawn. A total of 31,200 HHs (26 districts multiplied by 1,200 HHs) were selected in 1,040 clusters (26 districts multiplied by 40 clusters) in that process. The 1,040 selected clusters were all visited during the fieldwork period (2006 MICS Report).

3.6 Instrumentation and data collection

The MICS study, conducted from July to November, 2006, used four questionnaires that were translated into Chichewa and Tumbuka vernacular languages to collect data. One questionnaire, termed household questionnaire, administered to the head of the household or any person who was able to provide the information was used to identify all eligible persons for the specific forms. It collected information regarding household listing, education, water and sanitation, household characteristics, insecticide treated nets, orphan-hood, child labour, and salt iodization.

The other questionnaire, called under-five children questionnaire, administered to mothers or caretakers of under-five children collected information on Vitamin A, breastfeeding, care of illness, diarrhoea, malaria, immunization, and anthropometry. Another questionnaire, termed women questionnaire, administered to women aged 15-49 years gathered data on child mortality, birth history, tetanus toxoid, maternal and newborn health, marriage/union, contraception, sexual behaviour, HIV/AIDS, and maternal mortality. The fourth questionnaire, called men questionnaire, administered to men aged 15-49 years collected data on marriage/union, contraception, sexual behaviour, and HIV/AIDS.

3.7 Confidentiality and ethical clearance on data use

The MICS data do not show identities and particulars of its respondents. Thus, this study has maintained confidentiality of participants in reporting of results. The data was used with permission from the National Statistical Office of Malawi which was granted

through the Coordinator of the Master of Science in Biostatistics Programme at Chancellor College, University of Malawi.

3.8 Data analysis procedures

Analysis of data is one of the crucial stages of the research process. It is the properly analysed data whose results become easy to interpret and understand. Hence, this study partitioned this stage into further sub-stages, as indicated below for purposes of straightforward interpretation.

3.8.1 Baseline analysis

The sample data were examined in Stata package to check if all variables under study had complete values for all the data points or if there were some missing values. The children with incomplete data in some variables were dropped from analysis, with randomness assumption. Further, the baseline characteristics of the children with complete information were analysed in Stata Version 10 package. These included the totals and percentages of studied children based on the individual, household, cluster location, and regional characteristics.

The variable-specific estimates of two-weeks diarrhoea incidences were calculated in Stata Version 10 package. This explored the incidences before applying the statistical models to the data in light of the objectives to this study.

3.8.2 Cross-tabulations with outcome variable

The crude odds ratios (ORs) estimating a child's risk to diarrhoea given two levels of a particular factor were calculated in Stata Version 10 package. This aimed at foreshadowing findings to the third objective in this study before fitting the stated statistical models to the data.

3.8.3 Fitting Logistic, Poisson and Bayesian models to data

To achieve the objectives of this study, the logistic regression model (with logit link), the Poisson regression model (with log link), and the Bayesian semiparametric regression model were fitted to the data, using Stata Version 10 package for classical models, and BayesX package for the Bayesian model.

The results from logistic model are reported as odds ratios (ORs) of effects of levels of the factors on child diarrhoea together with their corresponding 95% CIs. The logarithms of expected diarrhoea cases under each factor, with their 95% CIs, are reported from the Poisson model. On the other hand, the results from the Bayesian method are reported as estimates of the posterior mean effects of factors on child risk, together with their corresponding 95% CrIs, and contextual non-parametric effects, with CrIs are reported for the non-linear variable age.

3.8.4 Comparative analysis for different models

The logistic and Poisson models were compared based on chi-square's goodness-of-fit test results. This answered the second objective stated in this study. The test was

preferred to the usual coefficients of determination (R^2) since the two models were non-nested and the data used was enumerative (counts) in nature. A goodness-of-fit test is a statistical test of how well the data at hand support an assumption about the distribution of a population or random variable of interest (Aczel and Sounderpandian, 2002). The test determines how well an assumed distribution fits the data. If the data are collected in a table of *k* cells with at least 5 counts per cell, and observed counts in cell *i* are denoted O_i while expected counts are denoted E_i, then the statistic,

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

has chi-square distribution with *k-1* degrees of freedom (that is, $E = np$ for a binomial random variable).

For a 1-tailed test, if the computed X^2 > chi-square ($k-1, \alpha$) from distribution tables, then the null hypothesis for a particular assumed distribution is rejected at α level, otherwise the null hypothesis is accepted. The closer the value observed in each cell to the expected value in that cell from the assumed distribution the higher the chances of accepting the distributional assumption of the model. Further, model adequacy statistics, such as pseudo-R^2 and parameter *p*-values, for individual models were studied before each model was compared with another.

The consistency of estimates between the Bayesian semiparametric model and either Binomial or Poisson model was compared through estimates for sizes of credible and confidence intervals. This answered the first objective in this study.

For the Bayesian model, the adequacy was checked via the Deviance Information Criterion (DIC) and posterior predictive checking was done via posterior credible intervals. The DIC is a generalization of the Akaike information criterion (AIC) and Bayesian information criterion (BIC), also termed Schwarz criterion. It is most applicable in Bayesian model selection problems where the posterior distributions of the models have been obtained through Markov chain Monte Carlo (MCMC) simulation.

The DIC is an asymptotic approximation as the sample size gets large, just like the AIC or BIC. It is only valid when the posterior distribution is approximately multivariate normal. Deviance can be defined as $D(\theta) = -2\log(p(y \mid \theta)) + C$, where $\boldsymbol{y}$ is the data, $\boldsymbol{\theta}$ are the unknown parameters of the model and $p(y \mid \theta)$ is the likelihood function. C is a constant that cancels out in all calculations that compare different models and, which therefore, does not need to be known. The expectation $\overline{D} = E[D(\theta)]$ is a measure of how well the model fits the data; the larger this is, the worse the fit. The effective number of parameters of the model is computed as $p_D = \overline{D} - D(\overline{\theta})$, where $\overline{\theta}$ is the expectation of θ. The larger this is, the better it is for the model to fit the data. The deviance information criterion is calculated as

$$DIC = p_D + \overline{D}$$

The idea is that models with smaller DIC should be preferred to models with larger DIC. Models are penalized both by the value of $\overline{D}$, which favors a good fit, but also (in common with AIC and BIC) by the effective number of parameters p_D. Since $\overline{D}$ will decrease as the number of parameters in a model increases, the p_D term compensates for

this effect by favoring models with a smaller number of parameters. Hence, DIC is a compromise between model fit and complexity (Mesele, 2009).

The advantage of DIC over other criteria, for Bayesian model selection, is that it is easily calculated from the samples generated by the MCMC simulation. AIC and BIC require calculating the likelihood at its maximum over θ, which is not readily available from the MCMC simulation. But to calculate DIC, simply compute $\overline{D}$ as the average of $D(\theta)$ over the samples of θ, and $D(\bar{\theta})$ as the value of D evaluated at the average of the samples of θ. Then the DIC follows directly from these approximations.

3.8.5 Checking randomness of outcome variable

The models were fitted with an assumption that the diarrhoea outcome variable, as well as the error resulting from fitting each parametric model was a random variable. This assumption had to be proved in the process of fitting the models. A procedure to employ depends on several factors, such as type of outcome variable (discrete or continuous), the way in which the data are observed and recorded (sequentially or not), and the nature of the study design (cluster or not), among others.

One simplest method used for a binary variable recorded sequentially and randomized individually is a nonparametric test called Runs Test for randomness. A run is a sequence of like elements that are preceded and followed by different elements or no element at all (Aczel and Sounderpandian, 2002). By arranging the diarrhoea cases and non-cases in the order they were recorded, it was easy to come up with the Runs, and, hence, the probabilities of obtaining any number of runs. The logic behind the Runs Test for

randomness is that if one obtains an extreme number of runs (too many or too few), then it can be decided that the elements in the sequence under study were not generated in a random fashion (Aczel and Sounderpandian, 2002). Thus, it sufficed to prove randomness using the Runs Test in this study.

The test was performed in StataSE 10 package with the assumption that data was recorded sequentially and randomized individually. A two-tailed hypothesis test that was conducted was as follows: $\boldsymbol{H_0}$: Diarrhoea observations were generated randomly *versus* $\boldsymbol{H_1}$: Diarrhoea observations were not randomly generated.

The test statistic is R = number of runs. The decision rule is to reject H_0 at level α, if $R \leq C_1$ or $R \geq C_2$. In this case, C_1 and C_2 are critical values obtained from cumulative distribution function $F(r)$ for the total number of runs R in samples of sizes n_1 for cases and n_2 for non-cases, with total tail probability $P(R \leq C_1 + R \geq C_2) = \alpha$.

3.9 Validity and reliability of estimates

The investigators in MICS study pre-tested the questionnaires during the month of June 2006 in Chichewa and Tumbuka speaking areas of the country and in both urban and rural settings. Based on the results of the pre-test, modifications were made to the wording and translation of the questionnaires (2006 MICS Report). This ensured internal validity of the findings that can be gotten using MICS data. The fact that random sampling techniques were used to collect MICS data, external validity as well as reliability of results can also be assumed.

However, in case the randomness test disapproves of assumption of randomness of diarrhoea variable then use of the Bayesian semiparametric model strengthens external validity and reliability of estimates from classical models where the results from the two types of models tallied, as the Bayesian model did not need the randomness assumptions.

3.10 Summary

This study applied statistical models of binomial, Poisson and Bayesian semiparametric to analyse two-week incidence variations of child diarrhoea in Malawi. The 2006 MICS data was used to that effect, with permission from NSO. The analysis of data was performed in Stata Version 10 package for the two classical models and in BayesX for Bayesian semiparametric model, with some descriptive statistics done in SPSS as well.

CHAPTER 4: RESULTS AND INTERPRETATIONS

4.1 Baseline analysis results

There were 22, 994 under-five children who were interviewed in the 2006 MICS study. A total of 15, 018 (65.3%) of these had complete information on all the studied variables and hence, their data was analysed in this study. The incomplete data was dropped based on randomness assumption. That is, dropped data points could produce similar results if analysed separately. Further, the large sample that remained ensured that dropping incomplete data points could not seriously distort the study findings.

The results presented in Table 1 show that the study involved almost equal numbers of female (50.4%) and male (49.6%) children. Further, it is shown that most of the studied children were in the age group 12-23 months (23%), with mean age of 28 months and a standard deviation of 16 months. In addition, a large proportion (56.1%) of the children was weaned. Furthermore, there were more children (87.1%) residing in rural areas. Likewise, most children (85.1%) had mothers whose highest education was primary. It is also indicated that most children studied (39.5%) were living in the southern region of Malawi. Besides, more children (62.2%) were living in families that were not sharing toilets. Similarly, a large proportion of the children (71.7%) were drinking water from piped source.

Finally, the results show that more children (53.1%) were residing along Lake Malawi and Shire Valley areas. In this respect, Lake Malawi and Shire Valley districts included Karonga, Rumphi, Nkhata-bay, Nkhotakota, Salima, Dedza, Mangochi, Balaka, Machinga, Zomba, Mwanza, Blantyre, Chikhwawa, Thyolo, and Nsanje. Whereas, Chitipa, Mzimba, Kasungu, Ntchisi, Dowa, Lilongwe, Mchinji, Ntcheu, Phalombe, Chiradzulu, and Mulanje were regarded as highland districts.

On incidence rate, the results indicate that out of the 15,018 children analysed 3,282 (21.85%) had diarrhoea at some time in two weeks preceding the survey. In addition, it is shown that the incidence rate was proportionally distributed in males (10.97%) and females (10.88%). Further, the rate was highest in age group 12-23 months (8.5%). It was also high in the breastfed children (13.4%). Furthermore, the rate was proportional between children who were living along Lake Malawi and Shire river valley (10.86%) and those from the highlands (10.99%).

Additionally, the rate was highest in central region of the country (9.36%) compared to the other two regions. Similarly, the rate was higher in children who were living in rural areas of the country (19.38%). Likewise, incidence was higher in children whose families were not sharing toilets (12.45%). Besides, the rate was highest in children whose mothers' highest education was primary (18.87%) compared to other studied levels of education. Finally, the incidence rate was higher in children who were drinking from piped water (15.28%) compared to other three sources.

Table 1: Baseline analysis results for child diarrhoea cases

Characteristic		Total (%)	Incidence (%)
Overall		15, 018 (100)	3,282 (21.85)
Gender:	Male	7,450 (49.61)	1,648 (10.97)
	Female	7,568 (50.39)	1,634 (10.88)
Age:	0-11	2,826 (18.82)	682 (4.54)
	12-23	3,458 (23.03)	1,277 (8.5)
	24-35	3,400 (22.64)	698 (4.65)
	36-47	3,054 (20.34)	399 (2.66)
	48-59	2,280 (15.18)	226 (1.5)
Breastfeeding:	Breastfed	6,585 (43.85)	2,013 (13.40)
	Weaned	8,433 (56.15)	1,269 (8.45)
Area of residence:	Rural	13,082 (87.11)	2,911 (19.38)
	Urban	1,936 (12.89)	371 (2.47)
Altitudinal locale:			
Lakeshore/riverine		7,981 (53.14)	1, 631 (10.86)
Highland		7,037 (46.86)	1,651 (10.99)
Region:	Northern	3,650 (24.30)	604 (4.02)
	Central	5,429 (36.15)	1,405 (9.36)
	Southern	5,939 (39.55)	1,273 (8.48)
Mother's education:	Primary	12,779 (85.09)	2,834 (18.87)
Secondary		2,165 (14.42)	437 (2.91)
	Higher	74 (0.49)	11 (0.07)
Family toilet:	Shared	5,670 (37.75)	1,413 (9.41)
Not shared		9,348 (62.25)	1,869 (12.45)
Drinking water source:	Piped	10,766 (71.69)	2,294 (15.28)
	Protected well	818 (5.45)	188 (1.25)
	Unprotected well	2,455 (16.35)	584 (3.89)
	Surface water	979 (6.52)	216 (1.44)

4.2 Cross-classification results

The results from Table 2 show that female children were as likely as male children to catch diarrhoea, although gender is not a significant factor in determining child's risk. Further, it is indicated that weaned children had 59.8% reduced odds of catching diarrhoea than children who were breastfed. The age variable results show that children aged 12-23 months had 84.1% higher odds of catching diarrhoea than those aged 0-11

months. While children aged 24-35; 36-47; and 48-59 months had respectively 18.8%; 52.8%; 65.4% reduced odds of catching diarrhoea compared to those aged 0-11 months.

Furthermore, it is shown that children from rural areas had 20.7% higher odds of catching diarrhoea than those from urban areas. Likewise, children who were living along the shores of Lake Malawi and Shire river banks had 16.2% reduced odds of catching diarrhoea than those from highlands. On region of stay, the results show that children who were living in the central and southern regions had respectively 76.1% and 37.6% higher odds of catching diarrhoea compared to those who were living in the northern region. In addition, a child from secondary educated mother had 11.3% reduced odds of catching diarrhoea than the one from a primary educated mother. But the results show no difference between odds of a child from primary educated and tertiary educated mother catching diarrhoea.

Besides, the results show that children whose families were sharing toilets had 32.8% increased odds of catching diarrhoea than those whose families were not. Finally, it is shown that children who were drinking from unprotected well had 15.3% increased odds of catching diarrhoea compared to those who were drinking from piped water. But there was no significant difference in odds of catching diarrhoea between children who were drinking from piped water and those drinking from protected well or surface water.

Table 2: Unadjusted Odds Ratios of child diarrhoea for selected predictors

Variable	Odds Ratio (OR)	95% CI of OR	P-value
Gender: ref (Male)	0.969	0.897-1.047	0.432
Breastfeeding: ref(breastfed)	0.402	0.371-0.436	<0.001
Age in months: ref(0-11)			
12-23	1.841	1.646-2.058	<0.001
24-35	0.812	0.72-0.916	<0.001
36-47	0.472	0.412-0.542	<0.001
48-59	0.346	0.293-0.408	<0.001
Area of residence: ref(urban)	1.207	1.07-1.362	0.002
Altitudinal locale: ref(highland)	0.838	0.775-0.905	<0.001
Region: ref(Northern)			
Central	1.761	1.582-1.96	<0.001
Southern	1.376	1.236-1.532	<0.001
Mother's education: ref(primary)			
Secondary	0.887	0.793-0.994	0.038
Higher	0.613	0.322-1.164	0.13
Family toilet: ref(not shared)	1.328	1.228-1.437	<0.001
Drinking water source: ref(piped)			
Protected well	1.102	0.93-1.305	0.26
Unprotected well	1.153	1.039-1.279	0.007
Surface water	1.045	0.893-1.224	0.58

4.3 Logistic and Poisson model results

The results, for a logistic regression model, presented in Table 3 show that the model as a whole fits the diarrhoea data significantly better than an empty model, that is, a model with no predictors (LR = 985.24, $p < 0.001$). However, chi-square's goodness-of-fit test result leads to rejection, at 5% level, of the binomial distribution assumption of total number of cases at any time of observation (GoF = 1019, p = 0.0015).

For Poisson model, the output for unconditional mean and variance of diarrhoea cases give mean of 0.2185 and variance of 0.1708. The values, though for unconditional mean and variance, indicate slight under-dispersion. However, the variance is not substantially smaller than the mean, $E(\Sigma(Y=1)) \approx var(\Sigma(Y=1) \approx 0.2$, and thus the predictor variables

could be of help. Further, using Microsoft Excel 'rand' function, random samples of 100, 1000, 5000, 10000 and 15000 generated from the diarrhoea variable produced prevalence rates of 0.27, 0.251, 0.242, 0.241, and 0.239 respectively, indicating that increasing sample size resulted in reduction of prevalence rate. So, it was reasonable to approximate binomial model with logit link by Poisson model with log link, but with robust standard errors to account for clustering of data. The results shown in Table 3 for Poisson model with robust (residual-based) standard errors, taking into account of the clustering, indicate that the model is significantly better than an empty model (LR=973, $p<0.001$). Further, the goodness-of-fit test is accepted at 5% level (GoF= 9225, $p = 1.00$), showing that the data give no statistical evidence that the diarrhoea cases does not follow Poisson distribution.

The estimates from Logit and Poisson models show that, adjusting for other factors, a weaned child had respectively 30.5% and 23.2% reduced odds and risk of catching diarrhoea compared to a breastfed child. In addition, the two models show that children who were living in the central region had respectively 67.5% and 47.2% higher odds and risk of catching diarrhoea than those who were living in northern region, adjusting for other factors. Likewise, children from southern region had respectively 36.5% and 27.2% adjusted higher odds and risk of catching diarrhoea compared to children from the north. Furthermore, it is indicated that odds and risk of catching diarrhoea increased by 27.3% and 19.2% respectively in children whose families shared toilets compared to those whose families did not, controlling for other factors.

The results also show that adjusted odds and risk of catching diarrhoea in children aged 12-23 and months were respectively higher by 92.8% and 57% than in children aged 0-11 months, while there was no difference in odds or risk between age group 24-35 and age 0-11 months. However, the adjusted odds and risk were respectively lower by 33.1% and 30% in children aged 36-47 and lower by 50.4% and 46.2% in age 48-59 compared to those aged 0-11 months. Similarly, both models showed that children living in families that shared toilets had 27.3% and 19.2 % respectively higher odds and risk of catching diarrhoea.

Finally, the two models showed no evidence of difference in adjusted odds and risk of catching diarrhoea between children living in rural and urban areas, lakeshore/riverine areas and highlands, primary educated and higher than primary educated mothers, and in children drinking from piped and other sources of drinking water. These results were in agreement with the crude OR reported before.

Table 3: Logit and Poisson model, adjusted OR and RR, results

Variable	Logit, OR (95%CI, p-value)	Poisson, RR (95%CI, p-value)
Breastfeeding: ref(breastfed)	0.695 (0.6-0.8, p<0.001)	0.768 (0.693 - 0.85, p<0.001)
Age in months: ref(0-11)		
12-23	1.928 (1.17-2.16, p < 0.001)	1.57 (1.452-1.699, p<0.001)
24-35	1.09 (0.93-1.29, p = 0.29)	1.055(0.936-1.19, p=0.378)
36-47	0.669 (0.55-0.81, p < 0.001)	0.7(0.604-0.813, p<0.001)
48-59	0.496 (0.4-0.61, p < 0.001)	0.538(0.453-0.637, p<0.001)
Area of residence: ref(urban)	1.122 **(0**.985, 1.28**,** p=0.08)	1.09(0.989-1.202, p=0.083)
Altitudinal locale: ref(highland)	0.918 (0.84, 1.002, p=0.055)	0.939(0.881-1.001, p=0.052)
Region: ref(Northern)		
Central	1.678(1.495-1.88, p<0.001)	1.472 (1.348-1.608, p<0.001)
Southern	1.365(1.22-1.528, p<0.001)	1.272 (1.167-1.387, p<0.001)
Mother's education: ref(primary)		
Secondary	0.922 (0.818, 1.04, p=0.185)	0.941(0.861-1.029, p=0.182)
Higher	0.783 (0.404, 1.52, p=0.47)	0.822 (0.485-1.395, p=0.468)
Family toilet: ref(not shared)	1.273 (1.17-1.38, p<0.001)	1.192 (1.123-1.266, p<0.001)
Drinking water source: ref(piped)		
Protected well	0.997(0.84, 1.19, p=0.97)	0.995 (0.878-1.129, p=0.942)
Unprotected well	1.036(0.93, 1.16, p=-0.53)	1.025 (0.947-1.109, p=0.539)
Surface water	1.084(0.92, 1.28, p=0.343)	1.063 (0.943-1.2, p=0.321)
Overall model fit	GoF=1019, p=0.002; LR=985, p<0.001	GoF=9225, p=1.00; W=973, p<0.001

4.4 Runs Test for Randomness results for diarrhoea variable

The results from Runs Test for Randomness of the diarrhoea outcome variable analysed in Stata Version 10 for *n* = 15, 018, using either continuity or split mean as cut-off points, with or without continuity correction produced the number of runs statistic, r = 4, 963 (z = -3.99, p < 0.0001). Hence, the data provide no evidence, at 5% level of error, that the diarrhoea observations were generated in a random way. This was expected as 2006 MICS sampling was done at cluster level and not individual level of a child. Since the

analysis of the data is done list wise, it is very likely to find that observations are not a random sample viewed from case by case situation rather than cluster by cluster.

4.5 Bayesian semiparametric model results

The results presented in this section were run in BayesX package Version 2.0.1, using the following code:

> *b.regress ca1 = cage_11(psplinerw2) + bf2 + ed3a + ws8 + water3 + dist3 + hh6 + ufreg, family=binomial iterations=12000 burnin=2000 step=10 predict using d.*

The Markov chain Monte Carlo (MCMC) simulations were run on the set of full conditional posterior distributions in order to derive the full posterior estimates for all the parameters of interest (see Ferreira da Silva, 2010c). The options *iterations, burnin* and *step* define the total number of iterations, the burn in period, and the thinning parameter of the MCMC simulation run (Brezger, Kneib and Lang, 2005). Specifying *step=10* as above forces BayesX to store only every 10th sampled parameter which leads to a random sample of length 1000 for every parameter in this case. Therefore, a sample of 10000 random numbers is obtained with the above specifications. It should be noted that the choice of iterations also affects computation time.

4.5.1 Bayesian model, fixed-effect results

The model presented in Table 4 has the following estimation results for the DIC: DIC based on the un-standardized deviance results are; Deviance (bar_mu) = 14847.347, pD =

11.887, DIC = 14871.121, and DIC based on the saturated deviance results are; Deviance (bar_mu) = 14847.347, pD = 11.887, DIC = 14871.121.

The effects displayed in Table 4 show that posterior mean amount of diarrhoea cases were expected to be low in weaned children, in children whose mother's education was higher than primary, and in children who lived close to Lake Malawi or Shire River. However, the posterior mean amount of cases were expected to be high in children whose families were sharing toilets, in children who were drinking from non-piped water source, in children who were living in rural areas, and in children who were living in other regions than northern region. These results were supported by direct fixed-effects results for each categorical variable that were analysed in BayesX as well.

Finally, it is clear that the 80% credible interval indicates significance of all variables studied. While the 95% credible interval shows that mother's education, source of drinking water and area of residence were not significant factors for determining a child' risk to diarrhoea. These results once again agree with those from logit and Poisson models.

Table 4: Bayesian model, fixed-effects results

Variable	Posterior mean	95% CrI	80% CrI
Constant	-1.137	-1	-1
Breastfeeding: ref(breastfed)	-0.376	-1	-1
Mother's education: ref (primary)	-0.111	0	-1
Family toilet: ref(not shared)	0.258	1	1
Drinking water source: ref(piped)	0.029	0	1
Altitudinal locale: ref(highland)	-0.202	-1	-1
Area: ref(urban)	0.110	0	1
Region: ref(northern)	0.123	1	1

4.5.2 Bayesian model, non-linear effects results

From Table 5, it is clear that expected posterior mean cases of diarrhoea was low in age groups 0-11 months, 36-47 months, and 48-59 months, but high in age groups 12-23 months and 24-35 months. However, the 95% credible intervals show that age groups 0-11 and 24-35 months have no significant effects. But the most vulnerable age group to diarrhoea is 12-23 months as found in logit and Poisson models.

Table 5: Bayesian model, non-linear effects results

Age group in months	Posterior mean	95% CrI	80% CrI
0-11	-0.016	0	0
12-23	0.624	1	1
24-35	0.078	0	1
36-47	-0.411	-1	-1
48-59	-0.714	-1	-1

CHAPTER 5: DISCUSSION OF RESULTS

5.1 Introduction

This study sought to establish consistency of estimates found using Bayesian semiparametric model and classical models, as well as comparing the classical models used. This was achieved by fitting the Bayesian semiparametric additive model, logistic and Poisson regression models to the 2006 MICS diarrhoea data which was collected by National Statistical Office of Malawi between June and December of that year with an aim of estimating key indicators of child and women health in each district. The analyses were done in SPSS, Stata, and BayesX packages as earlier alluded to.

5.2 Consistency of estimates found by Bayesian and Logit/ Poisson models

The results presented in Chapter 4 have shown that significance and direction of estimates from Bayesian semiparametric model and Poisson or logit model were generally similar. The exception is in closeness to lake/river variable which was found to be statistically significant using the Bayesian semiparametric model but insignificant factor using logit or Poisson models. The three models have coincidentally ruled out usefulness of mother's education, area of residence (rural or urban), and source of drinking water in determining child diarrhoea.

5.3 Classical models' comparison

The chi-square's goodness-of-fit tests' results presented in Chapter 4 suggest that Poisson log-linear regression model, with robust standard errors, fits the diarrhoea data set well than the logistic regression model. This was expected as unlike the logistic model, Poisson model with robust standard errors takes into account household correlations due to clustering of data.

5.4 Risk factors for child diarrhoea

The results presented in Chapter 4 suggest that gender of a child has little (if any) to do with a child's risk to diarrhoea as female children were as likely as male children to catch diarrhoea. This may imply that the biological make-up of a child's body gives no bias or advantage to any gender in terms of likelihood of catching diarrhoea.

Further, it has been found out that breastfeeding status of a child is a useful factor in determining a child's risk to diarrhoea. Thus, weaned children were found to have lower chances of catching diarrhoea than still breastfeeding children. This may reflect low possibilities of gastro transmission in a weaned child who chooses what to put to the mouth independent of the mother. It may also reflect on low hygiene considerations in breastfeeding mothers when giving food items to the breastfeeding babies in the country.
In addition, age of a child was found to be a useful factor in estimating child's risk to diarrhoea. To that effect, age group 12-23 months has been found to be the most risky group to diarrhoea compared to all other age groups studied. The results also suggest that the risk is lower in age 0-11 months and after 23 months of a child's life.

These variations across age groups may reflect breastfeeding stages of a child. For instance, the weaning time which is reported to pose more threats of diarrhoea attacks to a child (Kourtis et al, 2007) is around 17.6 months (Kazembe, 2008), that is, age group 12-23 months spans weaning time. Further, the low risk in age 0-11 months may reflect the fact that the data had a mixture of exclusively predominantly breastfed children who are reported to be at low risk of morbidity and mortality due to diarrhoea (Arifeen et al, 2001; Betran, 2001; WHO, 2000; Yoon, 1996; Hanson, 1994; Victora, 1992) and the general breastfed children. The 2006 MICS, whose data was analysed in this study, reported that approximately 56% of children aged less than 6 months were exclusively breastfed. Then the observed low risk of catching diarrhoea in age group 0-11 months, which overlaps age 0-6 months, reflect the high percentage of exclusively breastfed children analysed in the study.

However, the results show a shift of most risky age group upward from age 6-11 months reported in 2004 Malawi DHS and age 6-8 months reported by Kandala et al (2008) for the 1999 and 2003 Nigerian DHSs to age 12-23 months reported in this study. The shift seems to mimic the trend of diarrhoea in breastfeeding children reported recently by researchers in Malawi. Although this study was not intended to explore interaction of studied factors, there seems to be an interaction between child's age and breastfeeding status. Studies in Malawi have shown an increase in diarrhoea during and following weaning time among exclusively breastfed infants reportedly weaned at 6 months (Clayden, 2007). The fact that weaning time in Malawi is around 17.6 months (Kazembe, 2008) which is within 12-23 months age group then these results are not a surprise. As

per tradition, weaning time entails introduction of complementary infant foods which, may in turn spread diarrhoea to the child if not hygienically prepared by the mother. This is why researchers have recommended that greater emphasis should be placed on hygienic preparation of weaning foods and water purification in order to decrease infant diarrhoeal morbidity in resource-limited settings (Kourtis et al, 2007).

Furthermore, region of stay has been found to be a significant factor for determining child's risk to diarrhoea. Thus, the results suggest that children from the central and the southern regions are at higher risk compared to those from the northern. Compared with the southern region, children from central region have higher chances of catching diarrhoea. The causes of such differences can be far from speculation. However, the findings agree with the 2004 Malawi DHS results and a study report by Kazembe et al (2009).

Likewise, the findings have shown that children whose mothers' highest education qualification is secondary have marginally lower chances of catching diarrhoea compared to those with primary educated mothers. But the findings suggest no difference in the risk of catching diarrhoea between children with primary educated mothers and those with tertiary educated mothers, as well as between children with secondary educated mothers and tertiary educated mothers. However, mother's education was found to be statistically insignificant factor for determining child diarrhoea. This may reflect the way the study was designed, which just sought differences in academic qualification and not in health education of mothers.

Although other studies in sub-Saharan Africa have supported influence of mother's academic qualification on a child's chances of catching diarrhoea (see Kandala et al, 2008), there cannot be any immediate reason as to why one can think that mere differences in levels of academic or formal education achievements (other than health education) can result in differences in child's risk to diarrhoea. No wonder there was no difference in effects between secondary and tertiary education, the same results could be expected if levels of tertiary education were compared (for instance, diploma and degree). What is felt to have an effect on child's health is the mother's knowledge in health, which is richly provided in the primary education curriculum in Malawi. But also, mere health education literacy of the mother provided through attendance of antenatal or postnatal care services could serve the purpose of controlling child's health.

Besides, the findings have indicated that children from rural areas have high chances of catching diarrhoea compared to those from urban areas, although area of residence was found to be statistically insignificant factor in determining a child's risk. This agrees with results from the 2004 Malawi DHS. The situation may reflect low rates of exclusive breastfeeding practices in rural areas of the country. It is reported that exclusive breastfeeding reduces diarrhoea threats in under-five children (WHO, 2000). A study report by Kerr et al (2007) has indicated that only 4% of Malawian children are exclusively breastfed for 6 months in rural areas of Ekwendeni, Mzimba district. Thus, a majority of mothers living in rural areas of Ekwendeni do not practice exclusive breastfeeding during the first 6 months of a child's life. If the situation is true in all rural parts of the country, then the high risk to diarrhoea findings for rural children noted in this study may not be a surprise.

Furthermore, the results have found closeness to lake/river as a useful factor in determining a child's risk to diarrhoea. The findings suggest that children living along the Lakeshore or the river banks have reduced chances of catching diarrhoea compared to those from highlands. The opposite was expected, but these results may reflect high utilization of water sanitation interventions rolled by government and other stakeholders, such as free water guard in drinking water and improved drinking water sources-such as piped water and boreholes (Kumwenda, 2009), targeted to lakeshore/riverine dwellers in recent years who were previously believed to be at high risk of catching diarrhoea than highlanders. Thus, high use of safe and clean drinking water by residents of lakeshore or shire valley has reversed the old trend of child diarrhoea cases between highlands and lakeshore areas.

The findings also suggest that there is no difference in tendencies of catching diarrhoea in children who drink from protected well and surface water to those who drink from piped water. But children who drink from unprotected well were found to have marginally increased chances of catching diarrhoea compared to those drinking from piped water. However, source of drinking water was found to be statistically insignificant factor in determining a child's risk to diarrhoea. But the findings may reflect splash effects of the water sanitation interventions projects, such as free water guard, which were underway in many parts of the country around or during the time of MICS study which could not bring significant differences in diarrhoea cases in children who were drinking from different water sources.

Finally, the findings suggest family toilet facility is a useful factor in estimating a child's risk to diarrhoea. Thus, children from families that shared toilets were found to have increased chances of catching diarrhoea than those whose families did not share toilets. This may reflect high possibilities of gastro transmission from other people who use the same toilet as the child or her mother.

5.5 Summary

The study has revealed that most findings from both classical models were consistent with those from the Bayesian model. However, the analysis has ruled out the binomial distributional assumption about child diarrhoea data but supported Poisson assumption.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The findings suggest that estimating child's risk to diarrhoea using Bayesian semiparametric model is as good as using logistic or Poisson model. This is the case since the two groups of models have agreed in isolating most significant as well as insignificant factors for determining the child's risk to diarrhoea. But of the two classical models used on the data, the goodness-of-fit of Poisson regression, with robust standard errors, is better than logistic model.

It can further be concluded that the region from which a child comes (northern, central, or southern), age of a child, whether or not a child is still breastfeeding, whether or not a child comes from a family that shares toilet with other families, and closeness to lake/river are statistically significant factors in determining likelihood of the child suffering from diarrhoea. But, under-five child diarrhoea has little (if any) to do with area of residence (rural or urban), source of drinking water, and mother's education.

6.2 Implications of findings

The findings suggest that applying Bayesian semiparametric models together with classical models can help to confirm classical model estimates or this can provide

alternative estimates which can be trusted when one is not sure of level of satisfaction of classical model assumptions. Thus, various approaches to data analysis should be seen as complementary to one another rather than as competitors for outright domination (Julian Besag in McCullagh, 2002).

6.3 Limitations of the study

The study sample had 7,976 (34.69%) children with missing values in at least one variable of interest. This may have influenced the results in this study in one way or another as the nature of variability of the dropped data was not known, but was just assumed to be random.

In addition, the study did not exhaust all possible models for the diarrhoea data since it was just an application study on use of statistical models in explaining under-five diarrhoea incidence. It is important to mention that other models, such as Negative Binomial, Generalized Estimating Equations were possible, especially in situations where serious under-dispersion could be noted when fitting the Poisson model and where inter-cluster correlations were possible.

The study findings on actual epidemiology of under-five child diarrhoea incidence in the Malawian population may not be accurate since the survey data used are from 2006 which is not the most current one in the country. Thus, focus of the study was on whether a statistical model can be used to explain/predict the likelihood of a child suffering from diarrhoea rather than on whether the study findings on diarrhoea situation in the country reflect the true current situation on the ground. Thus, much attention was on formal

theory of the applied statistical models and their practical outcomes rather than on diarrhoea findings.

The models applied were not extended to capture seasonality of child diarrhoea in Malawi although it is obvious that findings on seasonality of the disease could add more meaning to the study, as study reports have indicated that there is a higher probability of infant diarrhoea in the rainy, compared to the dry season in Malawi (Clayden, 2007).

Finally, the households selected per each cluster were systematically sampled in the MICS study and the district was regarded as the universe, but data in this study was analysed at an aggregated national level. Moreover, the Runs Test for randomness ruled out randomness of the diarrhoea variable. This may have biased the results in one way or the other in this study. However, consideration of clusters using district as the universe of sampling in the MICS study made it difficult to find a shortcut way of proving randomness of diarrhoea outcome variable at national level standpoint in this research, as randomness at unit level may not necessarily imply randomness at aggregated level for some distributions other than the normal.

6.4 Recommendations

6.4.1 Bayesian semiparametric regression models should be employed in parallel with classical models as a checking tool when the researcher is in doubts of meeting classical model assumptions.

6.4.2 Researchers should consider fitting Poisson regression model with robust standard errors when analysing child diarrhoea data that is randomized at cluster level compared to ordinary logistic regression model.

6.4.3 More interventions in child diarrhoea are needed in central region of the country by government and other stakeholders in health in order to contain the problem in the region.

6.4.4 MoH and other stakeholders should continue mobilising for high hygiene practices in breastfeeding mothers in the country, especially around weaning period.

6.4.5 MoH and other stakeholders may initiate campaign for independent family toilets in the country as child diarrhoea is associated with sharing of toilets.

6.4.6 There is need for another study that may try to find causes of high risk to diarrhoea in children from central region of Malawi.

REFERENCES

Aczel, A. D., & Sounderpandian, J. (2002). *Complete business statistics, (5th ed.)*. New York: McGraw-Hill.

Adèr, H. J. (2008). Chapter One: Modelling. In H. J. Adèr & G. J. Mellenbergh (Eds.). *Advising on Research Methods: A consultant's companion* (pp. 271-304). Huizen, The Netherlands: Johannes van Kessel Publishing.

Agrest, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley.

Albert, J. (1995). Teaching inference about proportions using Bayes and discrete models. Bowling Green State University: *Journal of Statistics Education, 3*(3).

Albert, J. (2007). *Introduction to Bayesian thinking*. (Unpublished lecture notes).

Arifeen, S., Black, R. E., Antelman, G., Baqui, A., Caulfield, L., & Becker, S. (2001). Exclusive breastfeeding reduces acute respiratory infection and diarrhoea deaths among infants in Dhaka slums. *Pediatr,* 108, 67.

Banerjee, S. (2011). *MCMC algorithms for fitting Bayesian models*. University of Minnesota. (Unpublished lecture notes)

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis, (2nd ed.)*. New York: Springer.

Bernado, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: John Wiley.

Betran, A. P., Onis, M., Lauer, J. A., & Villar, J. (2001). Ecological study of effects of breastfeeding on infant mortality in Latin America. *Br Med J,* 323, 1–5.

Boadi, K. O. & Kuitunen, M. (2005). Childhood diarrheal morbidity in the Accra Metropolitan Area, Ghana: socio-economic, environmental and behavioral risk

determinants. *Journal of Health and Population in Developing Countries (ISSN 1095-8940).*

Brezger, A., Kneib, T., & Lang, S. (2005). BayesX: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software,* 14(11).

Burton, L. M., Kemp, S. P., Leung, M., Matthews, S. A., & Takeuchi, D. T. (2011). *Communities, neighborhoods, and health: Expanding the boundaries of place*. New York: Springer Science + Business Media.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. New York: CRC Press LLC.

Clayden, P. (2007). Diarrhoea in uninfected infants of HIV-positive mothers who stop breastfeeding at 6 Months. *HIV i-Base*.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models, (2nd ed.).* New York: Chapman & Hall/CRC

Faraway, J. J. (2006). *Extending the linear model with r generalized linear, mixed effects and nonparametric regression models.* New York: Taylor & Francis Group.

Ferreira da Silva, A. (2010). cudaBayesreg: Bayesian computation in CUDA. *The R Journal, 2*(2).

Gelman, A. (2005). *Multilevel (hierarchical) modelling: what it can and can't do*. New York: Columbia University.

Gelman, A., & Hill, J. (2011). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis, (2nd ed.)*. New York: Chapman & Hall/CRC.

Hamer, H. D, Simon, J., Thea, D., & Keusch, G. T. (1998). Childhood diarrhoea in sub-Saharan Africa. *Child Health Research Project Special Report, April 1998.*

Hansen, M. H., Hurwitz, W. N. & Madow, W. G. (1996). *Sample survey methods and theory, vol. II*. New York: John Wiley.

Hanson, L. A., Ashraf, R., Zaman, S., et al (1994). Breast feeding is a natural contraceptive and prevents disease and death in infants, linking infant mortality and birth rates. *Acta Paediatr,* 83, 3–6.

Index Mundi (2011). *Malawi age structure-demographics*. Accessed on February, 10, 2012 from: http://www.indexmundi.com/malawi/population.html.

Jerak, A., & Wagner, S. (2003). *Estimating probabilities of EPO patent oppositions in a Bayesian semiparametric regression framework.* Department of Statistics, Munich: University of Munich.

Kandala, N. B., et al (2001). Semiparametric analysis of the socio-demographic determinants of under nutrition in two African countries. *Research in Official Statistics, EUROSTAT, 4*(1), 81-100.

Kandala, N. B., Ji, C., Stallard, N., Stranges, S. & Cappuccio, F. P. (2008). Morbidity from diarrhoea, cough and fever among young children in Nigeria. *Annals of* Tropical *Medicine & Parasitology, 102*(5), 427-445.

Kazembe, L. N. (2008). Spatial Modelling of Initiation and Duration of Breastfeeding: Analysis of Breastfeeding Behaviour in Malawi – I. *World Health Population. 10*(3), *14-31.*

Kazembe, L. N., Muula, L. S., & Simoonga, C. (2009). Joint spatial modelling of common morbidities of childhood fever and diarrhoea in Malawi. *Health & Place, 15*(1), 165-172.

Kerr, R. B., Berti, P. R., & Chirwa, M. (2007). Breastfeeding and mixed feeding practices in Malawi: timing, reasons, decision makers, and child health consequences. *Food and Nutrition Bulletin, 28*(1).

Kleinbaum, D. G., & Klein, M. (2005). *Survival analysis – A self-learning text (2nd ed.)*. NY: Springer Science + Business Media, Inc.

Kleinbaum, D. G., & Kupper, L. L. (1978). *Applied regression analysis and other multivariable methods.* Belmont, CA: Wadsworth Publishing Company, Inc.

Kourtis, A., Fitzgerald, D., Hyde, L. et al (2007). Diarrhoea in uninfected infants of HIV-infected mothers who stop breastfeeding at 6 months: the BAN study experience. *14th CROI, Los Angeles. Abstract 772*.

Kumwenda, S. (2009). *Assessment of water guard use at household level in Chikwawa district*. Masters' Thesis, University of Malawi, College of Medicine.

Lawson, A. B. (2009). *Bayesian Disease Mapping-Hierarchical Modelling in Spatial Epidemiology.* New York: Taylor & Francis Group.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Thousand Oaks: Sage Publications.

Malawi's Ministry of Gender, Youth and Community Services (2003). *National policy on early childhood development*. Lilongwe: Ministry of Gender, Youth and Community Services.

Mccullagh, P. & Nelder, J. A. (1989). *Generalized linear models, (2nd ed.)*. NY: Chapman and Hall.

McCullagh, P. (2002). What is a statistical model? *The Annals of Statistics*, *30*(5), 1225–1310.

Medical News Today (December, 2011). *What is diarrhoea? What causes diarrhoea?* Accessed on December 30, 2011 from: http://www.medicalnewstoday.com/articles/158634.php.

Mesele, T. (2009). *Bayesian approach to identify predictors of children nutritional status in Ethiopia.* Unpublished masters thesis, Addis Ababa University, Addis Ababa.

Munthali, A. C. (2005). Change and continuity in the management of diarrhoeal diseases in under-five children in rural Malawi. *Malawi Med Journal*, 16(2), 43-46.

Mwambete, K. D. & Joseph, R. (2010). Knowledge and perception of mothers and caregivers on childhood diarrhoea and its management in Temeke Municipality, Tanzania. *Tanzania Journal of Health Research, 12(1).*

National Statistical Office of Malawi & UNICEF-Malawi (2008). *Monitoring the Situation of Children and Women: Malawi Multiple Indicator Cluster Survey 2006, FINAL REPORT*. Zomba & Lilongwe, Malawi: UNICEF & NSO.

National Statistical Office of Malawi (2005). *Malawi Demographic and Health Survey, 2004.* Calverton, MD: ORC Macro.

National Statistical Office of Malawi (2008). *Population and Housing Census, 2008*. Zomba: NSO.

PATH (2011). Charting the course for integrated diarrhea control in Malawi: A way forward for policy change. Washington, DC 20001 USA: PATH.

Ridall, G. (2007). *Bayesian Inference*. (Unpublished lecture notes).

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of complex surveys*. Chichester: Wiley.

Smith, A. F. M. (1984). Present position and potential developments: some personal views; Bayesian statistics. *J. Roy. Statist. Soc. Ser. A* 147, 245–259.

Spanos, A. (2003). *Probability theory and statistical inference: Econometric modelling with observational data.* New York: Cambridge University Press.

UBM Media (NZ) Ltd. (July 2009). What is diarrhoea? Accessed on November, 10, 2011 from: http://www.everybody.co.nz/.

UNICEF-Malawi (2010). *Health and Nutrition Issue*. Lilongwe: UNICEF. Accessed on November 15, 2011 from: http://UNICEF-Malawi/2010/Noorani.

United Nations Population Information Network (POPIN) (2011). *Guidelines on Tracking Child and Maternal Mortality*. Geneva: UN.

Victora, C. G., Huttly, S. R., Fuchs, S. C., et al (1992). Deaths due to dysentery, acute and persistent diarrhoea among Brazilian infants. *Acta Paediatr suppl,* 381, 7–11.

WHO (2000). Collaborative study team on the role of breastfeeding on the prevention of infant mortality, effect of breastfeeding on infant and child mortality due to infectious diseases in less developed countries: A pooled analysis. *Lancet,* 355, 451–5.

WHO, UNICEF, UNFPA & World Bank (2007). *Maternal and infant mortality estimates, 2005*. Accessed on July 26, 2011 from: http://www.who.int/making-pregnancy-safer/topics/maternal-mortality/en/.

Yoon, P. W., Black, R. E., Moulton, L. H., & Becker, S. (1996). Effects of not breastfeeding on the risk of diarrhoea and respiratory mortality in children under 2 years of age in Metro Cebu, The Philippines. *Amer J Epidemiol*, 143, 1142–8.