

Bayesian Generalized Linear Mixed Modeling of breast cancer data in Nigeria



UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Ropo Ebenezer Ogunsakin

November, 2017

Bayesian Generalized Linear Mixed Modeling of breast cancer data in Nigeria

by

Ropo Ebenezer Ogunsakin

A thesis submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements for the degree
of
DOCTOR OF PHILOSOPHY
in
STATISTICS

Thesis supervisor: Siaka Lougue (PhD)



**UNIVERSITY OF
KWAZULU - NATAL**

**INYUVESI
YAKWAZULU-NATALI**

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Ropo Ebenezer Ogunsakin, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then
 - (a) their words have been re-written but the general information attributed to them has been referenced, or
 - (b) where their exact words have been used, then their writing has been placed in italics and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

Ropo Ebenezer Ogunsakin (Student)

Date

Siaka Lougue (PhD) (Supervisor)

Date

Disclaimer

This document describes work undertaken as a PhD programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

Abstract

Breast cancer is the world's most prevalent type of cancer among women. Statistics indicate that breast cancer alone accounted for 37% out of all the cases of cancer diagnosed in Nigeria in 2012. Data used in this study are extracted from patient records, commonly called hospital-based records, and identified key socio-demographic and biological risk factors of breast cancer. Researchers sometimes ignore the hierarchical structure of the data and the disease when analyzing data. Doing so may lead to biased parameter estimates and larger standard error. That is why the analyses undertaken in this study included the multilevel structure of cancer diagnosis, types, and medication through a Generalized Linear Mixed Model (GLMM) which consider both fixed and random effects (level 1 and 2). In addition to the classical statistics approach, this study incorporates the Bayesian GLMM approach as well as some bootstrapping techniques. All the analyses are done using R or SAS for the classical statistics approaches, and WinBUGS for the Bayesian approach. The Bayesian analyses were strengthened by advanced analyses of convergence and autocorrelation checks, and other Markov Chain assumptions using the CODA and BOA packages. The findings reveal that Bayesian techniques provide more comprehensive results, given that Bayesian analysis is a more statistically strong technique. The Bayesian methods appeared more robust than the classical and bootstrapping techniques in analyzing breast cancer data in Western Nigeria.

The results identified age at diagnosis, educational status, grade tumor, and breast cancer type as prognostic factors of breast cancer.

Breast cancer, Bayesian, Bootstrapping, GLMM, Hospital, Multilevel, CODA/BOA.

Acknowledgements

First and foremost, my utmost gratitude goes to ALMIGHTY GOD, who in His infinite mercies inspired the conception of this research work and made it possible for the entire research work to be a great success despite all daunting odds. I would like to thank the many people who have supported me in so many different ways in bringing this research into being.

I would like to express my deepest gratitude to my supervisor, Dr. Siaka Lougue, who has given me much support and inspiration and for sharing with me his knowledge, for the invaluable guidance he offered me and for his encouragement. My thesis would not be as refined without his suggestions for improvement. My training has benefited greatly from his expertise, enthusiasm and encouragement. I would also like to thank other members of the Statistics Department University of Kwazulu Natal who have supported me along the way, in particular Danielle Roberts and Aler-shnee Pillay.

I would like to thank my wife and my wonderful boy Bernard, for providing their past, present and future support. Without this support, I would not be able to study at this prestigious University. I would also like to thank the University of Kwazulu Natal, for relieving my parents' financial burden by way of a fee remission.

I would also like to convey thanks to the Cancer registry of federal medical teaching hospital, Ido Ekiti, Ekiti State, Nigeria for their cooperation and permission to use the real data. I obligated to them every greeting and respect.

I am grateful to my fellow PhD students I shared this journey with, especially those in Oliver Tambo building, Westville Campus, UKZN.

Wonderful appreciation goes to all members of Deeper Life Campus Fellowship, University of Kwazulu Natal, Westville Campus for their love and prayer throughout this journey. God will not forget the labour of their love in Jesus name.

A very special thank goes to my family especially Prof Femi Tinuola. Although not exactly familiar with my scientific way of life, his lifelong encouragement and unconditional support financially always kept me on track and taught me what being true to yourself really means. Words cannot convey my gratitude. In particular, deepest thanks are given to Jesus my Lord for giving me such chance in pursuit of

my dream in a rigorous competition.

Finally, I appreciate the moral support of all my friends, particularly Evans, Precious Iwalaye and encouragement through this process all the way from beginning to the end. Thank You for everything!

Contents

	Page
List of Figures	ix
List of Tables	xi
Abbreviations	xii
Chapter 1: INTRODUCTION	1
1.1 Background	1
1.1.1 Breast Cancer	2
1.2 Research Problem	5
1.3 Aim and Objectives	7
1.4 Significance of the Study	7
1.5 Thesis Layout	8
1.5.1 Thesis Contributions	9
Chapter 2: LITERATURE REVIEW	11
2.1 Context of the Study	11
2.1.1 Area of Study	12
2.1.2 Global Prevalence of Breast Cancer	12
2.2 Breast Cancer Risk Factors	17
2.2.1 Socio-Demographic Factors	18
2.2.2 Medical factors	19
2.2.3 Context of Breast Cancer in Nigeria	20
2.2.4 Epidemiology of Breast Cancer	22
Chapter 3: DATA AND RESEARCH METHODOLOGY	25
3.1 Data	25
3.2 Classical and Bayesian Logistic Regression Model	26
3.2.1 Maximum Likelihood Estimate (MLE) of the parameters	27

3.2.2	Other Classical Logistic Regression Statistics	32
3.2.3	Assumptions on Model Adequacy	35
3.2.4	Bayesian Prior Distributions	36
3.2.5	Bayesian Posterior Distribution via Markov Chain Monte Carlo	39
3.2.6	Markov Chain Monte Carlo Simulation	46
3.2.7	Some key points derived from the classical statistics	49
3.2.8	Some key points derived from the Bayesian inference	50
3.3	Advantages and Disadvantages of the Bayesian Approach	50
3.3.1	Advantages of Bayesian Statistics over Classical Statistics	50
3.3.2	Advantages of Classical Statistics over Bayesian Statistics	51
3.4	Generalized Linear Models	51
3.4.1	Model Formulation	52
3.4.2	Parameter Estimation	54
3.5	Family of Binomial Logit Link Function with Hierarchical Structure	55
3.5.1	Multilevel Generalized Linear Models	56
3.5.2	Multilevel Logistic Regression Model	57
3.6	The Multinomial Logistic Regression Model	58
3.6.1	Model and Parameter Estimation in Multinomial	61
3.7	Bayesian Generalized Linear Mixed Model (GLMMs)	62
3.7.1	Generalized Linear Mixed Model	63
3.8	Estimation Methods	65
3.8.1	Maximum Likelihood Estimation	65
3.8.2	Bayesian Multilevel	65
3.8.3	Priors for Bayesian Multilevel Models	67
Chapter 4:	RESULTS	69
4.0.1	Descriptive Analysis Result of Breast Cancer	69
4.1	DETERMINANT OF BREAST CANCER TYPES	73
4.1.1	Introduction	73
4.2	Materials and Methods	75
4.2.1	Data Collection	75
4.3	Results	76
4.3.1	Descriptive analysis of breast cancer	76
4.3.2	Result of classical logistic regression	77
4.3.3	Result of Bayesian logistic regression model	77
4.3.4	Assessing the performance of Markov Chain Monte Carlo (MCMC) chains in WinBUGS	79

4.3.5	Discussion and partial conclusion	84
4.4	DIAGNOSTIC DETERMINANT OF PREFERRED CANCER TREATMENT	87
4.4.1	Introduction	88
4.4.2	Materials and Methods	89
4.4.3	Ethics	89
4.5	Results	90
4.5.1	Socio-demographic characteristics of participants	91
4.5.2	Result of the classical multilevel logistic regression	91
4.5.3	Result of Bayesian multilevel model with non-informative prior	92
4.5.4	Discussion and partial conclusion	97
4.6	MULTILEVEL MULTINOMIAL LOGIT REGRESSION MODEL WITH RANDOM EFFECTS	101
4.6.1	Introduction	101
4.6.2	Methods	102
4.6.3	Study participants and methods	102
4.6.4	Ethical considerations	103
4.6.5	Descriptive statistics	104
4.6.6	Result of classical multilevel multinomial model	104
4.6.7	Bayesian Estimation for Multinomial Logistic Regression	106
4.6.8	Bayesian multilevel multinomial results	110
4.6.9	Discussion and partial conclusion	111
4.7	SOCIO-ECONOMIC DETERMINANTS OF BREAST CANCER RISK FAC- TORS IN WESTERN NIGERIA: A MULTINOMIAL MODEL	114
4.8	Introduction	115
4.8.1	Ethical Approval	116
4.8.2	Methods	116
4.8.3	Model and Parameters Estimation in Multinomial	116
4.8.4	Bootstrapping technique	118
4.8.5	Implementations of Markov chain Monte Carlo (MCMC)	120
4.8.6	Bayesian implementation in WinBUGS	121
4.9	Results and Discussion	121
4.9.1	Classical multinomial model	121
4.9.2	Bayesian multinomial model	122
4.9.3	Discussion and partial conclusion	125
Chapter 5:	DISCUSSION AND RECOMMENDATIONS	126
5.1	Summary of Discussion	126

5.2	Limitations of the study	129
5.2.1	Data Limitations	129
5.2.2	Methodological Limitations	129
5.3	Scientific papers and articles	130
	References	182
	Appendix A	183
	Appendix B	186

List of Figures

Figure 2.1	Map showing geopolitical zone of Nigeria	12
Figure 4.1	Running Quantiles for the Posterior Parameters in the case of Female Benign and Malignant Breast Cancer Patients.	80
Figure 4.2	Auto-correlation plots for the Female Benign and Malignant Breast Cancer Patients.	81
Figure 4.3	The plot of the Brooks-Gelman MPSRF for three chains of 49,749 iterations.	82
Figure 4.4	Gelman Rubin convergence diagnosis for independent variables	83
Figure 4.5	Gelman Rubin convergence diagnosis for independent variables	84
Figure 4.6	WinBUGS' output of Gelman Rubin Statistic for some independent variable	95
Figure 4.7	Auto-correlation plots for the preferred treatment given to patients.	96
Figure 4.8	Posterior probability density plot	107
Figure 4.9	Sample autocorrelation plots	108
Figure 4.10	Quantile plots	109
Figure 4.11	Time series plot of MCMC output	110
Figure 4.12	WinBUGS' output autocorrelation	124
Figure 4.13	WinBUGS' output autocorrelation	124
Figure 5.1	WinBUGS' output of Gelman Rubin Statistic for some independent variable	184
Figure 5.2	WinBUGS' output of posterior density for some independent variable	185
Figure 5.3	WinBUGS' output of time series for some independent variable	185
Figure 5.4	WinBUGS' output of Gelman Rubin Statistic for some independent variable	186
Figure 5.5	WinBUGS' output of posterior density for some independent variable	187
Figure 5.6	WinBUGS' output of time series for some independent variable	188
Figure 5.7	WinBUGS' output of Gelman Rubin Statistic	189

Figure 5.8 WinBUGS' output autocorrelation 190
Figure 5.9 WinBUGS' output autocorrelation 191

List of Tables

Table 3.1	Values of the logistic regression model when the independent variable is binary	29
Table 3.2	Summary of diagnostic plots for a fitted model evaluation	36
Table 3.3	Some common distributions of exponential dispersion family with their link functions (Ntzoufras, 2011)	53
Table 3.4	Some common distributions with their link functions	57
Table 4.1	Distribution of grades of breast cancer at presentation	70
Table 4.2	Distribution of histologic types of breast carcinoma	70
Table 4.3	Summary of the categorized age groups for the patients	71
Table 4.4	Summary of the marital status for the patients	71
Table 4.5	Summary of the employment status of the patients	71
Table 4.6	Frequency distribution of breast cancer grade by biological risk factors (n= 236)	72
Table 4.7	Contingency table for educational status and breast cancer type by age group	72
Table 4.8	The Result of Classical logistic regression for Patients diagnosed of Benign and Malignant	77
Table 4.9	Heidelberger and Welch Stationarity and half-width tests for the Bayesian chains used in the diagnosis of MCMC	78
Table 4.10	WinBUGS Posterior Summaries for Breast Cancer Patients	78
Table 4.11	Convergence diagnostics MCMC algorithm for two way ANOVA model	79
Table 4.12	Multilevel Logistic Regression: Model A	92
Table 4.13	Multilevel Logistic Regression Model B	93
Table 4.14	Parameter estimates of models M1 and M2 obtained from WinBUGS	94
Table 4.15	Demographic characteristics of women diagnosed with different histologic types of breast cancer	94
Table 4.16	WinBUGS output for the evaluation of logistic regression multilevel using pD and DIC	97
Table 4.17	Multilevel multinomial model estimates of histological type	105
Table 4.18	Posterior means, posterior standard deviations and 95% credible intervals	106
Table 4.19	WinBUGS output for the evaluation of logistic regression multilevel using pD and DIC	110

Table 4.20 Posterior means, posterior standard deviations and 95% credible intervals . . 111

Table 4.21 Comparative results of the estimated parameters and their standard errors
based on the classical, Bayesian and bootstrapping multinomial regression models 122

Table 4.22 Posterior Distribution Summaries of parameters from MCMC Multinomial
model 123

Abbreviations

AIC	Akaike Information Criteria
ASR	Age Standardized Rate
BC	Breast Cancer
BIC	Bayesian Information Criteria
BGLMM	Bayesian Generalized Linear Mixed Model
BGR	Brooks-Gelman-Rubin
BMI	Body Mass Index
BOA	Bayesian Output Analysis
BRCA 1	Breast Cancer Type 1
BRCA 2	Breast Cancer Type 2
BUGS	Bayesian Inference Using Gibbs Sampling
CODA	Convergence Diagnosis and Output Analysis
Cred.I	Credible Interval
CI	Confidence Interval
DIC	Deviance Information Criteria
DNA	Deoxyribonucleic Acid
EDA	Exploratory Data Analysis
ER+	Estrogen Receptor Positive
ER-	Estrogen Receptor Negative
GOBP	Gene Ontology Biological Process
GLM	Generalized Linear model
GLMM	Generalized Linear mixed model
GR	Gelman-Rubin
Gy	Gray (is the standard unit of absorbed ionizing-radiation dose)
HW	Heidelberger-Welch
HER 2	Human Epidermal Growth Factor Receptor2
HRT	Hormone Replacement Therapy
HGLM	Hierarchical generalized linear models
IARC	International Agency for Research on Cancer
IGLS	Iterative Generalized Least Squares
INLA	Integrated Nested Laplace Approximation
LMCs	Low Middle Income Countries
LMM	Linear Mixed Model
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation

MRC	Medical Research Council
PSRF	Potential Scale Reduction Factor
REML	Restricted Maximum Likelihood
SAS	Statistical Analysis System
SSA	Sub-Saharan Africa
SE	Standard Error
SES	Socio Economic Status
SEER	Surveillance Epidemiology and End Results
WHO	World Health Organization

Chapter 1

INTRODUCTION

1.1 Background

Human life is exposed to the risk of various diseases across the globe. More specifically, a large percentage of the Nigerian population is being hit by a number of diseases whose incidences are rising at an alarming rate. Wide-ranging efforts to curb the rising incidence of these diseases, enhance societal awareness concerning their causes, and suggest various lines of treatment are being made. The rapid advancement of medical technologies, treatments, and research have significantly aided these efforts and helped in bringing the world to a point where many of these diseases are no longer considered as threatening to human life and well-being because they can now be prevented and/or cured. However, in the case of breast cancer, prevention strategies still require significant improvement and strengthening so as to reduce its rise in Africa and the world at large. The main constraint in managing breast cancer patients in Nigeria is the limited availability of equipment due to insufficient finances of health sector. This makes patients bear the burden of paying unreasonably large amounts of money for cancer treatment in a country where financial resources are already severely limited.

Cancerous cells arise due to mutations and alterations in DNA of normal healthy cells (Jackson & Bartek, 2009). Sometimes, a cyst develops around these cells and cell replication stops. This is known as the benign state. When these cells continue to grow rapidly and out of control, they are referred to as being malignant. Breast cancer risk factors are sometimes explained in relation to incidence and mortality. Therefore, there is need to distinguish between these two and to put into consideration different ways in which they can be modified. Breast cancer incidence is the number of new cases and is modified by the level of exposure to carcinogens, while breast cancer mortality is the number of deaths recorded from breast cancer, representing the risk of developing malignant breast lesion from the disease and as well

as the probability of dying as a result of it.

Research over the years have focused on treatments and drugs to stop the spread of breast cancer. A correct cancer diagnosis is needed for adequate and effective treatment because different cancer types have different requirements of a treatment regimen that encompasses one or more modalities like surgery, radiotherapy, chemotherapy, and so on. The ability to determine the suitable and best modality of treatment and palliative care for breast cancer patients specifically in western Nigeria will go a long way to help women of this geopolitical zone to manage their lives. The primary aim is generally to cure cancer or to considerably prolong life. Improving the patient's quality of life is also an important goal. With advancements in medical technology, advances have been made in the treatment of breast cancer in Nigeria regardless of the site of the cancer. However, better treatment for breast cancer patients is difficult to define.

The literature has revealed that elderly women are sometimes not included in clinical treatment trials, probably because of their age (Townsend et al., 2005). Since breast cancer biology differs from patient to patient with respect to factors like age, variations in response to treatment, and substantial competing risks of mortality (Biganzoli et al., 2012; Mieog et al., 2012; Kiderlen et al., 2015), the exclusion of some patients might have led to invalid/unreliable results. This implies that the exclusion of elderly women in trials probably led to an untrue representation for the general older population (Jolly, 2015). Consequently, an evidence-based treatment strategy for women with breast cancer is needed.

1.1.1 Breast Cancer

A significant increase in the incidence of breast cancer has been noticed (GLOBOCAN, 2012). This increment cuts across all continents of the world, and particularly Africa. Cancer of the breast is now the most dangerous disease common to women globally, and is a significant cause of cancer-related mortality in women across the world (Agboola et al., 2012). Therefore, it has increasingly become a focal point of research across the globe. On comparing the latest versions of the World Health Organization (WHO) and International Agency for Research on Cancer (IARC) reports on breast cancer (GLOBOCAN, 2008 & 2012), it was discovered that the number of new cases increased from 12.7 million in 2008 to 14.1 million cases in 2012. In addition, approximately 1.4 million women were diagnosed with breast cancer in 2012. There are 6.3 million women alive who have been diagnosed in the previous 5 years in relation to 2015 (Torre et al., 2015). This corresponds to an age-standardized rate

of 43.3 new cases per 100,000 (Ferlay et al., b). Thus, since the 2008 estimates, breast cancer incidence has increased by more than 20%, while mortality has increased by 14%. In western Nigeria, breast cancer accounted for 37% of all the cancer cases among women (OLUGBENGA et al., 2012), while in other geopolitical zones of Nigeria, reports suggest that breast cancer alone constituted about 35.41% of all cancers in women Afolayan (2008). Although breast cancer is thought to be a disease of the developed world, literature reveals that about 58% of deaths that occur as a result of breast cancer are traced to less developed countries (Ferlay et al., 2010). There are countless factors contributing to breast cancer which vary with respect to socio-economic, area, race, and life style differences. The role of lifestyle factors is well-illustrated in migrant studies, where breast cancer risks were compared between female migrants from low-to high-incidence countries and their offspring, showing that the risk increases in the following generations as a change in lifestyle is adopted (Ziegler et al., 1993; Shimizu et al., 1991).

Breast cancer stages have been shown to be one of the major prognostic factors (Møller et al., 2016; Jedy-Agba et al., 2017), particularly in an African setting. In addition, in contrast to breast cancer diagnosed in developed countries, the stages of breast cancer diagnosis in Nigeria as well as other African countries have been reported to be too late. The factors associated with this late diagnosis of breast cancer in Africa, and particularly in Nigeria, need to be studied in order to reduce the menace of this disease. Moreover, other factors that influence the prevalence of breast cancer among women in western Nigeria need to be studied for a better view of its epidemiology. A statistical study of this disease will be helpful in determining its prognostic risk factors, socio-economic factors, and related medical factors. This research examined the risk factors of breast cancer in Nigeria. The Nigerian setup is quite different from other parts of the African continent in many respects. For instance, unlike in Sub-Saharan African (SSA), smoking is rarely observed among women of Nigeria.

Several studies have established that breast cancer risk increases with an increase in age. Thus, age is an established risk factor for breast cancer. It has been discovered that poor socio-economic condition is also one of the factors causing an increase in breast cancer in Nigeria (Ojewusi & Arulogun, 2016). Those with highest socio-economic status and education have a twofold greater incidence than those with the lowest status. Some studies further showed that higher socioeconomic status (SES) is associated with higher breast cancer incidence (Pudrovska & Anikputa, 2012; Krieger et al., 2010; Vainshtein, 2008). Additional studies have provided a possible explanation for these findings, in that women with high SES are more likely to obtain routine breast cancer screening as a result of better access to preventive healthcare,

hence, increasing the detection rate of breast cancer (Akinyemiju et al., 2015). This study will explore these breast cancer determinants among Nigeria women. Cancer of the breast can be diagnosed in a woman of any age. However, the literature shows that the risk of developing breast cancer increases with age, and most cases (approximately 80%) of breast cancer occur in women over the age of 50 years. Other risk factors attributed to the increase in womans risk of breast cancer include hormonal and reproductive factors, obesity, alcohol, and physical inactivity. According to (McPherson et al., 2000), early age at menarche, older age at first birth, reduced parity, lack of breast feeding, and late menopause have all been associated with an increased risk of developing breast cancer in the developed world. Also, in 2003, it was reported by (Collaborators et al., 2003) that the use of oral contraceptives and hormone replacement therapies (HRT) have been shown to increase the risk of breast cancer for up to five years after discontinued use before the risk returns to the same levels as women who have never used hormone therapy. Moreover, it was mentioned in the literature that obesity in post-menopausal women (Reeves et al., 2007) and excessive alcohol consumption (Baan et al., 2007) have been found to increase risk of breast cancer, while engaging in some physical activity offers some protection (Chan et al., 2007). Statistics have indicated that approximately 10% of all female breast cancer is attributed to a genetic mutation (McPherson et al., 2000). There are two main genes that account for most genetically-linked breast cancer: BRCA 1 and BRCA 2. They have also been linked to ovarian, colon, and prostate cancer. From the literature, it was observed that women with a genetic mutation are more likely to have a strong family history of breast cancer and to be diagnosed at a younger age than the general population (McPherson et al., 2000). This has raised many unanswered questions, which motivated the initiation of the current research.

One major research question is, "How do we establish prognostic factors associated with breast cancer among women of western Nigeria?" Many subsidiary questions spring up from this research question: What are the explanatory and socio-economic factors of the prevalence level of breast cancer in the national and geopolitical zone of Nigeria? What are the attempts already made in the reduction of breast cancer? And what are the current and future consequences of breast cancer if more attention is not given to the menace? This research will attempt to show how breast cancer is associated to the events that occur among Western women of Nigeria. It will also explore the risk of breast cancer due to the relationships among the risk factors themselves, and help in estimating the preferred treatment given to breast cancer patients in this geopolitical zone of Nigeria. The cancer data in Nigeria shows that breast cancer is hitting the largest proportion of the entire population. The high incidence and mortality rates for breast cancer may be attributable to number of risk

factors that are to be explored for the Nigerian woman population. We have selected two sets of data from two hospitals for this study. The two hospitals are entirely different in their management and facilities for the patients. One is managed by the Federal government, while the other one is managed by state government. The two hospitals jointly represent the patients from all the prognostic factors levels.

Cancer

The human body is made up of billions of cells, and every one of them contains twenty-three pairs of chromosomes. These chromosomes contain myriads of different information that inform the body how it should grow, function as well as behave. Given the fact that chromosomes reproduce themselves every time a cell divides, there are many opportunities for something to go wrong during this process. Sometimes, something goes wrong with some cells and they do not die. They divide out of control and may grow into a tumor known as cancer. During the cell division, a process called a mutation alters one of the genes. The altered genes now begin to send wrong messages to other parts of the body. Thereafter, a cell begins to grow and multiply until it forms a lump called a malignant tumor or cancer. This tumor may be benign or malignant in nature. It is only when these cells start to divide uncontrollably, forming tumors or growths, that we have one of the more than 200 diseases called cancer. The major differences between malignant (cancerous) and benign (non-cancerous) tumors are that malignant tumors can spread into the surrounding tissue, destroy the surrounding tissue, and cause other tumors to develop. The major ways through which cancer spreads are outlined below.

- Invasion: The tumor grows into surrounding tissues or structures.
- Spread via the bloodstream: Cancer cells break away from the tumor, enter the bloodstream, and travel to a new site within the body.
- Spread via the lymphatic system: Cancer cells break away from the tumor and enter via the lymph vessels and lymph nodes to other parts within the body.

There are various types of cancers in relation to the site, namely, lung cancer, skin cancer, prostate cancer, basal cell cancer, breast cancer, etc.

1.2 Research Problem

According to WHO, breast cancer is the most commonly diagnosed cancer among women worldwide, claiming the lives of thousands of women each year and affecting countries at all levels of modernization. Studies have shown that out of all

the cases of cancer diagnosed in 2016, breast cancer alone accounted for 29%. This makes breast cancer to be the most frequently diagnosed cancer in women globally. The fight against this disease is given priority all over the world. Statistics regarding the issue unveil the need for intensive efforts towards addressing the menace of this disease in Africa in particular, because this chronic disease also affects this continent. In 2012, statistics indicated that 1.7 million women were diagnosed with breast cancer and the prevalence stood at 6.3 million women with 40 Nigerian women dying as a result every day (WHO, 2012). Although in the past, cancer of the breast used to be called a disease of the developed world, the literature reveals that almost 58% of deaths that occur as a result of breast cancer are traced to less developed countries (Ferlay et al., 2010). The level of breast cancer incidence in the world reflects the huge gap between rich and poor countries. This situation creates an additional problem related to the monitoring and measuring of the impact of undertaken actions against some prognostic factors of breast cancer in the Western part of Nigeria. As such, this bring a lot of serious confusion among data users at all levels about the real incidence level of breast cancer in the country. Statistics indicate that the occurrence and distribution of breast cancer among the southwestern citizens accounted for 37% of all cancer cases. The consequences of high breast cancer prevalence are enormous in each of the geopolitical zones of Nigeria.

Although not only of concern to biostatisticians, most researchers are trying to find the suitable methodology, with the smallest possible bias, to modeling breast cancer. Indeed, finding a lasting solution to this menace requires deeper knowledge on the phenomenon and the provision of reliable indicators reflecting the certainty of different geographical locations of the country. Different techniques have been used such as classical and non-parametric statistics, but they still have important limitations and need to be improved. Few studies have used Bayesian techniques to analyze breast cancer data in western Nigeria.

As with most social issues, the reliability of approaches differs from one country to another, depending on the specific socio-demographic. In Nigeria, the menace of breast cancer is significantly difficult for researchers to tackle because of its complexity, lack of data, and where the data exists, its usual poor quality. Another concern for researchers and policy makers is the representation of breast cancer, which involves a statistical model that is difficult to implement because of the weaknesses of existent data. In addition to these difficulties, we noticed an important insufficiency of scientific works related to this issue in western Nigeria. However, estimating the prevalence of breast cancer in the country and by zones is an important scientific issue and researching the factors explaining the problem is a challenge. Therefore,

several techniques have been proposed for modeling and analyzing breast cancer data. These approaches differ in the conditions underlying their use, the data requirements, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required for their implementation. This heterogeneity means that no single approach can be considered as best, or even operational, in all situations. To contribute improving knowledge in breast cancer issues in Western Nigeria, this thesis has set out the main objective of providing statistical information about the incidence and prognostic of breast cancer in Western Nigeria.

1.3 Aim and Objectives

Specifically, the objectives of the research are:

- To describe the patterns and distribution of breast cancer prevalence among western women in Nigeria as well as to establish factors which most contribute to breast cancer prevalence.
- To explain the variations in breast cancer (via a modeling technique) using socio-demographic and biological factors of the population, taking into consideration the hierarchical structure of the data via both classical and Bayesian techniques.
- To assess the association between breast cancer patients and some socio-demographic and biological factors.
- To develop a model based on the socio-economic determinants of breast cancer cases for Western Nigerian women.

1.4 Significance of the Study

Different established factors in relation to cancer of the breast have been investigated from the literature that are not similar to the Nigeria situation. Meanwhile, there exists a wider knowledge on the correlation between socio-demographic, reproductive, and risk factors and breast cancer in Nigeria. It was discovered that there exists at least some variation in breast cancer prevalence rates globally (Bray et al., 2004). This variation is attributed to a range of socio-economically correlated differences in the population of many hormonal and reproductive factors. Previous research on immigrants has showed the aforementioned environmental determinants as part of the factors which may contribute to breast cancer prevalence based on observed international and inter-ethnic differences. The study by (Ziegler et al., 1993) highlighted the successive increases in risk for migrants from low-risk Asia to high-risk

United States, mostly when the migration occurred in childhood.

Globally, women in developing countries (or women living in poor areas) are prone to higher risks of breast cancer upon migration because of changing exposures to reproductive and nutrition-related determinants. The most significant increments are observed in developing countries, where breast cancer risk has historically been low relative to industrialized countries (Bray et al., 2004). Moreover, few attempts have been seen in the literature that quantify the magnitude of risk between populations that might be explained by unique factors (Parkin et al., 2005). Although the relationships between various breast cancer risk factors have been studied elsewhere, their findings may not be applicable to the Nigerian situation due to differences in population. Therefore, there is a need to explore the risk factors associated with breast cancer among women in western Nigeria using datasets from two different hospitals, and establish such associations. The establishment of such factors may help in the implementation of prevention strategies against the disease. In addition, the knowledge of the female population at risk will help target breast cancer screening interventions and improve advocacy for protective practices. This improved treatment outcome is necessary as most breast cancer diagnoses in developing countries are done at last stage of the disease.

In conclusion, knowledge about associated and established breast cancer-related factors will give insights about the possible causes of the disease among women of western Nigeria and as well strengthen the role of healthcare workers, improve prevention, early diagnosis and treatment modalities, thus reducing its menace. In addition, this study employs the Bayesian approach due to its wide range of applications and advantages over classical statistics. Most of the studies on breast cancer in Nigeria were based on classical statistics. However, researchers have recently found some shortcomings with the classical method of analysis and proposed the Bayesian method. For instance, in a study conducted by (Ojo et al., 2017) to established risk factors of tuberculosis (TB) in South Africa by comparing the result of Bayesian and classical statistics, they found that the Bayesian approach helps in selecting the more significant factors related to the risk of TB better than the classical approach.

1.5 Thesis Layout

In the light of the above-mentioned objectives, the following content has been assigned to the document. This thesis is concerned with the performance of the Bayesian Generalized Linear Mixed Model (BGLMM) for the breast cancer disease. It is divided into six chapters: Introduction, Literature Review, Research Methodology, Re-

sults, Discussion, and Conclusions and Recommendations based on the main results. The first chapter focuses on the presentation of the issue of breast cancer as well as the explanation of the rationale for undertaking a research on this disease. The scientific, social and economic importance of this study from international, national and sub-national perspectives is highlighted. The socio-political, socio-cultural, and economic context in which the study is conducted is also presented. In the second chapter, the issue under study is explored within its scientific context by reviewing existent literature on breast cancer. We also discuss the risk factors based on its socio-demographic and medical factors. We subsequently give an overview of extant scientific research on the prevalence of breast cancer in Nigeria as well as on the African continent. Generally, this chapter entails a summary of the existing body of knowledge about the causes of breast cancer and the identification of some aspects that have been of little interest in previous studies, but which demand close attention. In chapter three, we focus on the research methodology. A deeper explanation of the dataset chosen for this study is given, and the conceptual and analysis scheme as well as statistical techniques of analysis used are presented. We include in this chapter the reasons behind the choice of the datasets and methodology, the quality of the datasets, the methodological limits of the research, as well as their impact on the findings. In chapter four, results of the descriptive and multivariate analysis (classical and Bayesian approach) of breast cancer are provided. In addition to the identification of the significant determinants of breast cancer, their mechanism of action as well as contributions are pointed out. In chapter 5, results from the analyses in chapter four are reviewed in tandem with knowledge from literature reviews and the context of the study. Acceptance of results and contribution of these findings to knowledge are discussed in detail. The last chapter provides conclusions and recommendations for improved actions against breast cancer prognostics in Western Nigeria. The main results of this research and scientific contributions are highlighted in the conclusion. The WinBUGS code for the entire thesis is provided in appendix E. In appendix D, R software code is provided which is used in the classical statistics while in appendix E, SAS code for other part of classical statistics is presented.

1.5.1 Thesis Contributions

A summary of the contributions of this thesis is outlined below

- Previous studies on breast cancer modeling in Western Nigeria use ordinary logistic regression and χ^2 test to model breast cancer, which implies they didn't consider the hierarchical nature of the dataset. This research moves further by incorporating the hierarchical structure into our models.
- Another contribution of this thesis is a comparison between Bayesian multi-

nomial models, classical multinomial models, and bootstrapping technique. Their results were compared, and the findings highlight the fact that the Bayesian model performs better than bootstrapping, followed by the classical model.

Chapter 2

LITERATURE REVIEW

The problem of breast cancer is not a new scientific problem across many fields and among researchers. Most researchers from medical sciences, sociology, epidemiology and other disciplines have studied the problem for many years, and investigations are still ongoing. This chapter aims to provide a synthesis of the literature about the specific points of interest defined in our objectives. We aim to review the literature on breast cancer prevalence, its explanatory factors, as well as their mechanisms of influence. Highlighting methodologies and results of past researches on breast cancer could help to better orientate this study and make use of this information to push forward knowledge in the subject.

2.1 Context of the Study

A presentation of the context of this thesis is important to clarify the problem and deepen the understanding of the issue of breast cancer. The context within which this study is undertaken, and the location of the data used supports the importance of, and need to carry out this research. In this section, the context of Western Nigeria women in relation to breast cancer is highlighted.

2.1.1 Area of Study

Nigeria a country in West Africa bordered by the Bight of Benin and Gulf of Guinea in the south. It shares maritime borders with Equatorial Guinea, Ghana, and So Tom and Prncipe. With an area of $923,768 \text{ km}^2$, the country is almost four times the size of the UK or slightly more than twice the size of the U.S. state of California. According to the last census, the country had about 200 million people, with more than half its people under 30 years of age (Source: NPC).



Figure 2.1. Map showing geopolitical zone of Nigeria

2.1.2 Global Prevalence of Breast Cancer

As we go through different stages in life, our bodies are subjected to many negative things. One of such negative things is a disease called cancer. Cancer is a generic term for a large group of diseases that affect any part of human body. Breast cancer is a form of disease that characterized by cells in the breast, becoming abnormal, and multiply uncontrollably, resulting in a tumor. In a situation where the tumor is not

treated on time, malignant cells can spread beyond the original tumor to other parts of the body, leading to a process called metastasis. Different types of cancer of the breast are identified by the cells in the breast that become malignant. A tumor may be malignant (cancerous) or benign (not cancerous). A cancerous tumor is malignant in nature if it grows and spreads to other parts of the body, while benign tumor can grow, but will not spread.

Cancer is one of the leading causes of morbidity and mortality among women globally, with approximately 14 million new cases diagnosed in 2012 (Ferlay et al., 2013a). The incidence of major cancers is rising globally. There is a great disparity among population segments. The higher incidence and mortality rates of cancer among blacks as compared to whites represent social, economic, environmental and educational factors rather than racial or genetic (Ries et al., 2006). In addition, being black also correlates highly significantly with being poor, less educated and deprived of a safe healthy environment. In a situation where all these factors are put together, the risk of cancer rises significantly. For a higher percentage of newly diagnosed cases of female breast cancer across the world, it has been observed that cancer of the breast is a neglected disease in terms of other numerically more frequent health problems. Some other school of thought regarded it as an orphan disease because very detailed information about tumor characteristics and the necessary host biology capable of providing basic care is absent. However, in some developed countries, Ginsburg et al., (2011) have documented the progress with the declines in mortality of breast cancer.

Globally, (GLOBOCAN, 2012) as reported by (Torre et al., 2015) showed that the impact of breast cancer has been rising in most continents of the world. In the same vein, there are wide gaps between rich and poor countries as noticed by Wild (2013). Among the developed countries, the incidences remain highest, while mortality rates are relatively much higher in less-developed countries. Findings show that the burden of breast cancer will increase in the years to come, not only because of the steep increase in incidence, but because of the increase in population in these countries (Taib et al., 2014; Jemal et al., 2011; Porter, 2008) found that an increase in life expectancy is believed to be an outcome of a reduction in mortality from infectious diseases by 2020. Therefore, Wild (2013) requested for proper attention for prevention and control measures to offset lifestyle changes that makes cancer of the breast the leading cause of cancer death and malignancy among women, particularly in developing countries like Nigeria. To reduce the menace posed by breast cancer, (Ferlay et al., 2010) called for the implementation of practical and affordable methods of diagnosis, early detection, and treatment of breast cancer among women of less developed countries. According to (Vineis & Wild, 2014), it is crit-

ical that mortality and morbidity rates in developing countries be brought in line with the progress made in recent years in more developed parts of the world. Improvements on the management of breast cancer patients in the developed countries have resulted to a reduction in mortality rates in (Yip & Taib, 2014; Taib et al., 2014). However, in developing countries, death from breast cancer continues to rise due to late presentation of the disease and lack of access to appropriate healthcare (Coughlin & Ekwueme, 2009; Taib et al., 2014). The higher breast cancer mortality rate for women in less developed countries noticed in the literature is partly because clinical advances to combat the disease are not available for women (Forman et al., 2012). This is attributed to the lack of early detection and poor access to treatment as a result of lack of awareness, lack of education, and deficient infrastructural and healthcare facilities. According to (Youlden et al., 2014), the report of GLOBOCAN (2012) indicated an alarming disparity in breast cancer incidence and mortality between the United States and the rest of the world. In United States, incidence of breast cancer remained stable from 2001-2010, with a 15.6% reduction in mortality across the life span; whereas globally, breast cancer claimed almost 522,000 women's lives in 2008, a figure that increased by 14% in 2012. (Youlden et al., 2014) adds that while incidence rates remain highest in more developed countries, the GLOBOCAN (2012) data showed that mortality rates are highest in less developed countries. He provided a fourfold explanation for this observation, and called for the application of the strategies that were successfully employed in the West to bring down the breast cancer mortality rate in developing countries, so as to save millions of lives among women.

It was mentioned in the literature that the most common cancer in the United Kingdom (UK) is breast cancer, which account for 31% of all cancers in women (Parkin, 2011a). The UK cancer research center estimated that about 47,693 women were diagnosed in 2008 with breast cancer (Coleman et al., 2011) and approximately 12,000 women died from breast cancer in the UK, representing about 16% of all female mortalities from cancer.

Despite the threat that breast cancer poses to human health, particularly Africa, (Vineis & Wild, 2014; Sylla & Wild, 2012) observed that few countries in this region have breast cancer-related data. For instance, most of the breast cancer incidence data in Sub-Saharan Africa (SSA) in recent times were based on reports from registries (Curado et al., 2007; Jedy-Agba et al., 2012a). It was further mentioned that the incidence rate of breast cancer in their study was higher than that reported by GLOBOCANs (2008) estimate of 38.7% per 100,000. Due to the prevalence of risk factors for these cancers, the reported increasing incidence may be real as reported

by (Forouzanfar et al., 2011). The need for high quality regional cancer registries to serve a vast country like Nigeria has been highlighted so as to adequately inform policy and the allocation of resources for breast cancer treatment (Jedy-Agba et al., 2012b). Past studies (Afolayan, 2008; Jedy-Agba et al., 2012a; Ferlay, 2004) have indicated and predicted an increment in breast cancer incidence and mortality rate for Nigeria. In developed countries such as Canada, reports showed that fewer of their women are dying from breast cancer. In the same vein, the Canadian Cancer Society reported a reduction by 42% since the peak in 1986 (Canadian Cancer Statistics, 2014). It was indicated further that Canadian women who are diagnosed with breast cancer are living longer than ever before, based on a 5-year survival rate of 88% (Canadian Cancer Statistics, 2014). In Nigeria, the situation is different where there is a prediction of more than a 100% increase in incidence and mortality rates of cancer of the breast by 2030 (Jedy-Agba et al., 2012b; Sylla & Wild, 2012). With the disparities in breast cancer outcomes between developed and developing countries, it is imperative that action be taken to understand some of the prognostic factors associated with late presentation of breast cancer in Nigeria and address them appropriately.

Based on the GLOBOCAN estimates in 2008, about 12.7 million cancer cases are estimated to have occurred, and out of these, breast cancer alone contributed almost 23%. Statistics show that it is the leading cause of cancer death among females in less-developed countries (Siegel et al., 2015). According to cancer statistics released in 2017 in the United States, breast cancer alone accounts for 29% of all cancers in women (Siegel et al., 2017). Approximately 1,700,000 new cases of breast cancer were diagnosed worldwide in 2012. This represents about 25% of all cancers among women (Thrift, 2016). No definite cause has been attributed to breast cancer, but some researchers outlined genetic factors, personal history, etc as role players. A report by (Ferlay et al., 2010), indicated that more than 1,100,000 cases of breast cancer are diagnosed and more than 410,000 result to death among the patients globally. The situation is however different in developed countries where about 55% of the global burden is currently experienced, and incidence rates are increasing more rapidly in developing countries. In another development, Stewart et al (Stewart et al., 2003), reported that most of the new cancer cases are now occurring among women from low and middle-income countries, where the incidence is increasing by as much as 5% per each year and there are about three-fourths of breast cancer deaths occurring globally. Out of the 411,000 breast cancer deaths that occurred in 2002 across the globe, 221,000 (representing 54%) occurred in low- and middle-income countries (LMCs) while the situation is different from China where the incidence rose from 126,227 cases in 2002 (Ferlay et al., a) to over 169,000 in 2008.

The incidence and mortality rates of female breast cancer vary greatly by region. In Western Europe, the mortality has declined from around 1980, but in parts of some European countries such as Georgia, the burden continues to rise (Autier et al., 2010; Forouzanfar et al., 2011). The age-adjusted incidence rate for breast cancer for all women in Georgia in 2008 was reported to be 38.5 per 100,000 women while the mortality rate was 19.5 per 100,000 (Forman et al., 2012). Available data from the Georgian National Cancer Center (GNCC) in 2010 (which is the most recent year for which data is available) mentioned that 1221 women were diagnosed with breast cancer, out of which 598 women died, indicating a 0.5 mortality-to-incidence ratio, which implies that there are approximately 5 deaths for every 10 newly diagnosed cases, compared to a ratio of 0.2 in the United States (DeSantis et al., 2011; Fritsch et al., 2012; Harford et al., 2011). The survival rate of breast cancer patient aged (14-49) years in Georgia was 18.9% compared to 89% in the United States (Fritsch et al., 2012). The situation is different in neighboring countries like Russia and Ukraine where the survival rates were found to be 50-54% between 1994-2004 and 80% in Western Europe for the same period (Hirte et al., 2007). The higher fatality rates of breast cancer in low- and middle-income countries have been attributed to a lack of awareness regarding the benefits of detection and treatment and late stage diagnosis among other risk factors (Shulman et al., 2010; Dvaladze, 2012).

In general, breast cancer survival rates have improved worldwide, though survival rates vary from country to country. This is because breast cancer is diagnosed at an earlier and localized stage in countries where patients have access to medical care, and there is progressive improvement in treatment strategies. In many countries with advanced medical care, the five-year survival rate of early stage breast cancers is 80 - 90 percent, which falls to 24 percent for breast cancers diagnosed at a more advanced stage. However, this is not the case in some parts of African countries where most of the public healthcare sector is poorly resourced and managed as a result of political unrest, corruption, and economic situations. For instance, in Tanzania, it was reported that 30% of the countrys healthcare professionals leave the health sector after receiving medical training because of poor incentives (Bryan et al., 2010).

On the African continent, a serious challenge is the lack of cancer registries by most countries. The World Health Organization (WHO) estimates that the incidence rate of breast cancer in Africa has increased steadily over the years (Ferlay et al., 2013b). As reported by (Anderson & Jakesz, 2008), there is also a higher mortality rate among breast cancer patients in most African countries. For instance, the five-year breast cancer survival rate in Gambia stood at 12.5% (Jemal et al., 2011). Considering the cases of breast cancer in low-to-middle-income countries like Nigeria, the

breast cancer burden and its mortality rates are increasing at an alarming rate as reported by (Jedy-Agba et al., 2012a; Pruitt et al., 2015). The delays in diagnosis and treatment in Nigeria have contributed to the rise of this menace (Oluwatosin & Oladepo, 2006). There are many factors that prevent Nigerian women from seeking treatment when they first notice a breast cancer symptom, such as misconceptions about breast cancer and its treatment outcomes, and economic and logistic obstacles (Anyanwu, 2008; Karayurt et al., 2008). Cultural and social factors such as stigma and inadequate healthcare infrastructure have also been linked to this behavior among women (Bello, 2012; Pruitt et al., 2015).

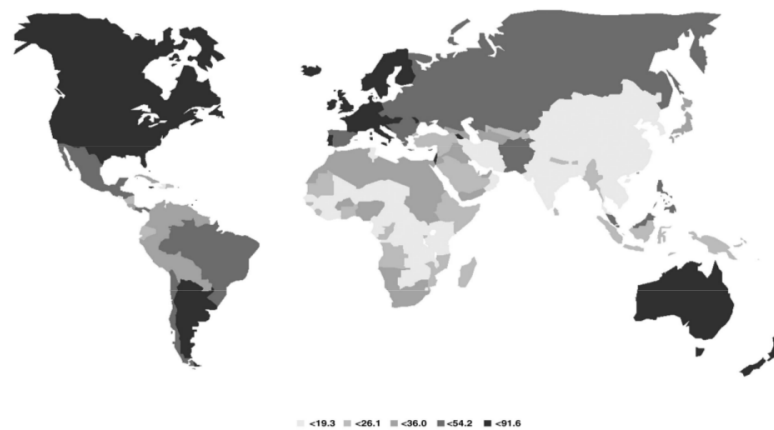


Figure 1: Breast cancer incidence worldwide: age-standardized rates (world population). Source: Ferlay *et al.*, 2001.

2.2 Breast Cancer Risk Factors

Available data from the World Health Organization (WHO) put breast cancer as the most frequent type of cancer in women, accounted for at least half a million deaths per year globally. Breast cancer is about a hundred times as frequent among women as among men, but survival rates are equal in both sexes. Several factors, including age, family history, age at first full-term pregnancy over 30 years, early menarche, late menopause, postmenopausal obesity, use of postmenopausal hormones, alcohol consumption, the use of contraceptives, hormonal treatment after menopause, lack of a breastfeeding history, obesity, and physical inactivity are associated with increased risk of breast cancer (Zare et al., 2013). It is estimated that 5-10% of breast cancer cases result from inherited mutations or alterations in the breast cancer susceptibility genes BRCA1 and BRCA2. In this thesis, breast cancer risk factors are categorized as socio-demographic and medical for a proper understanding.

2.2.1 Socio-Demographic Factors

- **Age:** In all countries of the world age is strongly related to breast cancer. It is extremely rare below the age of 20 years, but thereafter, the incidence steadily rises so that by the age of 90 years about 20% of women are affected by this disease.
- **Personal history of breast cancer:** An individual who has already been diagnosed with breast cancer has a higher risk of developing it again, either in the same breast or the other breast, than if they never had the disease.
- **Race and geographical areas:** From the literature, it can be observed that white women are slightly more likely to develop breast cancer than African American women. Asian, Hispanic, and Native American women have a lower risk of developing and dying from breast cancer.
- **Family history:** This is as a result of genetic mutation. Genes can be recessive or dominant, and when dominant in a patient, there is higher risk of them getting breast cancer. Hence, if they have a first-degree relative (mother, daughter, sister) who has had breast cancer, or multiple relatives affected by breast or ovarian cancer (especially before they turned age 50), they could be at a higher risk of contracting the disease.
- **Inherited factors:** the risk of breast cancer increases with age. If the trait is inherited, then as the age increases above 50 years, there is a tendency for some of these inherited genes to mutate (change). This mutation is abnormal, and the multiplication can cause breast cancer.
- **When a woman starts menstruation at an early age before 12 years or reaches menopause at an older age after 55 years, they have high risk of breast cancer due to a longer life time exposure to hormones. There are some hormones produced by the female organ, such as estrogen, which when produced in large quantity by the body will be deposited at the lymph nodes, multiply, and possibly cause breast cancer.**
- **Education and Exposure:** The risk of getting breast cancer can increase with increase in education or exposure to a more 'advanced' life. It has been observed that people in this class may one way or the other have exposed their chest or breasts to radiation therapy for treatment of diseases such as Hodgkins Lymphoma or even during registration in school using an X-ray machine. This radiation has an increased chance of breaking cells in the body, leading to these cells rapidly multiplying and becoming cancerous.

- Other factors are late or no pregnancy. Pregnancy after 30 years can increase the risk of breast cancer, as sex hormones like estrogen must have accumulated too much over the period of years instead of being used for pregnancy or child birth.
- Obesity and dense breasts can also be a factor. When there is too much connective fatty tissue in the breast resulting from fatty foods or an unhealthy feeding lifestyle, it becomes difficult for mammograms to detect any lump in the lymph nodes in the breast.

2.2.2 Medical factors

- Oral Contraceptive: Pills that control conception can also be a factor that increases chances of breast cancer. These pills work on the organs and hormones, some increasing blood flow for days, which only occurs when some hormones have been over produced. This abnormal multiplication can result in cancerous cells.
- Hormone Replacement Therapy (HRT): The International Agency for Research on Cancer (IARC) found that the use of combined oestrogen-progestogen hormone replacement therapy (HRT) for menopausal symptoms is as a highly probable cause of breast cancer. IARC has classified the use of oestrogen-only HRT as a possible cause of breast cancer, based on limited evidence (Chi et al., 2015). An estimated 3% of female breast cancers in the UK are linked to HRT use (Parkin, 2011b). A report from (Reeves et al., 2006) has shown that breast cancer risk is also higher in oestrogen-only HRT users. These findings are supported by the findings of (for the Women's Health Initiative Investigators et al., 2002), which indicated that 5 years of combined HRT was associated with a 26% increased risk of invasive breast cancer in post-menopausal women.
- X-rays: Ionizing radiation has been established as a risk factor of cancer. In the case of breast cancer, the risk increases linearly with the radiation dose according to (Ronckers et al., 2004). This has been noticed with diagnostic, therapeutic and accidental exposures to ionizing radiation. In addition, epidemiological studies of atomic bomb survivors and of medically irradiated populations reveal increased risk of female breast cancer with relative risks ranging from 1.0 4.3 per Gray (i.e. Gy, which is the standard unit of absorbed ionizing-radiation dose). The risk is higher if the exposure occurs in childhood and adolescence rather than in adulthood; it is minimum to zero if exposure occurred in the post-menopausal period (Andrieu et al., 2006).

- **Dietary factors:** According to (Holmes & Willett, 2004), the impact of specific dietary factors in breast cancer causation has not been completely dealt with it. In fact, trends show that breast cancer prevalence rates vary widely across the globe, and that the offspring of those who migrate from lower-incidence countries to those with higher incidence take on the higher rates of this disease. Moreover, observations promote the hypothesis that nutrition is an environmental determinant of breast cancer. This is supported by (Bray et al., 2004), who reports that there is a long established correlation between breast cancer and dietary fat intake, though the true relation does not seem to be strong and consistent.
- **Breast density:** The risk of breast cancer among women with the most dense breasts is higher compared with the least dense as reported by (Pettersson et al., 2014). Breast density is generally higher in younger, pre-menopausal women with lower body mass index (BMI) and lower parity, but there is also a genetic element (Boyd et al., 2002).
- **Exposure to estrogen.** Because the female hormone estrogen stimulates breast cell growth, exposure to estrogen over long periods of time, without any breaks, can increase the risk of breast cancer. Some of these risk factors are not under the individual's control, such as: starting menstruation (monthly periods) at a young age (before age 12) going through menopause (end of monthly cycles) at a late age (after 55) exposure to estrogens in the environment (such as hormones in meat or pesticides such as DDT, which produce estrogen-like substances when broken down by the body).

2.2.3 Context of Breast Cancer in Nigeria

The incidence of breast cancer in Nigeria has risen significantly (Jedy-Agba et al., 2012a). The age-standardized incidence rate for breast cancer in the period between 1960-1969 was 13.7 per 100,000. It rose to 24.7 per 100,000 by 1998-1999; more or less a doubling of incidence over four decades, or an approximately 25% increase in rate per decade. The rate in 2009-2010 was 54.3 per 100,000. This represents a 100% increase in the last ten years (Jedy-Agba et al., 2012a). The incidence of breast cancer in Nigeria in 2012 was (estimated age-standardized incidence rate (ASR) = 50.5 per 100,000) only half that in the United States (US) (ASR = 92.9 per 100,000), but estimates of mortality rates from this cancer were higher in this West African country than in the US (ASR = 25.9 vs. 14.9 per 100,000, respectively) (Jedy-Agba et al., 2017; Gathani et al., 2014), accounting for poorer survival (Jedy-Agba et al., 2017; Ferlay et al., b). In 2012 (OLUGBENGA et al., 2012) reported for western Nige-

ria the highest incidence of breast cancer for any western women while the hospital relative frequency data from the North-Central geopolitical zone in Nigeria also indicated a high frequency of breast cancer (Afolayan et al., 2012). A study conducted by (OLUGBENGA et al., 2012) in western Nigeria on the frequency of tumors of various sites, placed breast cancer at the top of the list of all types of cancers. The age-pattern of the incidence of breast cancer in Nigerian was similar to that found in studies of female breast cancer in other populations (Adebamowo & Adekunle, 1999; Ikpat et al., 2002; Ebughe et al., 2013a).

Breast cancer is the most common malignancy among Nigerian women (Jedy-Agba et al., 2012a). The incidence of breast cancer in Nigeria women increased from 13.5-15.3 per 100,000 in 1976 to 53.6 per 100,000 women in 1992, reaching the level of 116 per 100,000 women in 2001. As reported by (Akarolo-Anthony et al., 2010), it is now the leading cause of death among women in Nigeria. In the case of Afolayan et.al (Afolayan et al., 2012) they observed a steady increase in the incidence of breast cancer within a ten-year (1999-2008) period of study. There were occasional drops in incidence that coincided with periods of national or regional industrial unrest when public health facilities were closed. In the context of Nigeria, cancer of the breast is characterized by regional variation (Agboola et al., 2012). Nigeria is subdivided into different geopolitical zone. In the North Western geopolitical zone of Nigeria, breast cancer is reported to be second to cervix cancer. The situation is different in the Western geopolitical zone where breast cancer is the leading cause of malignancy among women (Jedy-Agba et al., 2012a; Afolayan et al., 2012). This situation is further different at Ilorin which is one of the cities in the North Central geopolitical zone, where breast cancer constituted 22.4% of new cancer cases registered in five years, and accounted for 35.41% of all cancers present in women. According to (Jedy-Agba et al., 2012a) cancer of the breast in Nigeria is associated with high mortality rates. As reported by (Forbes, 1997) with the adoption of Western lifestyles by African women, the incidence of cancer of the breast will continue to rise if proper measures are not taken, while an increase in mortality rate would follow. This growth as predicted by (Taib et al., 2014). According to (Afolayan et al., 2012), unfortunately, there is a no much data and sparse literature review on the trends of breast cancer in Nigeria. There are very few cancer registries in Nigeria, most of which are either hospital-based or pathology-based, instead of the preferred population-based cancer registries as in the case of developed countries. However, in low-resource countries like Nigeria, hospital-based cancer registries have been serving as a fundamental source of information on cancer prevalence (Curado et al., 2007).

Most of breast cancer cases in Nigeria are diagnosed late (Okobia et al., 2006; Anyanwu, 2008) as a result of poor utilization of screening facilities and lack of education on the part of patients. The peak age of breast cancer presentation in Nigeria is between 10-15 years later, compared to what is observed among Caucasians and African (and African American) women, where it occurs between the ages of 35-45 years (Okobia et al., 2006; Frempong et al., 2008). The major influences on breast cancer appear to be certain reproductive factors and diet. The level of cumulative incidence seemed to be higher in Nigeria compared to other African countries, most especially sub-saharan africa (SSA). The genetic, environmental, lifestyle, race, age and dietary differences could have contributed to these differences (OLUGBENGA et al., 2012; Ebughe et al., 2013a). Despite the threat that breast cancer poses to public health especially in sub-Saharan Africa, few countries in the region have data on breast cancer incidence (Sylla & Wild, 2012; Vineis & Wild, 2014).

2.2.4 Epidemiology of Breast Cancer

Age is an important factor in determining a woman's risk of having breast cancer (Ikpatt et al., 2002). The probability that a woman will be diagnosed of cancer of the breast increases throughout her life span, and most of it occurs during the postmenopausal period. It was mentioned in the literature that woman with high socio-economic status (SES) have higher risk of having breast cancer (Akinyemiju et al., 2015), and that it is common among women who delay having children into their 30's.

A late age of first birth, early age oral contraceptive use, as well as proliferative benign breast lesions have been associated with a higher risk of breast cancer among younger people (under 40) rather than older women (greater than 40). In addition, an inverse association was studied between the number of full term pregnancies and risk of breast cancer, and was found that women who gave birth four times or more had only one half the risk of women who gave birth just once. Many studies (Hulka & Moorman, 2001; Key et al., 2001; Colditz et al., 1993) have identified increasing age and a family history of breast cancer as established risk factors. A consistent increase in the risk of cancer of the breast was observed among women with a first-degree relative (mother, daughter, sister) who had the disease through pregnancy. In the case of women with a family history of breast cancer, the adverse effect was maintained up to age 70 years; parous women were at higher risk of breast cancer than nulliparous women. In the category of women without family history of the disease, first pregnancy was associated with a smaller increase in risk and early pregnancy and higher number of births were each correlated with a reduced

breast cancer prevalence (Antoniou et al., 2003). An antecedent history of breast cancer and benign breast lesions in a woman were found to be a significant factor that can increase the risk of the disease (Deane & Degner, 1998; Hislop & Elwood, 1981). Available statistics indicated that about half of the breast cancer in Italy are as a result of a high level of education and family history of the disease (Tavani et al., 1997). From the epidemiological evidence, breast cancer risk has also increased in subjects with family history of the disease (Antoniou et al., 2003).

It has often been stated that breast cancer risk is associated with socio-economic status (like income, educational status, etc). The reasons behind this are enormous. According to (Landman et al., 2010), when examining mortality, there is need to put into consideration the fact that survival is lower in lower classes, and secondly, that most of the gradient can be explained by the differing prevalence of common risk factors among the social classes. The variation in risk by educational status is almost entirely explained by the differential distribution of factors like parity, obesity and so on for the United State (US) case (Shaw et al., 2011).

The effect of reproductive factors as well as menstrual factors in the epidemiology of breast cancer have long been recognized.

The risk of breast cancer is higher in woman who had their first live birth at age 30 or higher, compared with nulliparous women (Nelson et al., 2012; Anderson et al., 2010). But older age at first delivery (greater or equal to 28 years) was shown to be moderate risk of breast cancer, mostly among premenopausal women, compared with the risk among women who had their first delivery at age less than or equal to 22 years old (Ritte et al., 2012). According to (DeSantis et al., 2011), they found that parity and age at first birth are correlated with the incidence of breast, as well as with the stage of diagnosis. In women aged 25-54, the opposite was observed, as there was no association between risk of breast cancer and age at first full term pregnancy (Cibula et al., 2010; Dite et al., 2010).

Moreover, a womans age at the time of her first full term pregnancy and the number of pregnancies have seen as an important determinant of breast cancer risk. (Nelson et al., 2012), in their stud,y found a significant reduction in the risk of breast cancer for women who had their first child at an early age. In the case of women whose first pregnancy occurred at ages 30-34, they were shown to have the same risk as nulliparous women. Pregnancies after 35 years are associated with an increased risk compared with nulliparous women (Flenady et al., 2011; Brown et al., 2010).

Some studies (Almeida et al., 2015; Buchanan et al., 2013) have found that the focus of etiology of exploration for cancer of the breast has shifted from dietary risk factors back to body size as well as reproductive factors. The difference in environmental risk factors was found to be the major difference in the incidence of breast cancer in Africa compared with Western countries (Jemal et al., 2010). Most of the studies on epidemiology (Van Den Brandt et al., 2000; Hsieh et al., 1990) found a positive association between obesity and breast cancer risk. Other studies (Saldova et al., 2014; Bhaskaran et al., 2014; Zheng et al., 2011) have found that a high body mass index (BMI) is positively associated with an increased risk of breast carcinoma. Some women add weight during menopause (Jensen et al., 2003; Simkin-Silverman & Wing, 2000; Mamun et al., 2010). The body mass index (BMI) remained insignificant with breast cancer for both premenopausal and postmenopausal women. A higher risk of breast cancer has been associated with women having a higher body weight and height, with the trend more pronounced for height (Hsieh et al., 1990; Onland-Moret et al., 2005). A study by (Colditz et al., 1995; Thompson et al., 2009) found that BMI, weight, and breast density were differently correlated with risk of breast cancer among postmenopausal women. It was also reported that height and obesity (BMI greater or equal to 30) were positively associated with breast cancer risk among postmenopausal women while in premenopausal women the opposite was observed (Ellison-Loschmann et al., 2013; Adebamowo et al., 2003; Amadou et al., 2013).

Socio-economic status has been observed as one of the risk factor of breast cancer. For example, age has been documented as one of the major risk factors for breast cancer across the world. In a study conducted by (Arora & Simmons, 2009), it was found that the age group 35-49 faced a major risk of breast cancer. A possible explanation for this may be the use of contraceptives and hormonal imbalance, which is common among women (Onyeanusi, 2015; OLUGBENGA et al., 2012).

Chapter 3

DATA AND RESEARCH METHODOLOGY

This chapter presents the research methodology used to reach the objectives of the study. It explains the adequacy of the methodology employed and the efficiency of the choices adopted. We also present our exploratory data analysis and the statistical methods. Furthermore, we present an explicit approach through which our data exploration and statistical tools helped to achieve the objectives of this study. In the first section, the datasets used in the study are presented and discussed extensively for a better understanding of their use, as well as to highlight their adequacy. In the second section, the methodological choices and their pertinence are developed. We present the data used for the analysis of this thesis. In particular, the variables and their respective significance to breast cancer research are explicitly detailed. The free software R, SAS and WinBUGS 14 are used to implement the statistical models, test the hypotheses, and plot data.

3.1 Data

This study aimed to study the performance of Bayesian analysis in modeling breast cancer among Western Nigeria women. The data was extracted from the cancer registry of Federal medical teaching hospital, Ekiti State, Nigeria. The data included information on demographic characteristics and socio-economic factors. Other information contained in our data is the number of cases of women diagnosed with histologically and pathologically confirmed cancer of the breast. The data extracted were for two different hospitals. One is federally owned, while the other is managed by a state government. An identification index, including each patient's name, residence, age, date of visit to the hospital and the identification number of each patient were used in order to uniquely identify the breast cancer patients from the cancer

records.

With respect to the quality of the data obtained from the records of the breast cancer cases, the major concern was the proportion of hospital records in which some of the relevant variables were missing. Information relating to variables like weight, height, age at first full term pregnancy, and age at menopause were missing. Factors considered in the analysis of this thesis include age at diagnosis, socio-demographic characteristics, marital status, and patients with pathologically and histologically confirmed cases. With regards to the quality of the data collected for the breast cancer patients, the primary concern was the high proportion of hospital records in which one or more variables were not recorded.

Research Scope and Methodology

In this thesis, the research is primarily focused on performance of the Bayesian Generalized Linear Mixed Model technique on breast cancer data in Nigeria. We employed an MCMC algorithm in estimating the required probability and likelihood functions, when exact computation were impractical. This technique shows great promise for providing a means of performing analyses which are practical in terms of computational time and which therefore can incorporate model complexities which currently are infeasible because of computational constraints. The contributions of this thesis is to be methodological with broad applicability of Bayesian technique compared to classical and Bootstrapping.

3.2 Classical and Bayesian Logistic Regression Model

Initially, all the categorical variables for the breast cancer data were examined in order to observe the true pattern of the data. The data was scrutinized so as to improve its quality before proceeding further in the analysis. The analysis was started with the descriptive statistics so as to exhibit the prevailing pattern in these variables. We performed logistic regression analysis by using both the exact logistic regression and conditional approaches. The categorical variables were studied at the stage of primary analyses by suitable statistical tests. Box-plots were used to examine the outliers or extreme values in where any appeared. The values obtained were then critically examined to rule out any distortion of the data. The criteria for including an individual variable into the model formulated was based on its statistical or biological significance. Logistic regression was used to test for the statistical significance of each variable. The conditional distribution of each response variable such as $\varphi = 1$ or 0 was also studied so as to decide relevant transformations of such variable in case of linear regression model. Generalized linear models (GLM) for cat-

egorical responses were initially developed as well in order to meet the need for the analysis of the datasets. These models explain the relationship between a response variable and some explanatory variables. Where the response variable is binary, the techniques for analyzing such responses are usually based on the assumption of a binomial distribution.

Logistic Regression Models

Suppose a binary random variable y follows a Bernoulli distribution, that is, y takes either the value 1 or the value 0 with probabilities $\eta_i(x)$ or $1-\eta_i(x)$ respectively, where $x = (x_1, \dots, x_k) \in \mathbb{R}^k$ is a vector of k explanatory variables. In addition, η_i represents the conditional probability $p(y = 1|x)$ of $y=1$, given x . Based on the binary outcome variable, we use the logistic distribution (Hosmer & Lemeshow, 2000b). The specific form of the logistic regression model with unknown parameters $\lambda_0, \dots, \lambda_k$ is

$$\eta(x) = \frac{e^{\lambda_0 + \lambda_1 x_1 + \dots + \lambda_k x_k}}{1 + e^{\lambda_0 + \lambda_1 x_1 + \dots + \lambda_k x_k}}.$$

Sometimes, it is convenient to change the notation by putting $x_0=1$. Hence, the model above becomes

$$\eta(x) = \frac{e^{(x^T \lambda)}}{1 + e^{(x^T \lambda)}}, \quad (3.1)$$

where $x = (x_1, \dots, x_k)^T$ and $\lambda = (\lambda_0, \dots, \lambda_k)^T$. The transformation of $\eta(x)$ is known as the logit transformation, and is represented as

$$\text{logit}(\eta(x)) = \log\left(\frac{\eta(x)}{1 - \eta(x)}\right). \quad (3.2)$$

Under the above transformation, the regression model in equation (3.1) is written as

$$\text{logit}(\eta(x)) = x^T \lambda. \quad (3.3)$$

3.2.1 Maximum Likelihood Estimate (MLE) of the parameters

Suppose we have a sample of n independent observations $\{(y_i, x_i)\}$, such that $i = (1, 2, \dots, n) \in (\{0, 1\} \times \mathbb{R}^{k+1})^n$, where y_i denotes the value of a dichotomous outcome variable, and x_i is the value of the explanatory variables for the i th individual. Assume

$$y_i \sim \text{Ber}(1, \eta(x_i)) \quad i = 1, 2, \dots, n.$$

Based on a set of data, we estimate the parameter vector $\lambda = (\lambda_0, \dots, \lambda_k)^T \in \mathbb{R}^{k+1}$ to fit the logistic regression model in equation (3.1). To obtain the ML estimator of λ , we define the likelihood function as follows

$$L(\lambda) = \prod_{i=1}^n \left\{ \eta(x_i)^{y_i} \times (1 - \eta(x_i))^{1-y_i} \right\}. \quad (3.4)$$

$$= \prod_{i=1}^n \left\{ \left(\frac{\eta(x_i)}{1 - \eta(x_i)} \right)^{y_i} (1 - \eta(x_i)) \right\}. \quad (3.5)$$

Using the notation in equation (3.1), the above expression becomes

$$L(\lambda) = \prod_{i=1}^n \left\{ \frac{\exp(y_i x_i^T \lambda)}{1 + \exp(y_i x_i^T \lambda)} \right\}. \quad (3.6)$$

Furthermore, we find the ML estimates, $\hat{\lambda}$, of λ by maximizing the log-likelihood function for the observed values of y_i and X_i . The log-likelihood function of the above expression yields

$$\ell(\lambda) = \sum_{i=1}^n \left\{ y_i x_i^T \lambda - \sum_{i=1}^n \ln(1 + e^{x_i^T \lambda}) \right\}. \quad (3.7)$$

Odds and Odds Ratio

The odds ratio is a measure of association which quantifies the relationship between being diagnosed with a particular disease or outcome and the health exposure under investigation. It is expressed as the ratio of the odds in favour of having the disease, if exposed, to the odds in favour of having the disease, if not exposed. Hence, it is necessary to introduce the terms odds and odd ratio so as to explain the binary data and to interpret the logistic regression coefficients. For a probability η of success, the odds are expressed as

$$Odds = \frac{\eta}{1 - \eta}.$$

In a 2 by 2 contingency table, the probability of success is η_1 in row 1 and η_2 in row 2. The odds of success within row 1 is expressed as

$$Odds_1 = \frac{\eta_1}{1 - \eta_1},$$

and odds of success within row 2 are given as

$$Odds_2 = \frac{\eta_2}{1 - \eta_2}.$$

Hence, the ratio from the two rows are called odds ratio and are represented by

$$Oddsratio = \frac{\frac{\eta_1}{1-\eta_1}}{\frac{\eta_2}{1-\eta_2}}. \tag{3.8}$$

Interpretation of the parameter

For the purpose of this study, we consider a single explanatory variable coded as either 0 or 1 for a better understanding of the logistic regression coefficients interpretation. The odds of the outcome being present among individuals with $x = 1$ are expressed as $\frac{\eta(1)}{1-\eta(1)}$. Also, the odds of the outcome being present among individuals with $x = 0$ are expressed as $\frac{\eta(0)}{1-\eta(0)}$. Therefore, the possible values of the logistic probabilities are shown in the table below.

Table 3.1: Values of the logistic regression model when the independent variable is binary

Response variable		Explanatory variable
y= 1	y= 0	
$\eta(1) = \frac{e^{\lambda_0 + \lambda_1}}{1 + e^{\lambda_0 + \lambda_1}}$	$1 - \eta(1) = \frac{1}{1 + e^{\lambda_0 + \lambda_1}}$	x= 1
$\eta(0) = \frac{e^{\lambda_0}}{1 + e^{\lambda_0}}$	$1 - \eta(0) = \frac{1}{1 + e^{\lambda_0}}$	x= 0

From the above table, we define the odds ratio as the ratio of the odds for $x = 1$ to the odds for $x = 0$, and is represented as

$$\begin{aligned}
 Oddsratio &= \frac{\frac{\eta(1)}{1-\eta(1)}}{\frac{\eta(0)}{1-\eta(0)}} \\
 &= \frac{\frac{\frac{e^{\lambda_0 + \lambda_1}}{1 + e^{\lambda_0 + \lambda_1}}}{\frac{1}{1 + e^{\lambda_0 + \lambda_1}}}}{\frac{\frac{e^{\lambda_0}}{1 + e^{\lambda_0}}}{\frac{1}{1 + e^{\lambda_0}}}} \\
 &= \frac{e^{\lambda_0 + \lambda_1}}{e^{\lambda_0}} = e^{\lambda_1}. \tag{3.9}
 \end{aligned}$$

Equation (3.9) means that the odds on the event that y equals 1 increase or decrease by the factor e^{λ_1} among those with $x=1$ than among those $x=0$. According to (Hosmer & Lemeshow, 2000b), this is the major reason why logistic regression has been chosen as a reliable statistical tool for medical research.

Bayesian Binary Logistic Regression Model

In the context of this study, we consider Bayesian logistic regression modeling from a generalized linear modeling (GLM) framework as described. In general, a generalized linear model (GLM) technique, as first introduced by (Nelder & Baker, 1972) and modified by (Fan & Gijbels, 1996), provides a flexible and unified approach to analyzing both normal and non-normal data. Initially, the application of the GLM often took a classical approach. However, the availability of complex and high-speed software routines has stimulated a rapid growth in Bayesian analyses carried out through GLMs. The fundamental idea of a GLM assumes that the underlying distribution of responses belong to the exponential family of distributions, and a link function transformation of its expectation is modeled as a linear function of observed covariates. Furthermore, it is assumed that the variance of the response is a specified function of its mean. The exponential family of distribution has a pdf which is generally of the form

$$f(\zeta, \theta, \phi) = \exp\left\{\frac{W(\zeta)\theta - b(\theta)}{a(\phi)} + c(\zeta, \phi)\right\} \quad (3.10)$$

where θ and ϕ are the location and scale parameters respectively, while $a(\phi)$ and $c(\phi)$ are known functions. Generally if $W(\zeta) = \zeta$, then

$$f(\zeta, \theta, \phi) = \exp\left\{\frac{\zeta\theta - b(\theta)}{a(\phi)} + c(\zeta, \phi)\right\}. \quad (3.11)$$

The expression above gives the mean and variance of ζ as $E(\zeta) = \partial b / \partial \theta$ and $\text{var}(\zeta) = a(\phi) \partial^2 b(\theta)$.

For the GLM method, the mean denoted by μ is related to the covariates through a link function $\psi(\mu)$. Hence, $\psi(\mu)$ is defined as follows

$$\psi(\mu) = \eta = \sum_{j=1}^p x_j \beta_j = X^T \beta$$

where β is the vector of parameters and is represented as $\beta = \beta_0 + \beta_1 + \dots + \beta_j$ and the maximum likelihood estimates are obtained as iterative solutions of the log

likelihood equations, as given below

$$\frac{\delta \ell}{\delta \beta_j} = 0.$$

and the observations as the vector

$$X = \begin{bmatrix} 1 \\ \vdots \\ x_j \end{bmatrix}$$

For the Bayesian paradigm, the conditional probability of a series of events approximates the product of the probability of events as well as the probability of a given event (Rekkas, 2009). Considering a given parameter θ denoted by $P(\theta)$, with y as a set of observed data. Therefore, $P(\theta|y)$ is the conditional probability of θ when y is true. Since the binary logistic regression model can only assume a y_i taking two values 1 or 0, $y_i (i = 1, 2, \dots, n)$, which follows a Bernoulli probability function.

$$P(Y = 1) = \rho$$

$$P(Y = 0) = 1 - \rho$$

where ρ is the proportion of patients in the category 1 of the response variable and $1 - \rho$ is the proportion of patients in the category 0 of the response variable. Assume we have n samples $\{(y_i, x_i), i = 1, 2, \dots, n\}$. The binary logistic regression model for the data under current study is represented by

$$y_i | \rho_i = Ber(\rho_i)$$

$$\rho_i = Pr(y_i = 1) = F(x_i^k \beta) \tag{3.12}$$

where $x_i = (x_{i1}, \dots, x_{ik})$ is vector of known covariates associated with i th individual, and $\beta = (\beta_1, \dots, \beta_k)$ is the regression parameter. The logistic function transformation is specified by:

$$F(x_i^k \beta) = \frac{e^{(x_i, \beta)}}{1 + e^{(x_i, \beta)}}. \tag{3.13}$$

The likelihood function of expression in (3.13) is derived as follows:

$$\begin{aligned}
 P(\beta|y, x) &= \prod_{i=1}^n \left[g(x_i, \beta) \right]^{Y_i} \left[1 - g(x_i, \beta) \right]^{1-Y_i} \\
 &= \prod_{i=1}^n \left[\frac{e^{(x_i, \beta)}}{1 + e^{(x_i, \beta)}} \right]^{Y_i} \cdot \left[\frac{1}{1 + e^{(x_i, \beta)}} \right]^{1-Y_i} .
 \end{aligned} \tag{3.14}$$

To continue with the Bayesian analysis, it is essential to derive a joint prior distribution over the parameter space. In reality, this is very hard to do because the relationship between the data and the parameters is very complex. The easiest way to overcome this difficulty is to propose either an informative prior or a non-informative prior, but with small precision, avoiding any complaint about the specification of subjective beliefs (O'hagan et al., 1990). In this thesis, we use non-informative normal priors with extremely small precisions which were set to the parameters. Therefore, the next procedure is to obtain the posterior distribution, since the inference is based on the information provided by the posterior distribution.

$$\begin{aligned}
 P(\beta|y, x) &= \ell(\beta) \cdot \prod_{i=1}^n \left[g(x_i^k, \beta) \right]^{y_i} \left[1 - g(x_i^k, \beta) \right]^{1-y_i} \\
 &\cdot \\
 P(\beta|y, x) &\propto P(\beta) \cdot \prod_{i=1}^n \left[\frac{e^{(x_i^k, \beta)}}{1 + e^{(x_i^k, \beta)}} \right]^{y_i} \cdot \left[\frac{1}{1 + e^{(x_i^k, \beta)}} \right]^{1-y_i} .
 \end{aligned} \tag{3.15}$$

Now, we found that expression (3.15) is a complex function of the parameters, and numerical approaches are required in order to obtain the marginal posterior distribution for each of the model parameters. Approximations can be obtained via numerical integration (Naylor & Smith, 1982). Simulation-based methods have proliferated in the last ten years or so, yielding two popular approaches known as importance sampling (Zellner & Rossi, 1984) and Gibbs sampling (Dellaportas & Smith, 1993; Albert & Chib, 1995). Re-sampling techniques, applied to logistic regression for randomized response data, were alternatively proposed by Souza and Migon (2004).

3.2.2 Other Classical Logistic Regression Statistics

In this section, we discuss the goodness of fit to our data. After fitting the logistic regression model parameters, there is need for assessing the significance of each variables with regards to predicting the response variable. Some of the tests we

employed are discussed as follows:

- **Deviance analysis of model fit**

The principle here is that the observed values of the response must be compared with the estimated values from the models with and without the variable under consideration. The comparison between the observed values of the response variable to the predicted values is based on the log likelihood function, and is represented as follows:

$$\sum_{i=1}^k \left\{ p_i \log[\pi(\eta_i)] + (1 - p_i) \log[1 - \pi(\eta_i)] \right\}. \quad (3.16)$$

The comparison of observed to the predicted values using log-likelihood is based on the statistic D written as

$$D = -2 \log \left[\frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}} \right]$$

which we expressed as

$$D = -2[\log(L_r) - \log(L_t)], \quad (3.17)$$

where $\log(L_r)$ is the log-likelihood for the extended model and $\log(L_t)$ is the log-likelihood for the simpler model, and D is the deviance. With large sample sizes, deviance (D) approximately follows a chi-square distribution with $(t - r)$ degrees of freedom. Where t and r are the number of parameters in the saturated and proposed models, respectively. Deviance is used for the assessment of the model's goodness of fit. The combination of the above expressions, deviance (D), can be represented as

$$D = -2 \sum_{i=1}^k \left\{ p_i \log \left[\frac{\pi(\eta_i)}{p_i} \right] + (1 - p_i) \log \left[\frac{1 - \pi(\eta_i)}{1 - p_i} \right] \right\}. \quad (3.18)$$

- **Akaike Information Criteria (AIC)**

This statistical test measures the relative value of a statistical model for a given set of data. AIC is expressed as

$$AIC = 2\delta - 2\ln(L),$$

where L is the likelihood of the model and δ indicates the number of parameters on the model under consideration. AIC is used to select the best model in a set of data. The model with the lowest AIC value of should be given prefer-

ence.

- Wald test

This statistic is used to assess the significance of individual logistic regression coefficients.

$$W = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}.$$

where $\hat{\beta}_1$ is the estimate of the coefficient of the explanatory variable and $SE(\hat{\beta}_1)$ is the standard error of $\hat{\beta}_1$.

- Cox and Snell Pseudo R^2 The R^2 for logistic regression is estimated by Cox and Snell Pseudo and is represented as

$$R^2 = 1 - \left[\frac{\ell(o)}{\ell(\hat{\beta})} \right]^{\frac{n}{2}}.$$

$\ell(o)$ is the log-likelihood of the initial model, and $\ell(\beta)$ is the log-likelihood of the current model.

$$\ell(o) = n \left[\hat{\pi}_o \log \left(\frac{\hat{\pi}_o}{1 - \hat{\pi}_o} \right) + \log \left(1 - \hat{\pi}_o \right) \right].$$

where $\hat{\pi}_o = \sum_{i=1}^k \frac{n_i p_i}{n}$, and n is a vector whose elements are the weight for the i th case.

- Likelihood Ratio Test

This statistics tests for the significance of all the variables included in the logistic regression model.

$$-2 \log \left(\frac{A_o}{A_1} \right) = -2 [\log(A_0) - \log(A_1)]$$

where A_o is the maximum value for the likelihood function of a simple model and A_1 is the maximum value for the likelihood function of a full model. According to (Hosmer & Lemeshow, 2000a,b), a simple model has one variable dropped, while a full model contains all the parameters of interest.

- Deviance Information Criteria (DIC)

The ability to fit complex multilevel models using Markov Chain Monte Carlo (MCMC) techniques presents a need for methods to compare alternative models. The standard model comparison techniques such as AIC and BIC require the specification of the number of parameters in each model. For multilevel

models which contain random effects, the number of parameters is not generally obvious and as such an alternative technique of comparison is demanded. The most widely used of such alternative technique is the Deviance information Criteria (DIC) as suggested by Spiegelhalter et al. (2002). The DIC statistic is a generalization of the AIC, and is based on the posterior mean of the deviance, which is also a measure of model complexity and fit. The deviance is defined as

$$D(\theta) = -2 \log f(y|\theta).$$

since DIC is a measure of model complexity, it considers a measure of the effective number of parameters in a model, and is defined by

$$pD = \bar{D}(\theta) - (\check{\theta}).$$

where $\bar{D}(\theta)$ is the posterior expectation of the deviance, given by

$$\bar{D}(\theta) = -2E \left[\log f(y|\theta) | y \right].$$

and $(\check{\theta})$ is the deviance evaluated at some estimate $\check{\theta}$ of θ . Therefore, we now define the deviance information criteria (DIC) by

$$DIC = \bar{D}(\theta) + pD = 2\bar{D}(\theta) - \hat{\theta}. \quad (3.19)$$

where \bar{D} is the posterior mean of the deviance that measures the goodness of fit, and pD represent the effective number of parameters in the model. In the case of the Bayesian and bootstrapping models, low values of \bar{D} imply a better fit, while small values of pD imply a parsimonious model. pD is higher for a more complex model, and DIC appears to select the correct model. The best fitting model is one with the smallest DIC, as suggested by (Spiegelhalter et al., 2002; Lesaffre & Lawson, 2012).

3.2.3 Assumptions on Model Adequacy

For the classical statistics, we examined different diagnostic plots to assess the performance of the fitted model. The Statistical Analysis System (SAS) produces a set of diagnostic plots as summarized in Table 3.2. The plots either indicate how close the fitted model is to the actual model that can produce the exact values obtained, or highlight the effect of some observed values on the model building.

Table 3.2: Summary of diagnostic plots for a fitted model evaluation

Diagnostic Plot	Usage
Cooks D statistic versus observation number	Evaluate influence of an observation on the entire parameter estimate vector
Dependent variable values versus predicted values	Evaluate adequacy of fit and detect influential observations
Externally studentized residuals (RStudent) versus leverage	Detect outliers and influential (high-leverage) observations
Externally studentized residuals versus predicted values	Evaluate adequacy of fit and detect outliers
Histogram of residuals	Confirm normality of error terms
Normal quantile plot of residuals	Confirm normality and homogeneity of error terms, and detect outliers
Residuals versus predicted values	Evaluate adequacy of fit and detect outliers
Residual-fit (RF) spread plot	side-by-side quantile plots of the centered fit and the residuals show how much variation in the data is explained by the fit and how much remains in the residuals

SOURCE: SAS system Guide 9.4: The RSREG Procedure

3.2.4 Bayesian Prior Distributions

This thesis make use of a full Bayesian approach in estimation, with prior distributions assigned to all the parameters. Bayesian statistics differs from classical statistics in the sense that parameters are regarded as random variables in the former, while a prior distribution has to be specified in order to make inference in the latter. The major challenge in Bayesian statistics is the correct specification of a Bayesian prior distribution, because appropriate prior specification is key in Bayesian modeling. (Gelman, 2002) indicated that the prior distribution is an important part of Bayesian inference, representing information about an uncertain parameter θ which is combined with the probability distribution of the likelihood of new data to produce the posterior distribution. This is then used for future inference on θ . Therefore, necessary precaution should be taken in selecting priors because inappropriate choices of priors may result to wrong inference (Institute et al., 2008).

In specifying priors, a number of points need to be considered. A key point among them is the fact that priors can be tentative. Because inference is assumed to be dependent on prior choice, alternative priors are examined to explore how sensitive the main conclusions are to alterations in the prior. Also, it is important and necessary to allow prior beliefs to be influenced by data. There are different types of prior distributions, some of which are discussed below:

- **Non-informative and Informative prior distributions**

As earlier mentioned, the most key important aspect of Bayesian statistics is the setting up of the right prior to include in the model. It is important at this point to explain the major differences between non-informative and informative prior distributions when specifying the Bayesian prior distributions. Non-informative (vague) priors are used if either little is known about the coefficient values, or one wishes to make sure that prior information plays very little role in the model. This simply means the data is allowed to remain influential in the analysis under consideration. As a result of the objectivity of non-informative priors, the majority of researchers in statistics prefer to make use of it compared to informative priors. The most common choice of non-informative priors is the flat prior, which assigns equal likelihood to all possible values of the parameters. On the other hand, an informative prior distribution summarizes the evidence about the parameters of interest from many sources and often has considerable impact on the posterior distribution.

In addition, for our data to remain influential, this thesis utilizes non-informative priors that will not influence the posterior distribution. We assume a multivariate normal prior on β with a large variance ($\sigma^2 = 1000$) and mean ($\mu_k = 1$), except otherwise stated.

$$\beta_0 \sim N(b_0, \Sigma_0^2).$$

The variance (σ^2) needs to be transformed before it is introduced into the model. Hence, we use $\tau = 0.001$ as the transformed variance. In the case of the Bayesian multilevel regression model, each random effect uses a gamma distribution with $\alpha = 0.1$ and $\beta = 0.01$. This thesis utilizes a multivariate normal (b_0, Σ_0) prior density for the parameter vector β . We also assumed that the prior for the i th component be normal (b_1, S_1^2), while the priors for each component are independent of each other. Therefore,

$$\Sigma_0 = \begin{bmatrix} s_0^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_1^2 \end{bmatrix}$$

and

$$b_0 = \begin{bmatrix} b_0 \\ \vdots \\ b_{1\cdot} \end{bmatrix}$$

Hence, we can write the general formula for prior distribution as follows:

$$p(\beta) \propto \exp \left[-\frac{1}{2} \left(\beta - b_0 \right)' \Sigma_0^{-1} \left(\beta - b_0 \right) \right] \quad (3.20)$$

Since a multivariate normal prior does not have to be made up from independent components, the posterior distribution will be a multivariate normal (b_1, Σ_1) , where

$$\Sigma_1^{-1} = \Sigma_0^{-1} + \Sigma_{ML}^{-1},$$

and

$$b_1 = \Sigma_1 | \Sigma_{ML}^{-1} | \hat{\beta}_{ML} + \Sigma_1 | \Sigma_0^{-1} | b_0.$$

where Σ_{ML} is the covariance matrix of the maximum likelihood estimate (MLE) vector, and its inverse is represented as

$$\Sigma_{ML}^{-1} = X \Sigma^{-1} X$$

while $\hat{\beta}_{ML}$ is the maximum likelihood vector and Σ_0^{-1} is the prior precision matrix.

The posterior distribution refers to the distribution of the parameters after data observation. The estimates of Bayesian inference are obtained by sampling from the posterior distribution. In terms of the Bayesian approach, we can write the posterior distribution as

$$p(\beta|y_i) \propto p(y_i|\beta)p(\beta)$$

where $p(y_i|\beta)$ is the likelihood function and is expressed as:

$$p(y|\beta) = \exp \left[-\frac{1}{2} \left(\beta - b_{LS} \right)' \Sigma_{LS}^{-1} \left(\beta - b_{LS} \right) \right].$$

Hence, the posterior distribution is represented as:

$$p(\beta|y_i) \propto \exp\left[-\frac{1}{2}\left(\beta - b_1\right)' \Sigma_1^{-1}\left(\beta - b_1\right)\right]. \quad (3.21)$$

- **Improper priors**

When the integral over the sample space does not converge, the probability distribution specified for θ is assumed to be improper.

$$p(\theta) \propto 1,$$

As argued by (Lancaster, 2004), an improper prior distribution can lead to posterior impropriety. To establish whether a posterior distribution is proper, past studies indicate that the normalizing constant $\int p(y|\theta)p(\theta)d\theta$ is finite for all y s. When an improper prior distribution leads to an improper posterior distribution, inferences based on the improper posterior distributions are not valid (Institute et al., 2008).

- **Prior for fixed effects**

Fixed-effect parameters have no constraints, and as such can assume any value. A prior distribution for such parameters will need to be defined over the whole real line. The conjugate prior distribution for such parameters is the normal distribution.

- **Normal prior with huge variance**

As the variance of the normal distribution is increased, the distribution becomes locally flat around its mean. As earlier mentioned, fixed effects can assume any value. However, a close examination of the data can narrow the range of values and a suitable normal prior can be found. Generally, the normal prior, $p(\theta) \propto N(0; 10^4)$ will be an acceptable approximation to a uniform distribution. If the fixed effects are very large however, a suitable increase in the prior variance may be necessary.

3.2.5 Bayesian Posterior Distribution via Markov Chain Monte Carlo

The Bayesian approach was first introduced by the Reverend Thomas Bayes. Today the Bayesian concept has gained popularity among many researchers across different fields as a result of its ability to handle complex models. Bayesian methods have also been embraced in other fields of science due to their ability to handle complexity in real-world problems. From the literature it was observed that Bayesian inference has many advantages over classical inference. In addition, Bayesian inference has means of incorporating prior knowledge about the parameters under

consideration since they influence the posterior inference. One of the major differences between classical and Bayesian inferences is that the former regards data as random and considers parameters to be fixed. This means that the values of the parameters are estimated from data using estimations that are random variables. In the case of Bayesian methods, parameters are assumed to follow a probability distribution while model parameters are considered as random variables. The main object of interest in Bayesian inference is the posterior distribution. Classical inference estimation depends solely on approximations as well as asymptotic results. In cases of missing data, classical inferences sometimes replace the missing observation with guesses and the analysis of the data on whether the missing observation were known. In this thesis, we develop some methodologies for the current data which deal with multilevel and complex likelihood functions.

It is important to emphasize the specification of both prior mean and variance. The inclusion of a prior mean provides a prior point estimate for the parameter of interest, while the uncertainty about the estimate is expressed by the variance. Therefore, it is necessary for priors to be selected carefully so that they represent the best knowledge about the parameters. Since we want our data to remain influential in all the analysis carried out, this thesis makes use of non-informative priors. Once there is enough information in the likelihood, a non-informative prior can be confidently used. By Baye's rule, we can express a posterior distribution, as presented by (Ntzoufras, 2011), as:

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \propto P(y|\theta)P(\theta). \quad (3.22)$$

where $P(\theta)$ is the prior distribution, and $P(y|\theta)$ is the posterior distribution. A posterior distribution embodies prior and observed data, and the likelihood is represented as (Ntzoufras, 2011):

$$P(y|\theta) = \prod_{i=1}^N P(y_i|\theta). \quad (3.23)$$

Since both classical Bayesian statistics have the same likelihood function, we can write

$$P(y|\theta) = \prod_{i=1}^N \eta^{y_i} (1 - \eta_i)^{1-y_i}. \quad (3.24)$$

In the context of simulation, sampling directly from the posterior may not be easy in complex models, and as such, there are some alternative sampling techniques which may be helpful. The most commonly used of these sampling techniques are:

- Gibbs sampling
- Markov Chain Monte Carlo (MCMC)

The above two techniques were discussed extensively by (Brooks et al., 2011; Ntzoufras, 2011; O'Neill, 2002). They recommended the Markov Chain approach for high-dimensional problems. Some of their recommendations for solving high-dimensional problems are utilized in this thesis. The MCMC technique is one that allows samples to be drawn from a distribution which has the posterior distribution as its stationary distribution. The MCMC techniques operate by defining a Markov chain whose equilibrium and target density are equal. After the chain has been simulated for the required amount of time for convergence occur, samples are drawn from the simulated chain. For the implementation of Markov Chain Monte Carlo, different algorithms have been developed, such as:

- Gibbs sampling algorithm
- Metropolis-Hasting algorithm

Markov Chain Monte Carlo Sampling Algorithms

The two most common Markov Chain Monte Carlo (MCMC) algorithms are the Metropolis Hastings and the Gibbs samplers. These two algorithms have many variants and extensions that have been developed. These forms and extensions are more advanced and sometimes more peculiar to some problems. In this section, we discuss in detail the Gibbs sampler together with its variants and extensions. The Gibbs sampling algorithm is a special case of the Metropolis-Hastings algorithm in which parameters are drawn from distributions with 100% acceptance rate. It is an alternating conditional sampling. The joint posterior distribution is decomposed into a sequence of simpler conditional distributions, where the target is to generate a data point from the conditional distribution of each parameter, conditional on the current values of the other parameter. The Gibbs algorithm implements MCMC methods using the WinBUGS software. The default option in the WinBUGS software for well-behaved models with log concave densities is the Gibbs sampling algorithm. The procedure for the Gibbs sampling algorithm is as follows:

Stage 1: From the conditional posterior distribution, sample a new set of fixed effects

$$p(\lambda|y, \vartheta_m^2, \vartheta_n^2, m) \propto \psi(y; \lambda, \vartheta_n^2)p(\lambda)$$

$$p(\lambda|y, \vartheta_m^2, \vartheta_n^2, m) \propto \left(\frac{1}{\vartheta_n^2}\right)^{\frac{Q}{2}} \prod_{i,j} \exp\left\{-\frac{1}{2\vartheta_n^2}(y_{ij} - m_j - (x_{ij}\lambda)^2)\right\}.$$

which implies that we can take the sample from the expression

$$\lambda \sim N(\hat{\lambda}, \Sigma).$$

where

$$\hat{\lambda} = \left[\sum_{i,j} X'_{ij} X_{ij} \right]^{-1} \left[\sum_{i,j} X'_{ij} (y_{ij} - m_j) \right].$$

and

$$\hat{\Sigma}^* = \vartheta_n^2 \left[\sum_{i,j} X'_{ij} X_{ij} \right]^{-1}.$$

Suppose we use a multivariate normality for λ with covariance matrix V_0 . Then we sample from

$$\lambda \sim N(\hat{\lambda}^*, \hat{\Sigma}^*).$$

Using the same explanation above, we have

$$\hat{\lambda}^* = \left[\sum_{i,j} X'_{ij} X_{ij} + \vartheta_n^2 V_0^{-1} \right]^{-1} \left[\sum_{i,j} X'_{ij} \check{y}_{i,j} \right].$$

and

$$\hat{\Sigma} = \vartheta_n^2 \left[\sum_{i,j} X_{ij}^T X_{ij} + \vartheta_n^2 V_0^{-1} \right]^{-1}.$$

Note that $\check{y} = \left[y_{ij} - (\Omega m)_{ij} \right]$.

Stage 2: In the second stage, a new set of residuals can be sampled and each residuals (m_j) is assumed to have a prior distribution $m_j \sim N(0, \vartheta_m^2)$. We can now derive the following posterior distribution

$$p(m_j|y, \vartheta_m^2, \vartheta_n^2) \propto \left(\frac{1}{\vartheta_n^2}\right)^{\frac{q_i}{2}} \prod_{i,j} \exp\left\{-\frac{1}{2\vartheta_n^2}(y_{ij} - x_{ij}\lambda - m_j)^2\right\} \times \left(\frac{1}{\vartheta_m}\right) \exp\left\{-\frac{1}{2\vartheta_m^2} m_j^2\right\}.$$

Hence, we can sample from

$$m_j \sim N(\hat{m}_j, \hat{\Sigma}).$$

where

$$m_j = \left[q_j + \vartheta_n^2 \vartheta_m^{-2} \right]^{-1} \left[\sum_{i=1}^{q_i} (y_{ij} - X_{IJ}) \lambda \right],$$

$$\hat{\Sigma} = \vartheta_n^2 \left[q_j + \vartheta_n^2 \vartheta_m^{-2} \right]^{-1}.$$

Stage 3: Here a new level 2 variance is sampled. Within this area of research, the choice of an appropriate prior distribution is an issue that hasn't been resolved. In this study, we adopted an inverse gamma prior, since it does not produce biased estimates like uniform priors. We can now sample from

$$\vartheta_m^{-2} \sim \text{Gamma}(\nu_m, \zeta_m).$$

where ϑ_m^{-2} is called precision.

$$\nu_m = \frac{j + 2\epsilon}{2}.$$

$$\zeta_m = \left(\epsilon + \sum_{j=1}^j \frac{m_j^2}{2} \right).$$

where j is the number of level 2 units. In the case of a uniform prior for ϑ_m^2 ,

$$\nu_m = \frac{j - 2}{2}.$$

$$\zeta_m = \sum_{j=1}^J \frac{\vartheta_m^2}{2}$$

Stage 4: a new level 1 variance is sampled:

$$\vartheta_n^{-2} \sim \text{Gamma}(\nu_n, \zeta_n),$$

$$\nu_n = \frac{N + 2\epsilon}{2},$$

and

$$\zeta_n = \frac{\sum_{ij} e_{ij}^2}{2}.$$

Stage 5: Compute the level 1 residuals.

For Bayesian analysis in WinBUGS, there is need to specify the model, likelihood, prior, and observed data, and set the initial values. WinBUGS then produces an appropriate Markov chain. WinBUGS requires much less MCMC programming than if one was to program the MCMC simulations by directly. In addition, there is need to ensure that a sequence has reached convergence before inferences are collated. The number of iterations performed for a practical convergence to equilibrium is a function of many factors such as:

- complexity of the model
- whether the prior and likelihood are conjugate
- closeness of the initial values to their respective posterior means
- kind of sampling method adopted

The Metropolis-Hastings algorithm is an iterative algorithm. It produces a Markov chain and allows empirical estimation of posterior distributions as well as the generation of samples from a probability distribution. The Metropolis-Hastings algorithm steps are outlined as follows:

- step1: starting values K for the parameter: $\theta^{i=0} = K$, where $i=1$
- step2: draw a candidate parameter θ^i from a proposal density $\propto (\cdot)$
- step3: compute the ratio $J = \frac{f(\theta^i) \propto (\theta^{i-1} | \theta^b)}{f(\theta^{i-1}) \propto (\theta^b | \theta^{i-1})}$
- step4: compare J with a $T(0, 1)$ random draw, t . If $J > t$, set $\theta^i = \theta^b$, otherwise, set $\theta^i = \theta^{i-1}$
- step 5: set $i = i + 1$ and return to step 1 until enough draws are obtained.

Metropolis- Hastings algorithm increase the efficacy with the utilization of a random walk process through the parameter space.

MCMC techniques via WinBUGS

WinBUGS is a product of Bayesian Inference using Gibbs Sampling (BUGS) that is managed by the Medical Research Council (MRC) of Biostatistics Unit in the United Kingdom. It is a free software that uses MCMC methods in complex Bayesian techniques (www.mrc-bsu.cam.ac.uk/bugs). WinBUGS models can be implemented in two ways namely by using the

- BUGS language
- graphical feature, Doodle BUGS, which allows the specification of models in terms of a directed graph.

The model types that can be fitted using WinBUGS are wide-ranging. A wide variety of models, including the multilevel, multinomial Dirichlet models can be fitted using WinBUGS. WinBUGS is a very powerful tool in fitting complex models, although it is a difficult and frustrating package to master. One of the main issues of WinBUGS is that there is a great amount of embedded statistical theory associated with MCMC. The other issue associated with Bayesian modeling is the choice of priors and initial values. The Bayesian analysis using WinBUGS requires three major tasks:

- model specification
- running the model
- bayesian inference using mcmc output

In WinBUGS, there are three types of nodes: constant, stochastic and deterministic. Constant nodes are meant for the declaration of constant terms, while stochastic nodes represent data or parameters that are assigned a distribution. Deterministic nodes are logic expressions of other nodes. In addition, WinBUGS expects each node to appear exactly once on the left-hand side of an equation. After MCMC simulation, WinBUGS provides several numerical and graphical summaries of the parameters. Convergence is assessed by either trace plots, Geweke, Gelman-Rubin or autocorrelation plots. CODA/BOA packages are also accessible from R for further convergence analysis. The Brooks-Gelman-Rubin (BGR) convergence statistic is used to assess the convergence of models and operates on multiple chains. Convergence is based on the equality of within-chain variability and between-chain variability, since simulations from multiple chains are independent.

Because of the complexity of the posterior distribution, it is difficult to directly sample from, and becomes even more complicated when random effects are included in the model. We employed the MCMC approach to the simulation of the random

numbers from the posterior distribution. In reality, a posterior distribution is often of higher dimension and analytically intractable, and when the posterior distribution comes from a distribution that is complex, the MCMC approach offers a better alternative for summarizing the posterior distribution. The quality of the MCMC sample depends on how quickly the sampling procedure explores the posterior distribution.

3.2.6 Markov Chain Monte Carlo Simulation

MCMC methods are an established suite of methodologies that enable samples to be drawn from some target density that is only known up to proportionality. It is a technique of drawing values of θ from approximate distributions and then correcting those draws to better approximate the main posterior distribution $p(\theta|y)$. These samples are drawn sequentially with the distribution of the sampled draw, depending on the last value drawn. This results into a Markov chain. The success of the MCMC technique depends on the approximate distribution which is improved by a simulation of each step until a convergence of the posterior distribution is achieved. The two most popular used algorithms: Gibbs sampler and Metropolis-Hastings are special cases of the MCMC algorithm.

The approach is to take a Bayesian framework and carry out the necessary numerical integrations using simulation. Instead of calculating exact or approximate estimates, the computer-intensive techniques generate a stream of simulated values for each quantity of interest. We focus here on one of such techniques, the Gibbs sampler, which has emerged as the most widely used. Gibbs sampling is an MCMC method for obtaining marginal distributions from a non-normalized joint density. It is a simulation tool for generating samples from the joint posterior distribution of unknown quantities in a model, conditional on the observed data.

In Bayesian inference, the model assumes that data are fixed and regards the parameters as random variables. Suppose we consider a given parameter θ and a set of observed data, the interest of the Bayesian approach is the probability of the parameter θ given the set of observed data y . This is written as $P(\theta|y)$. The next procedure involves obtaining the posterior distribution of the unknown parameter θ , given the observed data y . This procedure is achieved by multiplying the prior distribution with the likelihood function.

Markov Chain Convergence Diagnostics

For any simulation to be carried out properly, it is necessary that the MCMC algorithm be supplied with different starting values. The non-convergence problem in MCMC simulations may be as a result of the simulations not being a true represen-

tation of the desired distribution, based on the influence of initial values in the early part of the chain and the inability of the within-chain serial correlation. According to (Gelman et al., 2014b), these difficulties are dealt with by simulating multiple chains with different starting values distributed through the parameter space, monitoring convergence, and discarding the early iterations of the simulation. Autocorrelation can be reduced by applying thinning, which results in keeping some parts of the simulation after discarding some. There are several ways to assess or monitor whether parallel chains have converged or not, some which are explained by (Gelman et al., 2003). The Gelman-Rubin statistic is used to assess the convergence of chains separately for all parameters under consideration (Gelman et al., 2014a,b; Brooks & Roberts, 1998). In a situation where the sampling is to continue indefinitely, convergence will be monitored by estimating the factor by which the scale parameter might shrink.

Suppose we have N samples of ω from each of C chains and represent this as w^{nc} . We define the within-sequence mean \bar{w}^c and overall mean \bar{w} as:

$$\bar{w}^c = \frac{1}{N} \sum_{n=1}^N w^{nc}$$

$$\bar{w} = \frac{1}{C} \sum_{c=1}^C \bar{w}^c.$$

We now define the between chain variance (B) and within chain variance (T) as follows:

$$B = \frac{N}{C-1} \sum_{c=1}^C (\bar{w}^c - \bar{w})^2. \quad (3.25)$$

$$T = \frac{1}{C} \sum_{c=1}^C \left[\frac{1}{N-1} \sum_{n=1}^N (\bar{w}^{nc} - \bar{w}^c)^2 \right]. \quad (3.26)$$

After that, we construct two estimates of the variance of w . The first estimate of the variance of w is expected to underestimate the $\text{var}(w)$ if the chains have not ranged over the full posterior, while the second estimate, expressed as

$$\hat{v} = \frac{N-1}{N} B + \frac{1}{N} T$$

is an estimate of the $\text{var}(w)$ that is unbiased when equilibrium (stationary) conditions are reached, but is an overestimate when the starting points were over-

dispersed. The test statistic for the Gelman-Rubin diagnostic test can be estimated as follows:

$$\hat{R} = \sqrt{\frac{N-1}{N} + \frac{B}{NT}}.$$

\hat{R} is called the estimated potential scale reduction factor (PSRF) and measures the degree to which the posterior variance would decrease if we were to continue sampling by increasing N . If the potential scale reduction factor is greater, say above 2, then further simulations are required, but when $\hat{R} \rightarrow 1$, say $\hat{R} < 1.1$ is an indicator of convergence. This simply means the variance between the chains is similar to the variance within each chain. The corrected version of the \hat{R} statistic was defined by Brooks and Roberts (Brooks & Roberts, 1998), and is expressed as $\hat{R}_c = \frac{a+3}{a+1} \hat{R}$, where a is the estimate of the degrees of freedom for the pooled posterior variance estimate. Raftery and Lewis's diagnostic (Brooks & Roberts, 1998) determines the minimum number of iterations based on minimal autocorrelation, and the required sample size and length of burn-in process for a single chain. The Geweke diagnostic compares values in the first part of the Markov chain analysis to those in the latter stage to detect failure of convergence (Ntzoufras, 2011). Geweke's statistic has an asymptotically standard normal distribution, and is expressed as

$$Z_k = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{1}{K_1} S_1(0) + \frac{1}{K_2} S_2(0)}} \rightarrow N(0, 1), n \rightarrow \infty. \quad (3.27)$$

where $S_1(0)$ and $S_2(0)$ are respectively classical estimates of the respective variances. An inability to reach convergence in MCMC sampling may reflect problems in model identifiability due to overfitting. To overcome such problems, running multiple chains often helps by diagnosing poor identifiability. This is illustrated mostly when identifiability constraints are missing from a model, such as in discrete mixture models that are subject to label switching during MCMC updating (Frühwirth-Schnatter, 2001). One chain may have a different label from others, so that obtaining the Gelman-Rubin (G-R) statistic for some parameters is not sensible. A choice of diffuse priors tends to increase the chance of poorly identified models, especially in complex hierarchical models or small sample datasets as revealed by Gelfand et al. (1995). Correlation between parameters within the parameter set $= (\theta_1, \dots, \theta_k)$ increases the dependence between successive iterations, while informative priors may help in identification and convergence. As reported by (Zuur et al., 2002) re-parameterization measures aimed at reducing correlation such as centering predictor variables in regression usually improve convergence. According to (Heidelberger & Welch, 1983), the Heidelberger-Welch (HW) diagnostic is an automated test for checking the sta-

tionarity of the chain and further evaluating whether the length of the chain is sufficient to ensure the desired accuracy for the posterior means of the parameters in Bayesian analysis. This test is based on the Cramer-von Mises test statistic which decides to either accept or reject the null hypothesis that the chain is from a stationary distribution. The test will first check for stationarity, and thereafter determine the accuracy of the model parameters.

The advantage of the Bayesian method is that when the posterior distribution is simulated, the uncertainty of the parameter estimates is taken into account. That is, the uncertainty in the parameter estimates for the fixed part is taken into account in the estimates for the random part. Moreover, simulating a large sample from the posterior distribution is useful because it provides not only point estimates of the unknown parameters, but also confidence intervals that do not rely on the assumption of normality for the posterior distribution. Hence, credible intervals are also accurate for small samples dataset (Tanner & Wong, 1987). The number of MCMC iterations required are very large when the sample size is very small, since MCMC techniques do not perform very well with small datasets. This may lead to high autocorrelation among the parameters, particularly when estimating the mean. This is why up to 1,500,000 iterations were run in all our Bayesian models with a lag of 60 to reduce autocorrelation, which may have necessitated even more iterations.

Many other diagnostic tools have been proposed by (Ntzoufras, 2011) to assess convergence, and compared by (Brooks & Roberts, 1998). Convergence can also be implemented in the CODA/BOA package for R. Four different diagnostics are provided by CODA as indicated by (Little & Wang, 1996; Erkanli et al., 1999; Dunson & Colombo, 2003; Heidelberger & Welch, 1983). In this thesis, visual of diagnostic plots and Gelman-Rubin diagnostic ($\hat{R} \rightarrow 1$) were the main approaches to assessing convergence.

3.2.7 Some key points derived from the classical statistics

- The parameters of the population are unknown fixed constants.
- Statistical procedures have a long-term meaning, like an infinite repetition of the same experiment.
- Probabilities are interpreted as a frequency after a long number of experiments.

3.2.8 Some key points derived from the Bayesian inference

- Parameters are regarded as random variables, as we are not certain of the real values.
- The way to make inference is just the use of the rules of probabilities.
- The inclusion of a prior in the Bayesian model varies across individual researchers.
- There can be a continuous revision of our beliefs as data come to our hand.

That last two points of Bayesian inference makes them even more related to real life situations, and as a result, a more sensible and natural way of quantifying problems.

3.3 Advantages and Disadvantages of the Bayesian Approach

Bayesian and classical statistics both have advantages and disadvantages, as well as some similarities. When the sample size is large, Bayesian statistics often provides results that are equivalent to those obtained by classical statistics.

3.3.1 Advantages of Bayesian Statistics over Classical Statistics

Some advantages of using Bayesian statistics over classical statistics are outline as follows (Bolstad & Curran, 2007; Institute Inc, 2008)

- Bayesian statistics uses a single tool, Bayes Theorem, which is applicable in all situations. This contrasts to classical statistics that require many tools.
- Bayesian statistics allows for the incorporation of prior information in addition to the data that helps in strengthening inferences about the unknown parameters and can help in reducing necessary sample sizes.
- Bayesian statistics include uncertainty in the probability model, which leads to a more realistic prediction. Classical statistics do not include the uncertainty of the parameter estimates, which makes them produce less realistic predications compared with Bayesian statistics.
- Bayesian statistics estimate the full probability model, while classical statistics do not.
- Bayesian statistics have an axiomatic foundation that is uncontested by classical statistics. In addition, Bayesian statistics are coherent to a classical statistician.

- Bayesian statistics have the strength to compare multiple models with different techniques using Deviance Information Criteria (DIC), including multilevel models, but classical statistics cannot.
- Bayesian statistics are unaffected by the overfitting of a model, unlike classical statistics where the overfitting of model is a serious problem.
- Bayesian statistics use credible interval to make decisions, which makes their conclusions more robust than classical statistics which use confidence intervals.
- Bayesian inference via MCMC is unbiased with the size of sample size but classical statistics is biased when the sample size is small and the sample is sparse.
- Bayesian statistics via MCMC algorithms have a theoretical guarantee that the MCMC algorithm will converge if it is run long enough, unlike classical statistics where there are no guarantees for the convergence of the MLE.

3.3.2 Advantages of Classical Statistics over Bayesian Statistics

- Classical statistics models are not difficult to specify, since there is no need to specify prior distribution, initial values for numerical approximation, and the likelihood function, as is the case with Bayesian statistics.
- Bayesian statistics does not specify the approach to be taken in selecting a prior, which means there is no correct way to choose a prior.
- Bayesian statistics can produce posterior distributions that are strongly influenced by the priors. From a practical point of view, this leads to a lot of argument among many researchers.
- Classical models have a much shorter run time compared to Bayesian methods, particularly when the Bayesian method uses the WinBUGS software. For instance, simple classical statistics models may be run in minutes, which is not possible in the case of Bayesian statistics.

3.4 Generalized Linear Models

This section gives a brief description of the development of generalized linear models. In most cases, when responses are not distributed normally, a linear model may

not be appropriate. From a Bayesian point of view, the challenges of any model fitting are developing a methodology for estimating the parameters of the model, as well as predicting any unobserved random variables. The simplest model, and the most widely used, is the linear model. Its main assumptions are that (i) the observations are independent, (ii) the mean equals a linear combination of the predictors, and (iii) the variance of the response is constant for every observation. An additional fourth assumption is sometimes made, that (iv) the observations are a sample from the Normal distribution. Procedures for fitting linear models which are very easy to implement have been developed. However, the above assumptions are not always satisfied, therefore the use of more general models is necessary. Such models include the Generalized Linear Model (GLM) and the Generalized Linear Mixed Model (GLMM). The theory of linear model was generalized by (Nelder & Baker, 1972) to a family of models called generalized linear models (GLMs). In the concept of linear model, $E(\zeta) = \eta\beta$. The Link function $g(\cdot)$ was introduced by (Nelder & Baker, 1972) in order to transform the mean of the model to a linear scale. This is written as

$$g[\zeta] = \eta\beta. \quad (3.28)$$

The response variable ζ in a generalized linear model is no longer restricted to being normally distributed. Instead, it follows a wider class of distributions such as the exponential distribution.

3.4.1 Model Formulation

A member of the exponential family has a probability density function represented as follows:

$$f(\zeta) = \exp\{W(\zeta) \cdot Q(\theta) - B(\theta) + C(\zeta)\} \quad (3.29)$$

where $W(\zeta)$ is sufficient statistic, $Q(\theta)$ is the natural parameter and $B(\theta)$ is the normalization factor. Generally, an exponential distribution cannot have a support that varies according to a parameter; it must remain the same across all distributions in the family. If $Q(\theta) = \theta$, one can simply convert the exponential family distribution to canonical form. In addition, if vector θ in the above expression turns to a scalar, the expression becomes

$$f(\zeta, \theta, \phi) = \exp\left\{\frac{W(\zeta)\theta - b(\theta)}{a(\phi)} + c(\zeta, \phi)\right\} \quad (3.30)$$

where θ and ϕ are the location and scale parameters respectively, while $a(\phi)$ and $c(\phi)$ are known functions. Generally, if $W(\zeta) = \zeta$, then

$$f(\zeta, \theta, \phi) = \exp\left\{\frac{\zeta\theta - b(\theta)}{a(\phi)} + c(\zeta, \phi)\right\}. \quad (3.31)$$

The expression above gives the mean and variance of ζ as $E(\zeta) = \partial b/\partial\theta$ and $\text{var}(\zeta) = a(\phi)\partial^2 b(\theta)$.

Components of the Generalized Linear Model

The components of Y are independent normal variables with constant variance σ^2 and

$$E(Y) = \mu, \quad \mu = \pi\beta, \quad (3.32)$$

The expressions above are slightly rearranged to the three (3) components of a GLM: Random component: the components of Y are independent normal variables with constant variance σ^2 , Systematic component: covariates π_1, \dots, π_p produce a linear predictor written as:

$$\eta = \sum_1^p x_j \beta_j \quad j = 1, \dots, p.$$

Link function: The link between the two foregoing components of a generalized linear model is given by the identity $\eta = \mu$. Thus, the formula for the components of the linear predictor in terms of a so-called link function is written as

$$\eta_i = h(\mu_i) \quad i = 1, 2, \dots, n,$$

where $h(\cdot)$ is the link function. This is an expression which connects the parameters of the response ζ with the linear predictor and explanatory.

Table 3.3: Some common distributions of exponential dispersion family with their link functions (Ntzoufras, 2011)

Distribution	Link name	Inverse link	Variance function
Normal	identity	η	ϕ
Binomial	logit	$e^\eta/(1 + e^\eta)$	$\mu(1 - \mu)/N$
Poisson	logarithms	e^η	μ
Gamma	reciprocal	$e^{-\eta}$	$\phi\mu^2$

3.4.2 Parameter Estimation

Once a particular model is selected, there is need to estimate its parameters. In the case of the GLM, the estimators of the parameters are obtained by using a maximum likelihood method.

The Maximum Likelihood Method in GLM

The log-likelihood function for each of the components of the random vector ζ is

$$l_i(\theta_i, \phi, \zeta_i) = \exp\left\{\frac{\zeta_i - b(\theta)}{a(\phi)} + c(\zeta, \phi)\right\} \quad (3.33)$$

where we consider $l_i(\theta_i, \phi; \zeta_i)$ to be a function of θ_i and ϕ with ζ_i being given. The log likelihood components of response ζ is

$$\ell(\theta, \phi, \zeta) = \sum_{i=1}^n l_i(\theta_i, \phi, \zeta_i). \quad (3.34)$$

For the Maximum Likelihood Estimators (MLE) of model parameters β , we have

$$\begin{aligned} \psi = \frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^n \left[\frac{\partial l_i}{\partial \beta} \right] = \sum_{i=1}^n \left[\frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_i} \right] \\ &= \frac{1}{\phi} \sum_{i=1}^n \left[\frac{\zeta_i - \mu_i}{\text{Var}(\zeta_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) x_i \right]. \end{aligned} \quad (3.35)$$

The MLE $\hat{\beta}$ of the vector of parameters is then obtained as follows:

$$0 = \psi = \frac{1}{a(\phi)} \sum_{i=1}^n \left[\omega(\hat{\mu}_i) \left(\frac{\partial \hat{\eta}_i}{\partial \hat{\mu}_i} \right) x_i \right] \quad (3.36)$$

where the function $\omega(\cdot)$ is of the form

$$\omega(\hat{\mu}_i) = \frac{1}{\text{Var}(\zeta_i)} \left(\frac{\partial \hat{\mu}_i}{\partial \hat{\eta}_i} \right)^2, \quad i = 1, \dots, n$$

By adding this term

$$(X^T \hat{W} X) \hat{\beta} = \left[\sum_{i=1}^n \omega(\hat{\mu}_i) x_i x_i^T \hat{\beta} \right]$$

where $\hat{W} = \text{diag}\{\omega(\hat{\mu}_1), \dots, \omega(\hat{\mu}_n)\}$ to both sides of expression (3.33) and rearranging, we obtain

$$(X^T \hat{W} X) \hat{\beta} = X^T \hat{W} \hat{Z}$$

vector $\hat{Z} = (Z_1, \dots, Z_n)$ denotes an adjusted response, where $Z_i = x_i^T \hat{\beta} + \frac{\partial \eta_i}{\partial \mu_i} (\zeta_i - \hat{\mu}_i)$, $i = 1, \dots, n$

3.5 Family of Binomial Logit Link Function with Hierarchical Structure

Overview of Hierarchical Model

Multilevel regression models are common in many areas including political, social, and health research. Multilevel models have been illustrated (Gelman & Hill, 2007) in political research by developing a model to estimate state-level opinions from national polls, and also modeling the relationship between income and voter preference by state. Most examples of multilevel models in social science research have been put together by (Gelman & Hill, 2007), including a multilevel model of police stops data. An example of multilevel models in a health-related issue is given by (Gary-Webb et al., 2010), who modeled neighborhood and weight-related health behaviors. The usefulness of multilevel models has also been discussed in a research work credited to public health (Diez-Roux, 2000). Multilevel regression models give coefficients that vary by group a probability model, which allows the variation between groups to be modeled. This differs from the traditional regression approach which accounts for varying coefficients by using indicator variables. The probability models for the coefficients can themselves be given a probability model, and so on. Suppose we have a simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon, \quad i = 1, \dots, n \quad (3.37)$$

Assuming the data is divided into J groups, then the regression model in expression (3.37) becomes

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, J \quad (3.38)$$

The model can be expanded to a multilevel by giving probability models to β_{0j} , β_{1j} or both. For instance, the second level model could be

$$\beta_{0j} \sim N(\beta_0, \sigma_{\beta_0}^2)$$

and

$$\beta_{1j} \sim N(0, \sigma_{\beta_1}^2).$$

A more complex model sometimes places regression models in the intercepts as well as slopes, such as

$$\beta_{0j} = \alpha_0 + \alpha_1 w + \vartheta_j$$

and

$$\beta_{1j} = \pi_0 + \pi_0 w + v_j$$

for some predictor, w , and error distributions for ϑ_j and v_j .

3.5.1 Multilevel Generalized Linear Models

Generally, a generalized linear mixed model (GLMM) assumes the distribution of the random effects, b , to be multivariate normally distributed. Such an assumption can be relaxed to allow other random effect distributions. An approach was proposed by (Lee & Nelder, 1996), called hierarchical generalized linear models (HGLM), to loosen assumptions about the distributions of random effects in the models. Hierarchical generalized linear models (HGLM) extend GLMMs in such a way that the distribution of random effects needs not be normally distributed, but can be a member of the exponential family of distributions with the link function. Other ways by which HGLM extended GLMM are as follows: the mean and dispersion can be modeled jointly; additional random effects are allowed to be added into the linear predictor for the mean so that the random effects can follow any conjugate exponential family of distributions, termed the generalized linear model distribution; and structured dispersion components that depend upon covariates and the visualization of model diagnostics can be made. Given the dispersion components, estimating both the fixed and random effects reduces to estimating an augmented GLM. Given the fixed and random effects, estimating dispersion components with the use of an adjusted profile h-likelihood reduces to a second interlinked GLM. That is, HGLM can be expressed as two interlinked GLMs: one GLM is for the mean, while the other is for the dispersion, ϕ . HGLMs have several merits, two of which are that a single algorithm is sufficient to fit all models of the class, and that model fitting requires neither numerical integration methods nor prior distributions as used in marginal likelihood and Bayesian methods, respectively. Table 3.4 below shows the commonly used distributions and link functions (Lee & Nelder, 1996)

Table 3.4: Some common distributions with their link functions

Model	Link function, $g(\mu)$	b distribution	Link, $v(b)$
LMM	identity	normal	identity
Binomial GLMM	logit	normal	identity
Poisson GLMM	logarithms	normal	identity
Gamma GLMM	logarithms	normal	identity

3.5.2 Multilevel Logistic Regression Model

Multilevel models are regarded as extensions of linear regression models, which accounts for the nesting of data within higher order units as in the case of this study. In this section, we consider a model where the hospital is regarded as a level-2 unit, while the patient is a level-1 unit. Failure to account for the nesting of observations can result in wrong interpretation of results. The multilevel solution is to improve on the ordinary regression model by including error an parameter, so that we can capture group level dependencies in our data. Let j denote the level-2 units (hospitals), and i , the level-1 units (individuals). We assume that there are $j = 1, \dots, T$ level-2 units and $i = 1, \dots, t_j$ level-1 units nested within each level-2 unit. Then, the total number of level-1 individuals across level-2 units is represented by $t = \sum_{j=1}^T t_j$. Assume Y_{ij} is the value of the dichotomous response variable, coded as 0 or 1, associated with a level-1 unit, i , nested within a level-2 unit, j . The logistic regression model can now be written in terms of the log odds of the probability of a response, represented as $P_{ij} = \Pr(Y_{ij}=1)$. Therefore, considering the normal logistic regression model with a single random effect gives

$$\log \frac{p_{ij}}{1 - p_{ij}} = x'_{ij}\beta + \sigma_j \tag{3.39}$$

where x_{ij} is a $cx1$ covariate vector (includes a 1 for the intercept), β is a $cx1$ vector of unknown regression parameters and σ_j is the random effect. These are assumed to be distributed as $N(0, \sigma_\delta^2)$. Considering the fact that our datasets are categorical in nature, for computational simplicity mostly in models for categorical outcomes, the random effects are normally given in standardized form. Therefore, $\delta_j = \sigma_\delta \theta_j$ and the model in (3.39) is now expressed as follows

$$\log \frac{p_{ij}}{1 - p_{ij}} = x'_{ij}\beta + \sigma_\delta \theta_j \tag{3.40}$$

where σ_δ is the random effect variance term which is now included in the analysis of the regression model. For the multilevel representation of our model with one level-1 covariate and one level-2 covariate, we express the level-1 model in terms of

the logit as (De Leeuw et al., 2008):

$$y_{ij} = \log \frac{p_{ij}}{1 - p_{ij}} = \beta_{0j} + \beta_{1j} X_{ij} \quad (3.41)$$

where $\beta_{0j} = \beta_{00} + \beta_{01} X_j + \sigma_{0j}$ and $\beta_{1j} = \beta_{10}$.

Hence, model (3.41) is now

$$y_{ij} = \beta_{00} + \beta_{10} X_{ij} + \beta_{01} X_j + \sigma_{0j}. \quad (3.42)$$

As shown in the model (3.42), the log odds of a patient i in hospital j taking surgery is determined by the log odds of a particular patient in a particular hospital (β_{00}), the effect of the patient level ($\beta_{10} X_{ij}$) and the hospital level ($\beta_{01} X_j$), as well as the hospital level error [$\sigma_{0j}, \sigma_{0j} \sim N(0, \tau_{00})$]. The logistic cumulative (probability) is now represented as:

$$\Psi = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

where Ψ is the logistic cumulative distribution while η is the log odds of taking surgery.

3.6 The Multinomial Logistic Regression Model

The Multinomial logistic regression model is a technique of analysis which is applicable when the dependent variable under study consists of more than two categories. The multinomial response could be ordinal (ordered categories) or nominal (unordered categories). Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases. Sample size guidelines for multinomial logistic regression indicate a minimum of 10 cases per independent variable (Chan, 2005).

While a binary logistic regression model compares one dichotomy, a multinomial logistic regression model compares a number of dichotomies. A multinomial approach outputs a number of logistic regression models that make specific comparisons of the response categories. Considering a situation where we have j categories of the response variable, the model consists of $j - 1$ logit equations which are fit simultaneously. The probability of a categorical variable in a multinomial model is estimated using maximum likelihood estimation (Bayaga, 2010).

Model formulation

Unlike binary logistic models, multinomial logistic regression models pair each outcome category with a reference category. The last category or probably the most common category is assumed to be picked as reference J Agresti & Kateri (2011). Suppose $x_i = (x_{i0}, \dots, x_{im})^T$ denote the explanatory variables for individual $1 \leq i \leq n$ with m covariates X_1, \dots, X_m and the outcome variable has t categories ($t > 2$). The $\beta_j = (\beta_{j0}, \dots, \beta_{jm})$, ($1 \leq j \leq J - 1$), a row vector, represent the regression parameters for the j th category. Suppose $y_i = (y_{i1}, \dots, y_{iJ})$ denote a multinomial trial for individual $1 \leq i \leq n$. The trial y_{ij} is equal to 1 anytime a trial occurs in category j and $p(y_{ij} = 1|X_i)$ is the probability that the i th trial occurs in category j given a set of covariates x_i . Then the multinomial logistic regression model for $\Pr((y_{ij} = j)|X)$, ($1 \leq j \leq t - 1$) is represented as (Agresti & Kateri, 2011):

$$\Pr((y_{ij} = 1)|X) = \log\left(\frac{\Pr((y_{ij} = 1)|X)}{\Pr((y_{ij} = t)|X)}\right) = \beta_0^1 + \beta_1^1 b_1 + \dots + \beta_k^1 b_k \quad (3.43)$$

$$\Pr((y_{ij} = 2)|X) = \log\left(\frac{\Pr((y_{ij} = 2)|X)}{\Pr((y_{ij} = t)|X)}\right) = \beta_0^2 + \beta_1^2 b_1 + \dots + \beta_k^2 b_k \quad (3.44)$$

⋮

$$\Pr((y_{ij} = t - 1)|X) = \log\left(\frac{\Pr((y_{ij} = t - 1)|X)}{\Pr((y_{ij} = t)|X)}\right) = \beta_0^{t-1} + \beta_1^{t-1} b_1 + \dots + \beta_k^{t-1} b_k \quad (3.45)$$

where the category t is chosen as the reference category, β_0^j denotes the intercept, and $\beta_1^j, \dots, \beta_k^j$ denotes the regression coefficients of b_1, b_2, \dots, b_k in the j th category respectively, ($1 \leq j \leq t - 1$).

Cumulative logit model

The object of interest in ordinal data is to model $\Pr(Y = j|x_i)$, $j = 1, 2, \dots, t$, where t is the total number of ordered categories, with explanatory variables x_i . This thesis makes use of cumulative logistic regression models.

$$\text{logit}[\Pr((Y_i \leq j)|x_i)] = \frac{\Pr(Y_i \leq j|x_i)}{1 - \Pr(Y_i \leq j|x_i)} = \alpha_j + x_i^T \beta, j = 1, \dots, t - 1 \quad (3.46)$$

where $\alpha_j = \alpha_0 < \alpha_1 < \dots < \alpha_{t-1} < \alpha_t$, Y_i is the response for the i th individual and β is the fixed effects parameter. Suppose we have a model with t ordered, it implies that there are $t - 1$ logits. Then the j th cumulative probability is represented

as follows

$$Pr((Y_i \leq j)|x_i) = \frac{1}{1 + \exp[-(\alpha_j + x_i^T \beta)]}, \quad j = 1, \dots, t \quad (3.47)$$

when $j = 1, 2, \dots, t$, its probabilities are

$$\begin{aligned} \phi_j &= P(Y_i \leq 1) \\ &\vdots \\ P(Y_i = t - 1) &= P(Y_i \leq t - 1) - P(Y_i \leq t - 2) \\ P(Y_i = t) &= 1 - P(Y_i = t - 1) \end{aligned}$$

where $\sum_{j=2}^t P(Y_i = j|x_i) = 1$. The log likelihood function of y for the i th individual is expressed as

$$L(\zeta|y) = \prod_{i=1}^m \left\{ \prod_{j=1}^t \left[P(Y_i = j|x_i) \right]^{\delta_{i,j}} \right\} \quad (3.48)$$

$$= \prod_{i=1}^m \left\{ \prod_{j=1}^t \left[P(Y_i = j|x_i) - P(Y_i = j - 1|x_i) \right]^{\delta_{i,j}} \right\} \quad (3.49)$$

$$= \prod_{i=1}^m \left\{ \prod_{j=1}^t \left[\frac{\exp(\alpha_j + x_i^T \beta)}{1 + \exp(\alpha_j + x_i^T \beta)} - \frac{\exp(\alpha_{j-1} + x_i^T \beta)}{1 + \exp(\alpha_{j-1} + x_i^T \beta)} \right]^{\delta_{i,j}} \right\}$$

$$\ell = \log L(\zeta|y)$$

$$= \sum_{i=1}^m \sum_{j=1}^t \delta_{i,j} \log \left[\frac{\exp(\alpha_j + x_i^T \beta)}{1 + \exp(\alpha_j + x_i^T \beta)} - \frac{\exp(\alpha_{j-1} + x_i^T \beta)}{1 + \exp(\alpha_{j-1} + x_i^T \beta)} \right] \quad (3.50)$$

where $\delta_{i,j}$ is an indicator variable

$$\delta_{i,j} = \begin{cases} 1 & \text{if } y_i = j \text{ where } i = 1, 2, 3, \dots, m, j = 1, 2, 3, \dots, t \\ 0 & \text{if otherwise} \end{cases}$$

$$(\zeta|y) = (\alpha_1, \dots, \alpha_{t-1}, \beta)^T.$$

3.6.1 Model and Parameter Estimation in Multinomial

Suppose that Y is a categorical response variable with three categories, represented as 1, 2, or 3. Since the outcome variable has three categories, we need two logit models, as the logistic regression model uses a binary outcome variable which parameterizes in terms of the logit $y = 1$ against $y = 0$. We assume that there are k explanatory variables, $x = (x_1, \dots, x_k)$ in our model. The logit models for nominal responses pair each response category to a baseline category and the choice is arbitrary. If we set the last category as the baseline, then the baseline category logits are represented as follows

$$\ln \left\{ \frac{p(y = 1|x)}{p(y = 3|x)} \right\} = \lambda_{10} + \lambda_{11}x_1 + \dots + \lambda_{1k}x_k = \lambda'_1 x$$

$$\ln \left\{ \frac{p(y = 2|x)}{p(y = 3|x)} \right\} = \lambda_{20} + \lambda_{21}x_1 + \dots + \lambda_{2k}x_k = \lambda'_2 x.$$

Considering the above model, the response probabilities are set up as follows

$$p(y = 1|x) = \frac{\exp(\lambda'_1 x)}{1 + \exp(\lambda'_1 x) + \exp(\lambda'_2 x)},$$

$$p(y = 2|x) = \frac{\exp(\lambda'_2 x)}{1 + \exp(\lambda'_1 x) + \exp(\lambda'_2 x)},$$

$$p(y = 3|x) = \frac{1}{1 + \exp(\lambda'_1 x) + \exp(\lambda'_2 x)},$$

with parameters $\lambda = (\lambda_1, \lambda_2)$ as the unknown. In the context of this thesis, the outcome variables are recoded as follows

$$Y_1 = 1, Y_2 = 0, Y_3 = 0 \quad : Y = 1,$$

$$Y_1 = 0, Y_2 = 1, Y_3 = 0 \quad : Y = 2,$$

$$Y_1 = 0, Y_2 = 0, Y_3 = 1 \quad : Y = 3.$$

Irrespective of the value Y takes, the sum of these outcome variables is $\sum_{i=1}^k y_i=1$. The conditional likelihood function, given the covariates for independent observations of sample n , is expressed as

$$L(\lambda) = \prod_{j=1}^n \left\{ \left(\frac{e^{\lambda'_1 x_j}}{1 + e^{\lambda'_1 x_j} + e^{\lambda'_2 x_j}} \right)^{y_{1j}} \left(\frac{e^{\lambda'_2 x_j}}{1 + e^{\lambda'_1 x_j} + e^{\lambda'_2 x_j}} \right)^{y_{2j}} \left(\frac{1}{1 + e^{\lambda'_1 x_j} + e^{\lambda'_2 x_j}} \right)^{y_{3j}} \right\}. \quad (3.51)$$

If we take the log on both sides, the expression reduces to

$$\ell(\lambda) = \sum_{j=1}^n \left\{ y_{1j} \lambda'_1 x_j + y_{2j} \lambda'_2 x_j - \ln(1 + e^{\lambda'_1 x_j} + e^{\lambda'_2 x_j}) \right\}$$

as $\sum_{i=1}^3 y_{ij}=1$ for each j .

The maximum likelihood estimators are obtained by taking the first partial derivatives of ℓ with respect to each of the unknown parameters and setting these equations equal to zero. In addition, the estimates of the parameters and variance covariance matrix can be obtained by any standard statistical computer package like SAS and R (nnet package).

3.7 Bayesian Generalized Linear Mixed Model (GLMMs)

In this section, we discuss the approach of generalized linear mixed model (GLMM) in the context of Bayesian analysis. The dataset used in this thesis is clustered, and to account for this heterogeneity we use GLMMs which are an extension of generalized linear models (GLMs). The main focus is on the Bayesian generalized linear mixed model and prior distribution for Bayesian Hierarchical models, as it is applicable to the Bayesian GLMMs. Effort was also given to the estimation of parameters and the application of the model to the breast cancer data. The method of estimation considered is the Markov chain simulation or MCMC, and we applied this method on breast cancer data as discussed extensively in the results section.

3.7.1 Generalized Linear Mixed Model

This thesis considers a multilevel logistic regression model formulated from a generalized linear mixed modeling (GLMM) framework. The generalized linear model (GLM) is a collection of fixed-effects models that assume the non-dependence of outcome response observations. Considering a dataset with hierarchical structures, this assumption is not valid. In addition, the assumption of the random effects in a linear mixed model must be Gaussian. GLMMs are extension of GLMs to correlated data where observation dependence within a cluster is captured through the use of random effect. The properties of both GLM and LMM are incorporated by GLMM. In GLMM, the mean response is modeled not as a function of covariates alone, but as being conditional of unobservable random effects. The inclusion of random effects in GLMM is aimed at inducing the correlation between the observations marginally when averaged over the distribution of the random effects.

Model Formulation

Unlike the linear mixed model, generalized linear mixed models assume that the outcome responses are conditionally independent. Suppose the random effects b_i , and the set of covariates

$$E(Y_{ij}|b_i, X_i, Z_i) = h(X_i\beta + Z_ib_i). \quad (3.52)$$

where $i = 1, 2, \dots, m; j = 1, 2, \dots, n_i$, and $h(\cdot)$ is an inverse link function. Random effects, b_i are drawn from $MVN(0, D_i)$. Y_{ij} is the j th outcome, X_i is an $n_i \times p$ matrix, and Z_i is an $n_i \times q$ matrix (associated with random effects b_i).

In terms of the link function, the expression (3.56) can be represented as

$$h^{-1}[E(Y_{ij}|b_i, X_i, Z_i)] = X_i\beta + Z_ib_i,$$

where $h^{-1}(\cdot)$ is a link function. The pdf's of Y_{ij} are represented as

$$f(y_{ij}|b_i, x_{ij}, z_{ij}) = \exp\left\{\frac{(y_{ij}\theta_{ij}) - b(\theta_{ij})}{a(\phi)} + c(y_{ij}, \phi)\right\}. \quad (3.53)$$

where θ_{ij} is a canonical parameter of a linear predictor η_{ij} , ϕ is a dispersion parameter which may or may not be known but is a function of the linear predictor, η_{ij} written as

$$\eta_{ij} = x_{ij}^T\beta + z_{ij}^Tb_i$$

and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions. The conditional expectation and variance are stated respectively

$$\mu_{ij} = h(\eta_{ij}) = b(\theta_{ij}) = \frac{\partial b(\theta_{ij})}{\partial \theta_{ij}}$$

and

$$\text{Var}(Y_{ij}|\eta_{ij}) = \phi \frac{\partial^2 b(\theta_{ij})}{\partial \theta_{ij}^2}$$

We then obtain the estimation for random variance components by maximizing the marginal likelihood, and by integrating out the random effects, b_i 's. According to (Pinheiro & Bates, 1995), the contribution of the likelihood of the i th individual is represented as:

$$f_i(y_i|\beta, D_i, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, \phi) + f(b_i|D_i) db_i. \quad (3.54)$$

Hence, the likelihood for β , D and ϕ is expressed as

$$\begin{aligned} L(\beta, D, \phi) &= \prod_{i=1}^m f_i(y_i|\beta, D_i, \phi) \\ &= \prod_{i=1}^m \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, \phi) + f(b_i|D_i) db_i. \end{aligned} \quad (3.55)$$

The expression above involves the computation of m integrals over the q -dimensional random effects, b_i . This computation cannot be obtained directly. Estimation of parameters are obtained by approximations of the above expression. We found that in GLMMs, the handling of random effects indicates that they follow some distributions like gamma, etc. Meanwhile, when Bayesian approaches via MCMC are involved, such problems become unnecessary. Bayesian MCMC techniques will produce a huge number of random samples for fixed and random effects to the likelihood functions instead of integrating intractable random effects. This technique is robust and it produces a means of marginalizing random effects (Browne et al., 2006; Zhao et al., 2006) which is an accurate method for the GLMM.

3.8 Estimation Methods

According to (Raudenbush & Bryk, 2002), there is no single way to estimate the parameters in a multilevel model. The most widely-used techniques for the estimation include Maximum Likelihood Estimation (MLE), Restricted Maximum Likelihood (REML), and Bayesian estimation. Most of these techniques of estimation can be carried out through different algorithms. For example, Maximum Likelihood Estimation (MLE) estimation can be achieved using any of these mechanisms: Expectation Maximization (EM) algorithm, the Newton-Raphson algorithm, Fisher scoring algorithm and Iterative Generalized Least Squares (IGLS), while Bayesian estimation will be carried out using Gibbs sampling in WinBUGS.

3.8.1 Maximum Likelihood Estimation

Maximum likelihood estimation is the most widely-used estimation method for parametric models. The main idea is to choose the value of the parameters that maximizes the joint density of the observations, called likelihood. In the case of linear models, the calculations for deriving the MLE can be achieved analytically, but in the case of a GLM or GLMM, numerical methods are adopted.

In case of the GLMM, the log-likelihood of the parameters (β, γ) , given a set observations y , can be expressed in (3.60) as

$$\ell(\beta, \gamma|y) = -\frac{1}{2}\log|\Sigma(\gamma)| + \log \int \exp\left\{ \sum_i y_i \theta_i(\beta, z) - \sum_i b(\theta_i(\beta, z)) - \frac{1}{2}z^T \Sigma^{-1}(\gamma)z \right\}. \quad (3.56)$$

The integral in expression (3.60) does not have a closed-form solution, which makes it difficult to obtain the likelihood accurately. In order to overcome these difficulties, the MCMC algorithm is employed (Ngesa et al., 2014). The MCMC technique is one of the techniques employed to generate the estimates of unknown parameters θ and correct the values generated in order to have a better estimate of the desired posterior distribution, $p(\theta|\eta)$ (Ojo et al., 2017; Ntzoufras, 2011).

3.8.2 Bayesian Multilevel

Bayesian multilevel modeling is a statistical modelling approach that involves multiple levels (hierarchical form) and estimates the parameters of the posterior distribution using the knowledge of Bayesian technique. Bayesian multilevel modeling is used when information is available on several different levels of observational units. The sub-models combine to form the hierarchical model, and Bayes theorem is used

to integrate them with the observed data and account for all the uncertainty that is present. The result of this integration is the posterior distribution, also known as the updated probability estimate, as additional evidence on the prior distribution is acquired. Frequentist statistics users, the more popular foundation of statistics, have been known to contradict Bayesian statistics due to their treatment of the parameters as a random variable, and use of subjective information in establishing assumptions on these parameters (Gelman et al., 2004). However, Bayesianists argue that relevant information regarding decision making and updating beliefs cannot be ignored, and that hierarchical modeling has the potential to overrule classical methods in applications where respondents give multiple observational data. In addition, the Bayesian model has proven to be robust, in particular when the posterior distribution is less sensitive to the more flexible hierarchical priors. Bayesian analyses are especially well suited for the analysis of multilevel models because of their flexibility in specifying multilevel structures of parameters using priors, as well as their ability to handle small samples and model mis-specifications (an over-parametrization of the likelihood can be resolved with well-chosen priors).

Bayesian Multilevel Estimation Procedure

One fact about statistics is that it is all about unpredictability. In classical statistics, we express our unpredictability about how well an observed statistic estimates the unknown population parameter by examining its sampling distribution over an infinite number of possible samples. Based on the fact that it is only one sample that we have, the sampling distribution is assumed to be based on a mathematical sampling model. The literature puts it that the use of Bayesian estimation techniques for multilevel models for variance parameters when the fixed effects are estimated was introduced by (Congdon, 2010). The key feature of Bayesian multilevel estimation is that it provides an appealing choice for researchers working with sparse datasets. This approach can be carried out using MCMC algorithms described at the beginning of this chapter (i.e. Gibbs sampling algorithm). All MCMC algorithms are iterative, and at each iteration, they are designed to yield a sample from the joint posterior distribution of the parameters of the model. After running a specific number of iterations either in WinBUGS or R, we get a sample of values from the distribution of any parameter which can be used to derive any desired distribution characteristic like the mode, mean and standard deviation. The mathematical formulation for Bayesian multilevel models combines prior information about the fixed and random effects with the likelihood based on the data under consideration. These parameters are considered as random variables described by the probability distributions, and

the prior information for a parameter is incorporated into the model via a prior distribution. Prior distributions must be specified, but these specifications of priors vary from one researchers to another, as many researchers do not want the prior to influence their results. The algorithm of Bayesian estimation in WinBUGS is very computer-intensive and frustrating sometimes.

3.8.3 Priors for Bayesian Multilevel Models

Comparisons between classical and Bayesian techniques of fitting variance component and random effect logistic regression models gave rise to Bayesian methods (Best et al., 1995; Best, 1996). The impact the prior distribution can have has been indicated by simulation outputs with respect to bias, particularly in level 2 variances. Gelman (Gelman et al., 2014a) investigates this more compactly and gave a recommendation on the use of a uniform prior for non-informative priors.

Conclusion

This chapter introduced the different methodologies adopted in this research. It begins with the classical logistic models for both classical and Bayesian statistics. When classical regression techniques are applied to multilevel data sets, the classical regression models may fail to account for the dependency structure in the datasets. Therefore, multilevel models are applied to ensure valid inferences. Thus, we proceeded to the multilevel logistic regression model in order to account for the hierarchical structure in our datasets. We also observed that the multilevel regression model is more complicated than the standard single-level multiple regression model. This study employed two types of the multilevel models, namely, the classical and the Bayesian multilevel models, and their results were compared. We introduced the descriptions of both models and described how parameter estimation was carried out for each of them. One of the interesting things about multilevel modeling techniques is that they can deal with datasets in which the times of measurement differ from subject to subject. The greatest advantage of Bayesian over classical statistics is that Bayesian statistics help in selecting more significant predictors in a model compared to classical statistics. They also make use of MCMC methods of estimation to achieve a desired posterior distribution. But sometimes, it takes a long time to draw a final decision from Bayesian models, which means that models should be run with different priors on the model parameters. Moreover, to ensure that the choice of the prior distribution does not affect the final results, there is need to assess the convergence of the result to know whether it has reached a convergence level or not. The strength of a multilevel model is that it can account for nesting in any dataset, and

has a very good ability to separately estimate the predictive effects of an individual predictor as well as its group level mean.

Chapter 4

RESULTS

Exploratory data analysis

The main purpose of exploratory data analysis (EDA) is to help us understand the data in detail before proceeding to modeling and inference tasks. R Software was used to explore the data in order to have an insight into the nature of the variables used because of its flexibility, clear and neatness in multiple line and surface graphs in this thesis. In particular, the variables and their respective significance to HIV research are explicitly detailed. In this section, a detailed and extensive exploratory data analysis is presented. The EDA focuses on the following:

- Determine association between the response variable and predictor variables,
- Providing a basis for further data collection through hospital record,
- Assessing assumptions on which inference will be based.

4.0.1 Descriptive Analysis Result of Breast Cancer

The descriptive statistics served to present the cancer data in general and more specifically information related to breast cancer. This section comprised of univariate and bivariate analysis. At univariate level of analysis, the data is presented with focus on the measurement of central tendency (mean, median) and dispersion (standard deviation, inter-quartile range). At bivariate level, variables are taken two by two to measure the association between each independent variable and the dependent variable. We also made use of statistics hypothesis tests to check the association or difference of means of each of selected prognostic factor. The responses and their respective percentage frequencies to the measured variables are presented in this section. From the descriptive statistics, we found that 192 cases accounting for (81.01%) were malignant breast lesions, while 45 cases (18.99%) were benign giving

a ratio of 4.3:1 for malignant to benign breast lesion. The current study focuses on the investigating the effects of demographic, socio-economic and medical factors on breast cancer. We explore each of the factors that are thought to be risk factors of breast cancer. The information obtained is useful for understanding how these factors determine attitudes of women of this part of Nigeria towards general health services and behaviors that may reduce the spread of breast cancer.

Summary statistics shows that 50.21% (119) of patients were presented with grade 2

Table 4.1: Distribution of grades of breast cancer at presentation

Grade	Cases	%
I	75	31.65
II	119	50.21
III	43	18.14
Total	237	

disease, followed by grade 1 disease (31.65%), with grade 3 disease being least commonly seen (18.14%) in this part of Nigeria. Further descriptive statistics reveals that the grade of disease was not seen to be significantly different amongst the different age groups, tribe and breast cancer types (see Table 4.1).

Table 4.2: Distribution of histologic types of breast carcinoma

Grade	Cases	%
Infiltrating duct carcinoma	99	41.77
Lobular carcinoma	57	24.05
Mastocytosis	516	6.75
Others (mixed histologic type)	65	27.43
Total	237	

Infiltrating duct carcinoma was the commonest histologic type of breast cancer seen (41.77%). This is distantly followed by others (mixed histologic types) (27.43%), lobular carcinoma (24.05%) while mastocytosis is the rarest type seen in the study population (6.75%). (see Table 4.8).

Table 4.3 gives the number of respondents in each of the age groups categories. The results in the table indicate that the proportion of respondents in each age group decreases with increasing age. Approximately 41% of the respondents were in the 20-34

Table 4.3: Summary of the categorized age groups for the patients

Age group	N	Percentage (%)
20-34	97	41
35-49	65	27
50-69	58	25
70+	17	7

years age group and approximately 27% were from the 35-49 years age group. Relatively smaller proportions of respondents were drawn from the older age groups in order to mirror the age structure of the breast cancer women in western Nigeria.

Table 4.4: Summary of the marital status for the patients

Marital Status	N	Percentage (%)
married	221	93
single	14	6
separated	2	1

Table 4.4 gives the results of the marital status of the patients. The data show that almost all the women in population (93%) were married and those in categories single/separated constituted 6% and 1% of the respondents.

Table 4.5: Summary of the employment status of the patients

Employed	N	Percentage (%)
civil servant	46	19.4
retired	23	9.7
self employed	168	70.9

Approximately 70.9% of the respondents were self-employed whereas approximately

19.4% claimed to be civil servant with about 9.7% of the patients have retired from the active work as shown in Table 4.5.

Table 4.6: Frequency distribution of breast cancer grade by biological risk factors (n= 236)

Variables	Grade I	Grade II	Grade III	Total
	n(%)	n(%)	n(%)	n(%)
Age Groups				
20-34	29(30.21%)	47(48.96%)	20(20.83%)	96(100)
35-49	18(27.69%)	35(53.85%)	12(18.46%)	65(100)
50-69	19(32.76%)	29(50%)	10(17.24%)	58(100)
70+	8(47.06%)	8(47.06%)	1(5.88%)	17(100)
Tribe				
Igbo/Efik	2(28.57%)	5(71.43%)	0(0%)	7(100)
Yoruba	72(31.44%)	114(49.78%)	43(18.78%)	229(100)
Breast cancer types				
Malignant	59(30.73%)	98(51.04%)	35(18.23%)	192(100)
Benign	15(34.10%)	21(47.73%)	8(18.18%)	44(100)
Treatment				
Surgery	50(29.76%)	86(51.20%)	32(19.05%)	168(100)
Others	24(35.29%)	33(48.53%)	11(16.18%)	68(100)

Table 4.6 shows the distribution of histologic grade by biological factors of breast cancer. The grade of disease was not seen to be significantly different amongst the different age groups, tribe, type of breast cancer and treatment modality. Table 4.7

Table 4.7: Contingency table for educational status and breast cancer type by age group

	Pry& Secondary		Tertiary		Malignant		Benign	
	n	%	n	%	n	%	n	%
20 - 34	81	71.7	16	12.9	68	35.4	29	64.4
35 - 49	32	28.3	33	26.6	56	29.2	9	20.0
50 - 69	-	-	10	58	52	27.1	6	13.3
70+	-	-	10	17	16	8.3	1	2.2

displays a educational status by age group contingency table. It is evident that the patients aged 20 - 34 years had at least high school education constitute almost 71.7% of the patients whereas a substantial proportion of those who had tertiary education is 12.9% within the same age group.

4.1 DETERMINANT OF BREAST CANCER TYPES

Abstract

The aim of this paper is to show how to handle Bayesian analysis of the binary regression model using Gibbs sampler in WinBUGS and to illustrate some aspects of model building and diagnostic checking via Markov Chain Monte Carlo (MCMC). Bayesian inference assumes that there are specific free (parametric) distributions for the unknown parameters. Bayesian fits probability of interest by incorporating prior information concerning the unknown parameters as well as the likelihood function of the observed data which helps in obtaining the posterior distribution. We developed a Bayesian binary regression model to established patients diagnosed with malignant breast cancer in western Nigeria. Markov Chain Monte Carlo (MCMC) methods are used to make inference and to evaluate Bayesian binary regression models. By using R software and WinBUGS 14, a model for Bayesian binary logistic regression model was fitted. The results of the Bayesian binary logistic regression is compared with the classical logistic regression. Bayesian model offers better estimates than the classical inference. The aspects of assessing convergence in Bayesian analysis are extensively discussed, including the posterior distribution of parameters. The methodological contribution of this paper is the strength to fit a Bayesian binary logistic regression model with the current dataset using Bayesian inference and classical statistics. It was established that breast cancer types is dependent on one's level of educational status and the age at diagnosis.

4.1.1 Introduction

Breast cancer (BC) is the commonest cause of mortality and morbidity among women worldwide, and currently the most common cancer among Nigerian women (Adebamowo & Ajayi, 1999; Ebughe et al., 2013a; Ojewusi & Arulogun, 2016; Oladimeji et al., 2015; Banjo, 2004). The human breast is a pair of mammary glands composed of specialized epithelium and stroma in which both benign and malignant lesions can occur (Dauda et al., 2011). Previous findings in Nigewria affirmed that benign breast constitutes the larger of the breast lesions but much concern is given to malignant lesions of the breast since breast cancer is the most frequent malignancy in the majority of the women (Uwaezuoke & Udoeye, 2014). Available statistics show that the annual incidence of breast cancer is increasing globally, and this is occurring more rapidly in countries with a hitherto low incidence rate of breast cancer (Wilson et al., 2004). Globally, breast cancer accounts for 18.4% of cancers associated with women. In 2012, (Jedy-Agba et al., 2012a) reported that the incidence of breast can-

cer in Nigeria has risen significantly with the incidence in 2009 - 2010 reported to be at 54.3 per 100 000, thereby representing a 100% increase within the last decade. The report about patients diagnosed with breast cancer in eastern Nigeria suggested that every 1 out of 5, representing 23%, are malignant in nature (Yusufu et al., 2003). Incidence rates of breast cancer vary from 19.3 per 100,000 women in Eastern Africa to 89.7 per 100,000 women in Western Europe, and are high (greater than 80 per 100,000) in developed regions of the world. From literature, we found that previous studies only focused on benign breast cancer (Abudu et al., 2007; Adesunkanmi & Agbakwuru, 2000; Forae et al., 2014; Guray & Sahin, 2006; Kumar et al., 2014; Anyikam et al., 2008; Godwins et al., 2011; Ochicha et al., 2002).

The application of the Bayesian technique and its usage to analyze cancer data has proliferated in recent years. Several researchers such as (Acquah, 2013) studied the comparison of Bayesian and classical logistic regression estimation and found that classical statistics has some limitation which requires an alternative approach to overcome. (Yu & Wang, 2011) presented a work on the estimation of a mixed logit model for simulated experimental data using Bayesian and classical logistic regression. They found in the classical estimation some unidentified parameters which could not be estimated by a classical technique, but which the Bayesian technique could correctly identify. A study on the prediction of panicle and shoot blight severity of Pistachio in California was carried out by (Mila & Michailides, 2006) using the comparison of both Bayesian and classical logistic regression techniques suggested that Bayesian gave a better result than the classical. They also found that Bayesian method was able to produce new information which the classical approach did not give. Other studies that have also shown similar result (Albert, 1996; Congdon, 2014; Marrelec et al., 2003; Díaz & Batanero, 2016; LOZANO-FERNÁNDEZ, 2008; Gordóvil-Merino et al., 2010)

In most case, studies comparing both methods find that Bayesian technique proffers a better solution compared to classical and has greater capability to eliminate the limitations encountered in classical regression. The Bayesian technique assumes model parameters as random variables and not as constants, while the probability of the unascertained parameters can be obtained via Bayes theorem (Bedrick et al., 1997; Congdon, 2005; O'Neill, 2002; O'Neill et al., 2000; Wong & Ismail, 2016). This provides information regarding parameter uncertainty that might be very difficult to obtain using the classical technique. Classical technique fits the logistic regression by means of an iterative approach like maximum likelihood method, fishers scoring, or iterative proportional fitting. In some cases, as a result of this iterative approach, convergence may be difficult to achieve. But the maximum likelihood es-

timation has a significant bias for sparse data, a limitation which can be addressed by the use of Bayesian approach. The robustness and accuracy of the results produced by Bayesian approach makes its gain popularity in data analysis. As a result, this paper investigate the significant predictors as well as characterizing patients diagnosed of benign and malignant breast cancer lesion using both classical approach and Bayesian approach (with non-informative priors). We also presents diagnosis of Markov Chain Monte Carlo (MCMC) Convergence.

4.2 Materials and Methods

4.2.1 Data Collection

Ethical approval was obtained from the ethics committee of the Federal Medical Teaching Hospital, Ekiti State, Nigeria. This data was extracted from cancer registry of the Federal Medical Teaching Hospital. We accessed 237 records and 20 variables of breast cancer data. Some of these variables describe socio-demographic and cancer-specific information on the incidence of breast cancer. Each record contained patient-related tumor information. Extensive variable selection procedures were performed on the 20 variables. The records of patients aged 20 years and above were sorted out for this analysis. Information collected includes age, marital status, educational level, religion, race, type of breast cancer, occupation, Lab number, case number, site of the female breast cancer, type of diagnosis and histological type. Other information recorded was the modality of treatment received: surgery, chemotherapy, hormonal therapy, radiotherapy or combination of these.

With respect to the quality of the data obtained from the record of the breast cancer cases, the main concern was the proportion of hospital records in which some of the relevant variables were not recorded. Information relating to variables like weight, height, age at first full term pregnancy and age at menopause were missing. For input variable selection, we tried to limit the number of variables and select only the clinically relevant ones. Logistic regression models were fitted to obtain independent estimates of the risk of breast cancer. Modeling started with all the variables followed by sequential deletion according to their statistical importance. Each variable was assessed through the Shapiro- test so as to see whether the data follow a normal distribution. The R software was used for the classical statistical analysis and the WinBUGS14 software for the Bayesian analysis. As a requirement of the Bayesian approach, several diagnostics tests were performed to answer convergence of the Markov chain Monte Carlo (MCMC) algorithm and the true reflection of the

posterior distribution.

Assessing Bayesian Markov Chain Monte Carlo (MCMC) and Convergence

We employed Gibbs sampling algorithm in order to obtain posterior distribution of parameters (Ntzoufras, 2011). In this study, non-informative prior were assumed in order not to influence the posterior distribution and it was assumed that $\lambda_k \sim N(1, 0.0001)$. All the Bayesian analysis in this study were carried out using WinBUGS 14 (Lunn et al., 2000). In our model, 1,500,000 Markov chain Monte Carlo (MCMC) iterations were ran, with the initial 200,000 discarded to cater for the burn-in period. The 5,000 iterations left were used for assessing convergence of the MCMC. Several tests, both graphical and formal tests can be used to check convergence are available via convergence diagnostics. These diagnostics are used mainly to check for stationarity of the chain and verify the accuracy of the posterior summary measures. Various convergence diagnostics are available but for the current study, we utilize the trace plots, the Geweke plots and the Heidelberger-Welch test as described by (Brooks & Roberts, 1998; Lesaffre & Lawson, 2012). The diagnostic of Heidelberger-Welch is used to test if the chain is stationary and convergence had been reached during the burn-in period. This diagnostic test is applied by discarding the first 10% of the chain if the test is failed while the remainder of the chain is re-tested. This process is repeated until either the test is passed or only 50% of the original chain remains. In this paper, we focus more on the techniques for determining whether a particular Markov chain has converged to stationary or not. The technique require only the output from one or more Markov Chain Monte Carlo (MCMC) simulations algorithm.

4.3 Results

4.3.1 Descriptive analysis of breast cancer

The main goal of this paper is to investigate the significant predictors as well as characterizing patients diagnosed of benign and malignant breast cancer lesions and presents diagnosis of MCMC convergence for the analysis, comparing the classical approach and Bayesian approach. Various prognostic factors are considered which include: intercept(λ_0), marital status: separated(λ_1), level of education: at least high school(λ_2), religion: christian(λ_3), tribe: yoruba(λ_4), age: 35-49(λ_5), 50-69(λ_6), 70+(λ_7), occupation: retired(λ_8), self employed(λ_9). A total of 237 breast cancer patients' data was extracted for analysis in the current study. Of these, 192 cases accounting for (81.01%) were malignant breast lesions, while 45 cases (18.99%) were

benign giving a ratio of 4.3:1 for malignant to benign breast lesion. The mean age of the respondents was 42.2 ± 16.6 years with 52% of the women aged between 35-49 years. Table 4.9 shows the Heidelberger and Welch stationarity tests for the Bayesian Markov chain Monte Carlo. This statistic is used to test if the chain is stationary and convergence had been reached during the burn-in period.

4.3.2 Result of classical logistic regression

Table 4.8 shows the result of a classical logistic analysis of the breast cancer tumors. The results indicate that the type of breast cancer (malignant) was observed to be strongly associated with age and educational status. This indicate that women with at least high school education have a significantly higher risk of being diagnosed with malignant breast tumors.

Table 4.8: The Result of Classical logistic regression for Patients diagnosed of Benign and Malignant

	Est	Std Error	z value	Pr(> z)
λ_0	-2.421	1.2308	-1.967	0.0492
λ_1	1.2479	0.5976	2.088	0.4459
λ_2	0.5926	0.7774	0.762	0.0368
λ_3	1.0782	0.6392	1.687	0.0916
λ_4	1.2048	0.8515	1.415	0.1571
λ_5	1.1952	0.5732	2.085	0.0371
λ_6	0.5034	0.8188	0.615	0.5387
λ_7	1.0534	1.4439	0.73	0.4656
λ_8	0.4823	1.0432	0.462	0.6439
λ_9	0.9898	0.5658	1.749	0.0802

4.3.3 Result of Bayesian logistic regression model

For the Bayesian technique, we present MCMC diagnostics for the patients diagnosed with benign and malignant breast cancer in Table 4.10. The posterior means in the Bayesian technique were obtained after a burn-in period of 5000 with Monte Carlo error less than 2%. The posterior means and medians of the coefficient λ_2 (educational status) and λ_5 (age) indicate that it was significant. The results of the posterior provide some evidence about the important variable to be selected while profiling patients diagnosed with malignant breast cancer (Table 4.10). For educational level, Table 4.10 shows that those with at least high school education are 1.3 times more likely than others to have benign breast cancer. The results indicate that women with age ≥ 35 years were at a higher risk of been diagnosed with malignant breast cancer than those with age < 35 years. Our result showed no indication that

Table 4.9: Heidelberger and Welch Stationarity and half-width tests for the Bayesian chains used in the diagnosis of MCMC

Param.	Station-arity Test	P-Value			Half-Width Test	Mean			Half width		
		C1	C2	C3		C1	C2	C3	C1	C2	C3
λ_0	passed	0.927	0.888	0.308	passed	0.117	0.132	0.129	0.054	0.054	0.055
λ_1	passed	0.591	-0.219	0.364	passed	-1.148	-1.144	-1.148	0.01	0.009	0.01
λ_2	passed	0.0572	0.818	0.204	passed	1.348	1.343	1.35	0.01	0.011	0.011
λ_3	passed	0.394	0.51	0.994	passed	0.836	0.838	0.839	0.013	0.012	0.012
λ_4	passed	0.915	0.987	0.112	passed	1.22	1.216	1.22	0.016	0.016	0.016
λ_5	passed	0.893	0.815	0.313	passed	-0.216	-0.226	-0.228	0.044	0.044	0.044
λ_6	passed	0.808	0.914	0.237	passed	-0.977	-0.977	-0.983	0.038	0.037	0.038
λ_7	passed	0.824	0.94	0.507	passed	-1.536	-1.55	-1.551	0.044	0.044	0.044
λ_8	passed	0.64	0.954	0.163	passed	0.6168	0.613	0.612	0.02	0.042	0.019
λ_9	passed	0.40	0.896	0.092	passed	1.054	1.058	1.053	0.009	0.009	0.009

religion, race or marital status could be a major factor for women to be diagnosed with malignant breast cancer in the western part of Nigeria.

Table 4.10: WinBUGS Posterior Summaries for Breast Cancer Patients

	Mean	SD	MC error	2.5%	Median	97.50%	start	Sample
λ_0	0.126	1.973	0.01568	-3.514	0.049	4.251	5000	49749
λ_1	-1.147	0.669	0.003012	-2.453	-1.146	0.176	5000	49749
λ_2	1.347	0.637	0.002987	0.180	1.316	2.695	5000	49749
λ_3	0.838	0.815	0.003789	-0.637	0.796	2.561	5000	49749
λ_4	1.219	0.920	0.004805	-0.633	1.224	3.012	5000	49749
λ_5	-0.223	1.672	0.001283	3.910	-0.102	4.685	5000	49749
λ_6	-0.979	1.523	0.01116	-4.418	-0.836	1.603	5000	49749
λ_7	-1.545	1.686	0.01269	-5.250	-1.415	1.371	5000	49749
λ_8	0.614	1.133	0.005761	-1.454	0.554	3.002	5000	49749
λ_9	1.055	0.590	0.00283	-0.093	1.047	2.241	5000	49749

All the convergence diagnostics tests are obtained using CODA/BOA package in R software. Diagnostics are computed on 1500000 iterations, thinned by sixty iterations, after a 5000 iteration burn-in. The Geweke (Z score) diagnostic suggests stationarity for all the parameter as all the Z-scores in Table4.11 are less than 1.96. Heidelberger-Welch and Raftery-Lewis also suggest acceptance by the MCMC algorithm. For the Heidelberger-Welch diagnostic, the p-value is from the Cramer-Von-Mises test of stationarity. Raftery-Lewis N is the estimated number of MCMC iterates required in order to obtain an accurate estimate of the quantile of marginal posterior density and I is the dependence factor. The dependence factor above five (5) indicate the presence of high autocorrelation that may occur as a result of poor choice of starting values for the chains of high posterior correlations (that is an indicative of convergence failure).

Table 4.11: Convergence diagnostics MCMC algorithm for two way ANOVA model

Parameter	Geweke Z	Heidelberger-Welch P	Raftery-Lewis	
λ_0	-0.64	0.92	4567	1.22
λ_1	0.29	0.59	3857	1.03
λ_2	0.18	0.57	3690	0.99
λ_3	1.76	0.39	3708	0.99
λ_4	0.39	0.92	3973	1.06
λ_5	0.78	0.89	7016	1.87
λ_6	0.92	0.81	9406	2.51
λ_7	0.80	0.82	10244	2.73
λ_8	0.28	0.64	3915	1.05
λ_9	-0.72	0.40	3763	1.00

4.3.4 Assessing the performance of Markov Chain Monte Carlo (MCMC) chains in WinBUGS

When the results of the model are computed, it is necessary to check for the stationarity of Markov chain Monte Carlo algorithm. The performance of a diagnostic test can be examined in several ways. The diagnostics examined in this paper are those of Geweke (Geweke et al., 1991), Heidelberger and Welch (Heidelberger & Welch, 1983), and Raftery and Lewis (Raftery & Lewis, 1992), which look at convergence of an individual chain, and that of Gelman and Rubin (Gelman & Rubin, 1992), which bases convergence on analysis of multiple chains. These results are essential to achieve an adequate estimation of the parameters for the convergence of MCMC algorithm. After an initial burn-in has been discarded, it is advisable to use only a subset of the parameters of interest. This process is called thinning and the purpose is to improve the mixing of the chain. The best way to achieve this is by generating the autocorrelation for each parameters being sampled. Fig 4.2 was presented to demonstrate this process and it shows that there is no problem of autocorrelation among the MCMC chain. Both Fig 4.1 and Fig 4.2 revealed that the chain for each parameter is stationary and not correlated. A more formal technique for diagnosing the convergence of MCMC chains is reported in Fig 4.4. The blue and red lines denote the variance within and between chains. To support that the chain is converged, the ratio must converge to unity and the blue and red lines must converge to a stable value. It also displays the red lines representing the \hat{R} . Hence, Figure 4.4 indicates that all the $\hat{R} \rightarrow 1$ which suggests that the algorithm converges. Both Figure 4.3 and Fig 4.4 explain the same thing but one is achieved through CODA/BOA while the other one is through WinBUGS.

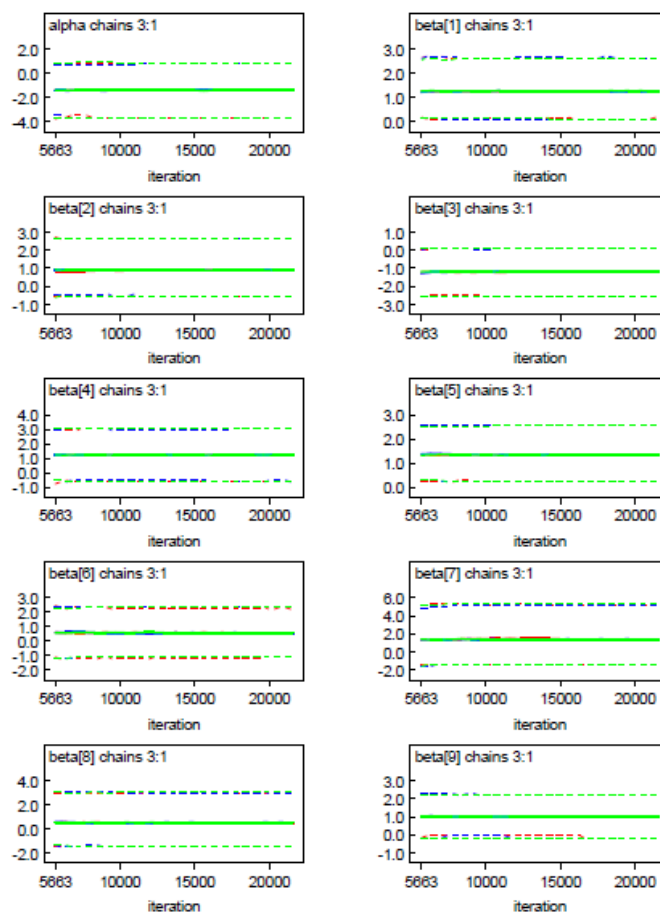


Figure 4.1. Running Quantiles for the Posterior Parameters in the case of Female Benign and Malignant Breast Cancer Patients.

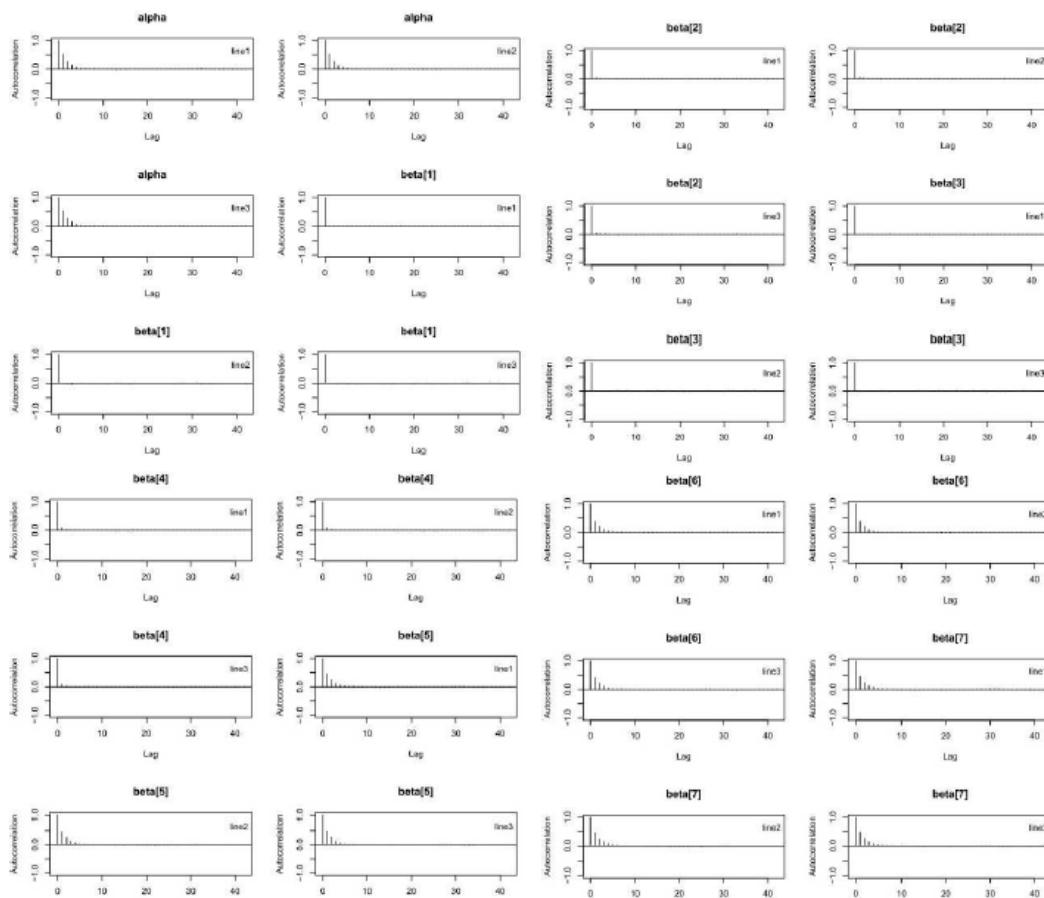


Figure 4.2. Auto-correlation plots for the Female Benign and Malignant Breast Cancer Patients.

Brooks & Gelman Multivariate Shrink Factors

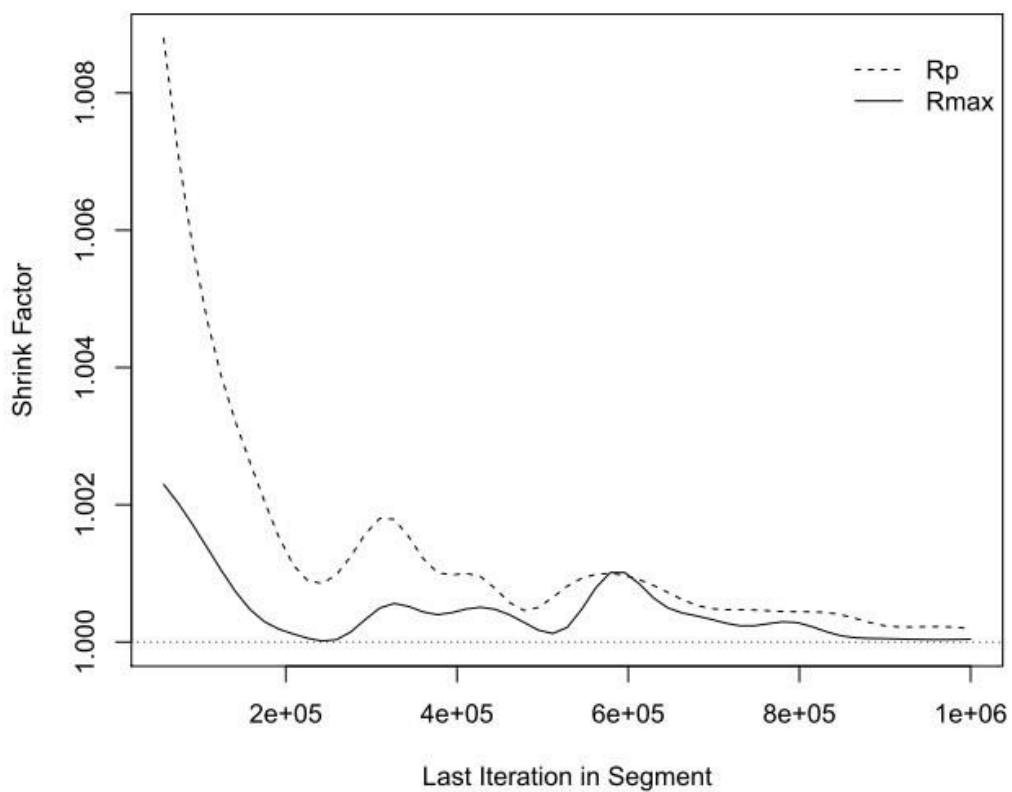


Figure 4.3. The plot of the Brooks-Gelman MPSRF for three chains of 49,749 iterations.

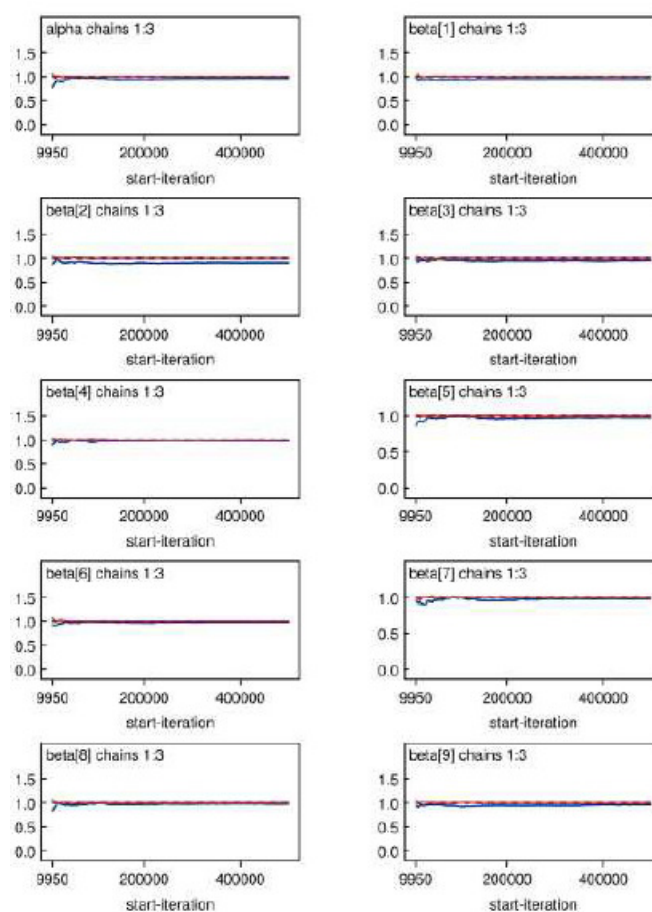


Figure 4.4. Gelman Rubin convergence diagnosis for independent variables

Result of Geweke plots for some selected parameters is given below. As a rule of thumb, a significant proportion of Z-scores outside the two-standard deviation bands is indicative of a chain that has not converged by iteration k . The results in Figure 4.5 show that all the Z-scores fall within the two-standard deviation bands for the parameters gender, and age groups 20-24, 25-29 and 30-44 years whereas there is a negligible proportion for the Z-scores under the intercept and age group 35-39 years that are outside the bands. This is a strong indication of a chain that has converged by iteration 1,500,000. We also considered the Heidelberger-Welch diagnostic test. The results of a test with $\epsilon = 0.1$ show that most of the parameters have passed the stationarity test except for male, age group 50-54 years, the married level for marital status, the literate level for literacy and the gender by age group interaction terms 35-39 :M and 40-44 :M. All the parameters passed the half-width test indicating that the chain was run sufficiently.

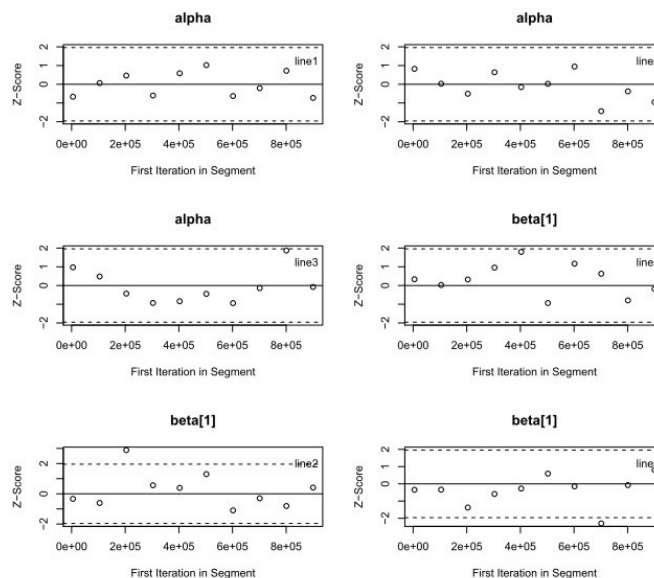


Figure 4.5. Gelman Rubin convergence diagnosis for independent variables

4.3.5 Discussion and partial conclusion

The present confirm findings from studies conducted in Nigeria over the past years, on Breast cancer among women in the western Nigeria (OLUGBENGA et al., 2012; Abudu et al., 2007). All these studies showed that age could be a risk factor for malignant breast lesion. Similar studies have been documented in other parts of Africa and the rest of the world (Arora & Simmons, 2009). From the results of analysis, patients age 35-49 years constituted the majority of patients (52%) in our study, indicating that women age 35-49 have a higher risk of developing breast cancer than their other counterpart in the group. Therefore, more attention on breast cancer treatment are necessary for this age group. This agrees with breast cancer facts and figures released between 2011-2012. However, this corresponds to the working class population and it is also the child bearing age for many women. This may be as a result of the use of contraceptive and hormonal imbalance which common among the women (Onyeanusi, 2015; OLUGBENGA et al., 2012). From the study, malignant breast lesions appeared to have higher distribution among those who had at least high school education, an observation which supports previous studies (Yüksel et al., 2017; Ibrahim et al., 2015), Zaria (Yusufu et al., 2003; Ntekim et al., 2009). This result was supported by the descriptive statistics which shows that 52.3% of those diagnosed had at least high school education meaning that that those who are educated were more interested in presenting their health problems to physicians (keeping in mind that this is an hospital data) rather than consulting quack medical doctors. The high proportion of malignant breast lesion might also be attributed

to lifestyle changes among those class. In addition, this may also be as a result of their exposure to advancement in life like the nature of occupation, diet, without observing caution to health management. We found that the mean age of breast cancer patients in Western is 42years; this is similar to several Nigerian institution based studies, Adebamowo reported 43 years (Adebamowo & Adekunle, 1999), Ikpat et.al 42.7 years (Ikpat et al., 2002) and 44.9 years by Ebughe.et al (Ebughe et al., 2013a). Although our variables' interactions did not categorize age and educational status as components of socioeconomic status (SES), our findings are similar to those of some studies which showed that higher socioeconomic status (SES) is associated with higher breast cancer incidence (Pudrovska & Anikputa, 2012; Krieger et al., 2010; Vainshtein, 2008). Additional studies have provided a possible explanation for these findings, that women with high SES are more likely to obtain routine breast cancer screening due to better access to preventive healthcare based on their level of education, higher income and increasing age, hence, increasing the detection of breast cancer (Akinyemiju et al., 2015).

Although our study did not investigate the risk factors for breast cancer in association with its sub molecular types, a recent study conducted by (Akinyemiju et al., 2015) evaluated the association between SES and breast cancer subtypes using a valid measure of SES and the Surveillance, Epidemiology and End Results (SEER) database. Socioeconomic status based on measures of income, poverty, unemployment, occupational class, education and house value, were categorized into quintiles and explored. Their findings showed that a positive association between SES and breast cancer incidence is primarily driven by hormone receptor positive lesion. Malignant breast lesions which can be subdivided into non-invasive and invasive tumors are documented to be more commonly diagnosed in postmenopausal women (Lehmann-Che et al., 2013). A molecular classification of breast cancer, with more than five reproducible subtypes (basal-like, ERBB2, normal-like, luminal A, and luminal B) has been defined through gene expression profiling and microarray analysis (Lønning et al., 2007). In addition, performing the gene set enrichment analysis (GSEA), a gene set linked to the growth factor (GF) signaling was observed to be significantly enriched in the luminal B tumors (Loi et al., 2009). Another study states that multiple pathways were identified by mapping gene sets defined in Gene Ontology Biological Process (GOBP) for estrogen receptor positive (ER+) or estrogen receptor negative (ER-); and among them, in a separate set, pathways related to apoptosis and cell division or G-protein coupled receptor signal transduction were associated with the metastatic capability of ER+ or ER- tumours, respectively (Jack et al., 2007)

Fig 4.2 measures the dependency among the Markov chain samples. The plot of Gelman Rubin convergence in Fig 4.4 suggesting that the MCMC sequence has converged on the posterior density as red line fall towards one for all parameters. Our findings are similar to the result obtained by Salameh.et.al (Salameh et al., 2014), Jackman.et.al (Jackman, 2000). Fig 4.3 shows a plot of the Brooks-Gelman MPSRF (denoted R_p) along with the maximum PSRF (denoted R_{max}) for successively larger segments of the chains. This plot suggests that although the chains differs significantly for the first few thousand iterations, they mix together after that and three chains of 1,500,000 iterations each is probably sufficient to ensure convergence of the chains (see (Sinharay, 2003) for a better understanding). It also suggests using a burn-in of about 200,000 each.

Findings from Bayesian and classical inference are not significantly different which could be due to the covariates or non-informative prior utilized in the model. Despite the similarities in their results, it is still difficult to compare the two approaches because classical inference make use of confidence interval to make decision while Bayesian uses credible interval. Moreover, when both techniques produce similar results, findings from Bayesian are given more attention because it is more robust compared to the classical. It is also possible to assess convergence of models under the Bayesian which could also make its result better than the classical inference. The second focus of this paper illustrate the importance of assessing convergence of Markov Chain Monte Carlo methods as well as presenting methods to show whether convergence has been reached. The model used in this paper updates quickly and adding complexity will also improve the required time for updating. These diagnostics are necessary to ensure that we are actually sampling from a chain that has converged after a desirable burn-in. Using the posterior mean as a point estimate, comparison of the classical statistics estimates with the simulation (MCMC) result. The estimated means and standard errors appear quite close with minimum results show a reduction of standard errors associated with the coefficients obtained from the Bayesian approach, hence resulting in higher stability to the coefficients. Other studies have also shown similar result (Gordóvil-Merino et al., 2010; Acquah, 2013). We have demonstrated a way to analyze small sample binary data with covariates, and have applied two approaches on breast cancer data. The classical statistics logistic regression model has some shortcomings which can be addressed with possible alternative approach. The purpose of this study was therefore to introduce Bayesian logistic regression as an alternative approach and demonstrate its application to parameter estimation in the setting of generalized linear model for comparative analysis with the classical statistics. Findings of this study shows that the Bayesian Markov Chain Monte Carlo (MCMC) algorithm proffers an alternative

framework to breast cancer data. Both the classical approach and Bayesian approach suggest that age of the patients and those with at least high school education are at higher risk of being diagnosed with malignant breast lesion than benign breast lesion in Western Nigeria. A comparison of the classical and Bayesian approach to modeling breast cancer reveals lower standard errors of the estimated coefficients in the Bayesian approach in the setting of generalized linear mixed model. Thus, the Bayesian approach is more stable. Malignant breast lesion is increasing alarmingly even though most people lack basic knowledge about its spread. The higher proportion of those affected by malignant breast lesion is found among the educated and younger women. Therefore, this shows that non-educated women do not patronize these services based on our findings. More efforts are required towards creating awareness and advocacy campaigns on how the prevalence of malignant breast lesions can be reduced, especially among women.

We recommend that governments, non-governmental organizations and other sectors involved in policy making to put in place policies, strategies and sensitization that target non-educated women to enhance their patronization of breast cancer screening in the health facilities, so as to access the appropriate management health assessment as well as providing financially supported treatments for breast cancer patients.

4.4 DIAGNOSTIC DETERMINANT OF PREFERRED CANCER TREATMENT

SUMMARY

Abstract

Breast cancer is one of the most common cancers among women and is the main cause of death among Nigerian women. Breast cancer treatment strategies in Nigeria need urgent strengthening to reduce mortality rate as a result of the disease. The objective of this study is to determine the relationship between the age at diagnosis and established the prognostic factors of modality of treatment given to breast cancer patient in western Nigeria.

The data was collected for 247 women who had breast cancer in two different hospitals. Model estimation is based on fully Bayesian approach via Markov Chain Monte Carlo (MCMC). In this study, a multilevel model based on the generalized linear mixed model is used to estimate the random effect. Considering 0.05 as the level of significance, the Bayesian data analysis was done using the WinBUGS14 software while SAS was used for the classical statistics.

The mean age of the patients (at the time of diagnosis) was 42.2 years with 52% of the women aged between 35-49 years in this study. Bayesian analysis showed a significant relationship between treatment modality and the age ($P=0.04$), and malignant breast lesion ($P=0.001$). The result also indicate that hospital a breast cancer patient attends in this part of Nigeria also played a significant role for the treatment modality. By fitting Bayesian multilevel models the variables age and breast cancer stages (malignant lesion) were significant.

The results showed that age, breast cancer stages (malignant) and hospital had a significant role in the treatment modality of breast cancer patients in Western Nigeria. The study showed the practicality and flexibility of Bayesian multilevel approach in analyzing breast cancer data.

Keywords: Bayesian, Bayesian inference, Multilevel, Generalized linear mixed modeling, CODA/BOA.

4.4.1 Introduction

Breast cancer refers to a malignant tumor that has developed from cells in the breast. Breast cancer cells are normal cells that at some point begin to present structural or functional alterations in nuclear deoxyribonucleic acid (DNA) and gradually became abnormal due to uncontrolled cell division. This uncontrolled division leads to the cells growing at an abnormal, uncontrolled rate. Breast cancer is the most common cause of malignancy among women worldwide (Ebughe et al., 2013a; Adebamowo & Ajayi, 1999; Chen et al., 2016), and is a public health challenges among Nigeria women. Although cancer of the breast is thought to be a disease of the developed world, literature reveals that almost 58% of deaths that occur as a result of breast cancer are traced to less developed countries (Ferlay et al., 2010). Available record by World Health Organization indicated that 30 Nigerian women died of breast cancer every day in 2008 and this had risen to 40 women in 2012. Statistics show that over 508 000 of Nigerian women died in 2011 as a result of breast cancer (Torre et al., 2015; Akarolo-Anthony et al., 2010; AO et al., 2013; Adebamowo & Ajayi, 1999). Moody et.al (OLUGBENGA et al., 2012) presented a profile of cancer patients attending tertiary health institution in South Western parts of the country. They observed the occurrence and distribution of cancers among the southwestern citizens, it was concluded that breast cancer alone accounted for 37% of all the cancer cases present in southwestern of the country.

With advancement in medical technology, treatment for breast cancer advances have been made regardless of the site of the cancer in Nigeria. Better treatment for breast

cancer patients is difficult to define and the literature has revealed that older women are sometimes excluded from clinical treatment trials, probably because of their age (Townnsley et al., 2005). Since breast cancer biology differs from patient to patient with respect to factors like age, variation in response to treatment, and substantial competing risks of mortality (Biganzoli et al., 2012; Mieog et al., 2012; Kiderlen et al., 2015), the exclusion of some patients might not be valid. This implies that those elderly who are included in trials are probably not a true representation for the general older population (Jolly, 2015). Consequently, an evidence-based treatment strategy for women with breast cancer is needed. Literature review on epidemiological studies of risk factors for breast cancer have reported that breast cancer is related to family history of breast cancer, early menstruation, late onset of menopause, old age, age at first pregnancy over 30 years, use of contraceptives, hormonal treatment after menopause, no history of breastfeeding and obesity (Zare et al., 2013).

Although there have been substantial published studies on prognostic factors for breast cancer in western Nigeria, population-based research is sparse. Therefore, we sought to determine risk factors for treatment given to female breast cancer in western Nigeria using generalized linear mixed models. This research focuses on the justification for the use of the conventional surgery method in the treatment of cancer, based on data from two understudied hospitals in Southwestern Nigeria, one federally owned, and the other state-owned. Consequently, the questions in the minds of people in this region regarding the method of breast cancer treatment employed by federal hospitals, as against that used by state hospitals, will have been addressed. Therefore, it is important to explore the relationship between these variables and treatment modality. Fundamentally, this study presents a comparison of the classical and Bayesian multivariate generalized linear mixed models.

4.4.2 Materials and Methods

4.4.3 Ethics

Ethical clearance to conduct the study was sought from the Ethical Review Committee of the Federal Teaching Hospital, Ekiti State, Nigeria. The data was extracted from the cancer registry of the Federal Medical Teaching Hospital. 237 records and 20 variables of breast cancer data were accessed, each containing patient-related tumor information. Extensive variable selection procedures were performed on the 20 variables, and the records of patients aged 20 years and above were selected for the analysis. The information collected included age, marital status, educational level, religion, race, type of breast cancer, occupation, Lab number, case number, site of the cancer, type of diagnosis, and histological type. With respect to the quality

of the data obtained, the main concern was the proportion of hospital records in which some of the relevant variables were absent. Information relating to variables like weight, height, age at first full term pregnancy, and age at menopause were missing. Other information recorded was the type of treatment received: surgery, chemotherapy, hormonal therapy, radiotherapy or a combination of these. For input variable selection, we tried to limit the number of variables and select only the clinically relevant ones. Exact Logistic regression models were fitted to obtain independent estimates of the risk of breast cancer. Modeling started with all the variables, followed by sequential deletion according to their statistical importance. SAS was used for classical statistical analysis, while WinBUGS was used for Bayesian analysis. A GLMM with a logit link function was performed for both classical as well as Bayesian method since this study considered two levels of analyses (hospitals). We also examine several diagnostics that have a very wide range of application. The diagnostics are those of Geweke (Geweke et al., 1991), Heidelberger and Welch (Heidelberger & Welch, 1983), and Raftery and Lewis (Raftery & Lewis, 1992), which look at convergence of an individual chain, and that of Gelman and Rubin (Gelman & Rubin, 1992), which bases convergence on analysis of multiple chains.

4.5 Results

This section introduces the treatment modality for breast cancer patient in western Nigeria as well as the results obtained from an analysis for both classical and Bayesian techniques. We applied the two methods of hierarchical models: classical multilevel and Bayesian multilevel approach and their results were compared. For us to fit the classical multilevel logistic model, we used lme4 and glmer package in R software. For each parameters in our models, we compute the estimates and the 95% confidence intervals (CI) and the deviance for the model were obtained. We ran Bayesian multilevel logistic models by fitting generalized linear mixed model (BGLMM) using Markov Chain Monte Carlo (MCMC) techniques in WinBUGS software (Ntzoufras, 2011). For each model, 1,500,000 Markov chain Monte Carlo (MCMC) iterations were ran, with the initial 500,000 discarded to cater for the burn-in period and thereafter keeping every 100-th iteration of the remaining one million sample value. The 1,000,000 iterations left were used for assessing convergence of the MCMC and parameter estimation. We assessed MCMC convergence of all models parameters by checking the Gelman-Rubin, trace and autocorrelation plots of the MCMC output. Our findings are in line with the previous literature by Clèries et al. (2012); Salameh et al. (2014); Gelman et al. (2003).

4.5.1 Socio-demographic characteristics of participants

A total of 237 breast cancer female patients extracted from the cancer registry of Federal Medical Teaching Hospital, Ekiti State, Nigeria, were used. Of these, 192 cases accounting for (81.01%) were malignant breast lesions, while 45 cases (18.99%) were benign giving a ratio of 4.3:1 for malignant to benign breast lesion. The mean age of the respondents was 42.2 ± 16.6 years with 52% of the women aged between 35-49 years. It was observed that 93.67% of those who participated in this study were Christians while 6.33% were Muslims. The percentage of breast cancer was higher among Christians. Various prognostic factors are considered which include: intercept(λ_0), breast cancer types: malignant(λ_1), age: 35-49(λ_2), 50-69(λ_3), 70+ (λ_4), level of education: at least high school(λ_5), religion: christian(λ_6), tribe: yoruba(λ_7), occupation: site of the cancer(λ_8), hospital(λ_9).

4.5.2 Result of the classical multilevel logistic regression

Results of the multilevel logistic regression model presented in Table 4.12 identified the following predictors as the statistically significant determinants of treatment given to breast cancer patients in western Nigeria: age groups, histological type, type of breast cancer diagnosed. These variables presents a statistical significant influence on breast cancer treatment at 95% confidence interval (CI) while other variables did not show any statistical evidence of significant influence at 95% CI. In this analysis we observed that the risk of breast cancer varies significantly between the socio economic and medical factors respectively. In addition, to establish the findings of our results beyond the kind of analysis that has been performed on breast cancer in western Nigeria, there is need to use advanced statistical techniques. To achieve this, we first performed a model called multilevel logistic regression model denoted by model A. The main aim of this model is to compare the prognostic factors associated with treatment given to breast cancer patients.

Findings from model A reveal that age, histological type (infiltrating duct and others) may likely be the major factors considered before treatment is given to breast cancer patient in this part of Nigeria. Results from multilevel logistic regression model B indicate that the addition of type of breast cancer as well as marital status in the model did not make histological type to be associated with treatment modality. Although, both age category (35-50) years and 51+ are now significant. Hence, from model A and model B, we found that histological type (infiltrating duct and others) was significant when type of breast cancer was not included in the model. It means that the influence of the histological type passes through the type of breast cancer a patients was diagnosed. Results from model A shows that most of the patients

Table 4.12: Multilevel Logistic Regression: Model A

Variables	Estimate	Std.Error	Z value	Exp(B)	Pr(> z)
Age groups					
35-50	-0.7145	0.4488	-1.5920	0.4890	0.1114
51+	-0.9721	0.4608	-2.110	0.3780	0.0349*
Histological type					
infiltrating duct	-1.8505	0.5528	-3.348	0.157	0.0008***
lobular	-0.4617	0.6783	-0.681	0.630	0.4961
others	-1.5655	0.5017	-3.121	0.209	0.0018**
Unit					
Gynaecology	-1.6050	0.5521	-2.907	0.200	0.0036**
Surgical	-0.2721	0.5811	-0.468	0.762	0.6396
Unit(mixed)	-0.7211	0.5282	-1.365	0.486	0.1723
Site of the cancer					
others	-0.8707	0.4613	-1.888	0.419	0.0591

Deviance: 209.3, AIC: 231.3, BIC: 269.4

were based on surgical operation. Moreover, the regression model also indicated that breast cancer type (malignant) is significant with treatment modality. This is an indication that the treatment given was based on the diagnosis of the breast cancer type whether benign or malignant lesion. Based on the value of AIC, we can say that the result of model B fit better than that of model A.

In addition, results of regression model A indicates that women aged 51+ years are 62.2% less at risk of taking surgery treatment than those aged 35-50 years. In case of model B, it was found to be 69.5% less at risk. The Table 4.13 shows that women not placed in a surgical unit but diagnosed of breast cancer have 1.4 times more risk of not surviving than others in this part of Nigeria. Patients with histological type (lobular) are 84.3% less at risk of surgery treatment than patient with histological type (others). Findings of model B presented in Table 4.13 (infiltrating duct) show that women having histological type (lobular) are 3.2 times more at risk of breast cancer than those with histological type (infiltrating duct and others).

4.5.3 Result of Bayesian multilevel model with non-informative prior

We make use of non-informative prior for the Bayesian multilevel model. The result is presented in Table 4.14. In line with the theory, the results from the Bayesian analysis with non-informative priors are similar to those of the classical analysis. In fact, the theory reveals that non-informative priors should not have effect on the posterior. The most significant aspect of the Bayesian statistics is its credible interval (Cred. I) is quite different from the confidence interval (CI) for classical statistics. In addition, the credible interval in Bayesian statistics is more robust than the confi-

Table 4.13: Multilevel Logistic Regression Model B

Variables	B	Std.Error	Z value	Exp(B)	Pr(> z)
Age groups					
35-50	-1.3466	0.6221	-2.165	0.260	0.0304*
51+	-1.1871	0.5935	-2.000	0.305	0.0455*
Histological type					
infiltrating duct	-0.1158	0.6840	-0.169	0.891	0.8655
lobular	1.1759	0.8134	1.446	3.241	0.1483
others	-0.3752	0.6119	-0.613	0.687	0.5398
Type of breast cancer					
malignant	6.4056	1.2627	5.073		$3.9210e^{-7}$ * **
Unit					
Gynaecology	-0.5399	0.5958	-0.906	0.583	0.3648
Surgical	0.3429	0.6687	0.513	1.409	0.6081
Unit(mixed)	-1.6301	0.8792	-1.854	0.196	0.0637
Site of the cancer					
NOS/lower-inner quadrant	-0.4938	0.6766	-0.730	0.610	0.465
Marital status					
single/separated	-0.6899	0.9406	-0.734	0.502	0.4633

Deviance: 142.9, AIC: 168.9, BIC: 214.0

dence interval in classical statistics which is sometimes affected by the sample size. Considering the credible interval in model M1, results of non informative prior for Bayesian multilevel model indicate that variables are the only significant predictors associated with modality of treatment given to breast cancer patient while are not significant determinants of treatment modality in western Nigeria. Result of model M2 also shows that patient diagnosed of benign breast cancer are 17.4% (OR= 0.826) less likely to receive surgery treatment compared with patients diagnosed of malignant breast lesion. Findings also highlight that patients who their breast lesion are not malignant in nature are not likely to be subjected to surgical operation unit. Therefore, model M2 indicates that such patients are normally retained in general unit section among the two hospitals.

Furthermore, patients with histological type (MORF3) are 1.51 times more likely [$\beta = 1.51(0.452-2.601)$] at risk of taking surgery treatment than others while as age increases, the risk of taking surgery treatment also increases by 1.17 (see model M1). Most of the predictors significant in model M1 are no longer significant in model M2 when type of breast cancer was introduced into the model. This indicates that type of breast cancer is a major factor considered before a patient is being placed on any treatment in this part of Nigeria.

Table 4.14: Parameter estimates of models M1 and M2 obtained from WinBUGS

Parameter	Model M1			Model M2		
	Estimate	2.5%	97.5%	Estimate	2.5%	97.5%
Intercept	-1.516	-10.650	7.472	-0.258	-4.586	3.664
λ_1	0.554	-0.302	1.424	0.822	-0.327	2.019
λ_2	1.165	0.274	2.095	1.124	-0.058	2.376
λ_3	1.303	0.293	2.416	0.034	-1.095	1.228
λ_4	1.510	0.452	2.601	-0.191	-10.690	-4.691
λ_5	0.111	-1.027	1.259	0.511	-0.721	1.705
λ_6	0.550	-0.462	1.572	-0.246	-1.639	1.155
λ_7	0.663	-0.219	1.583	2.099	0.336	4.212
λ				0.614	-0.688	2.054
τ	0.606	0.002	3.062	2.775	0.008	15.50

Table 4.15: Demographic characteristics of women diagnosed with different histologic types of breast cancer

	1		2		3		4	
	n	%	n	%	n	%	n	%
Age groups								
20 - 34	58	59.6	10	19.2	14	24.2	4	23.5
35 - 49	23	23.2	22	42.3	18	31.0	7	41.2
50 - 69	14	14.1	7	13.5	5	8.6	-	-
70+	4	4.0	13	25.0	21	36.2	6	35.3
Education								
At least high school	68	68.7	21	36.8	3	18.8	21	32.3
Pry& Sec	31	31.3	36	63.2	13	81.3	44	67.7

Assessing the performance of Markov Chain Monte Carlo (MCMC) chains in WinBUGS

To further establish the Bayesian results, it is necessary to check for the convergence of Markov Chain Monte Carlo (MCMC) once the results of the models are computed. This step is necessary so as to ensure that we are sampling from a better posterior distribution. Fig 4.6 illustrates the convergence of the Bayesian with non-informative prior. The algorithm converged after 1,500,000 iterations. To avoid autocorrelation, a lag of 60 was chosen which required up to 1,500,000 iterations and the first 500,000 iterations were discarded keeping every 10-th iterations for the remaining one million iterations. In order to assess whether a chain has converged or not, we plot the sampled value against its number in the chain. When the time series centered around a constant mean, it implies that the chain has reached convergence as reported in Fig 4.6 and have the chains converged to the same solution. Fig 4.7 displays the representation of the parameter behavior after 1,500,000 Monte Carlo repetitions. It was found that the kernel densities for shape and scale parameters exhibit approximately symmetric distribution. A more formal technique for diagnosing the convergence of MCMC chains is reported in Fig 4.6. The blue and red lines denote the variance within and between chains. To support that the chain is converged, the ratio must converge to unity and the blue and red lines must converge to a stable value. Hence, Fig 4.6 indicates that all the $\hat{R} \rightarrow 1$ which suggests that the algorithm converges.

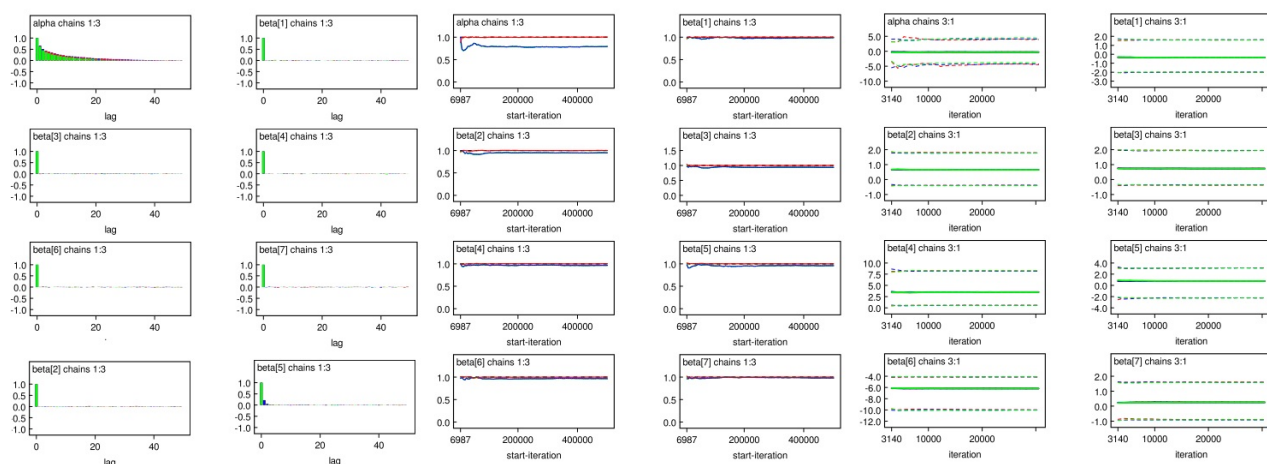


Figure 4.6. WinBUGS' output of Gelman Rubin Statistic for some independent variable

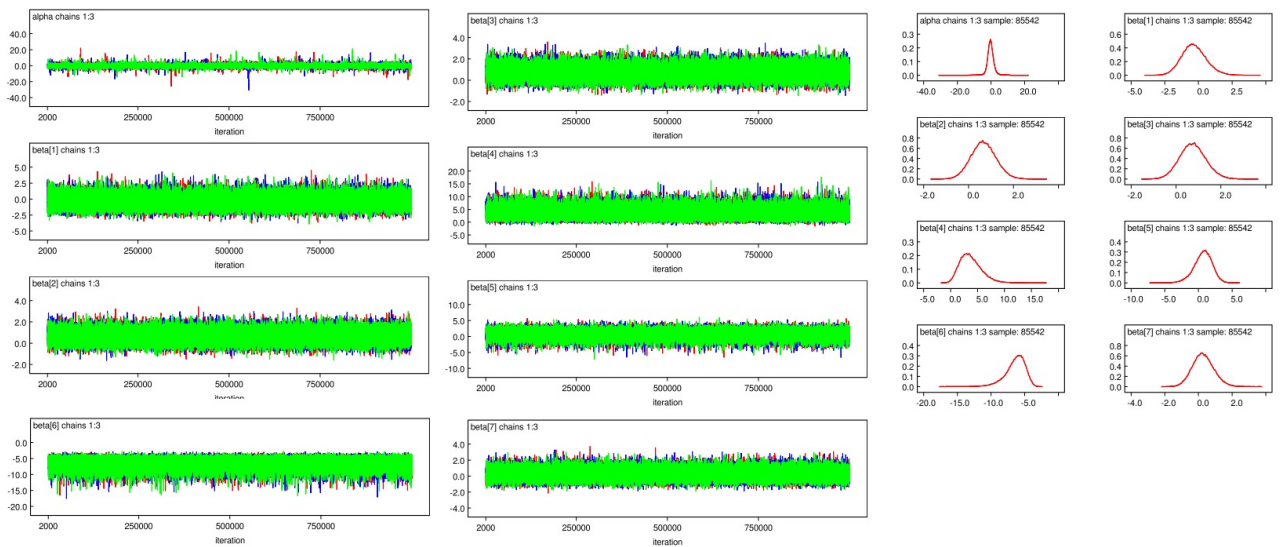


Figure 4.7. Auto-correlation plots for the preferred treatment given to patients.

Model Selection Procedure

Selection of best models from the same dataset can be achieved using summary measures of fit. The deviance statistic is available for both classical and Bayesian modeling, and is equal to minus the log-likelihood written as

$$deviance = -2 \log(L)$$

where L is the likelihood and is defined as the probability of the data given the models parameters to be estimated. That is,

$$L = p(y|\theta)$$

In case of classical statistics, the addition of parameters to the model is assumed to improve the fit. But if addition of one parameter to a model does not have impact on the model, the expected deviance is reduced by one. In addition, assume g predictors are added and the deviance is reduced significantly more than g , then the observed improvement is statistically significant. Moreover, adjusted deviance = deviance + number of predictors, where the difference between adjusted deviance and deviance is tested against the χ^2 distribution with degrees of freedom set as the number of extra predictors. In the case of non-nested models the Akaike information criterion (AIC) can be used as alternative. This is expressed as $AIC = deviance + 2(\text{number of predictors})$.

The model with the smallest AIC indicates better fit. Other measures of model se-

lection is the Bayesian information criteria (BIC) which is represented as

$$BIC = -2 \log(L) + g \log(n).$$

According to (Spiegelhalter et al., 2002), g is the number of estimated parameters and n is the sample size.

In the concept of Bayesian analysis, (Spiegelhalter et al., 2002) suggested the mean posterior deviance $\bar{D} = E(D)$ as a measure of model fit. Where deviance written as $D(\theta)$ is $D(\theta) = -2 \log p(y|\theta)$ for a likelihood expressed by $p(y|\theta)$. The mean posterior deviance does not account for the improvement of fit with increasingly complex models. Bayesian model comparison criteria called deviance information criteria (DIC) merges goodness of fit and model complexity to estimate the effective number of parameters denoted by pD . Deviance information criteria (DIC) is represented by

$$DIC = D(\bar{\theta}) + 2pD = \bar{D} + pD. \quad (4.1)$$

Models with smaller DIC fit better and are preferred

$$pD = \bar{D} - D(\bar{\theta})$$

where pD is the number of effective parameters, $\bar{\theta}$ is the posterior mean and $D(\bar{\theta})$ is the deviance for posterior mean. For all models, we initiated three chains with

Table 4.16: WinBUGS output for the evaluation of logistic regression multilevel using pD and DIC

model	Dbar	Dhat	pD	DIC
M1	219.503	210.411	9.092	228.596
M2	152.710	142.208	10.501	163.211

1500000 iterations for each chain. DIC and pD were calculated on the last 500,000 iterations. From above table, we can observe that for model M1 the effective number of parameters is approximately 9 showing the restrictions among the random effects. Comparing model M1 with model M2 shows that there is not much evidence for correlated random effects. This is also supported as can be seen from 95% credible interval (Cred. I) for $\tau = [0.0019, 3.062]$ (see Table ??).

4.5.4 Discussion and partial conclusion

For breast cancer treatment in Nigeria, the reason for surgery treatment on cancer patients is not clear as reported by Odetunmibi et al. (2013). The present study showed prognosis factors for breast cancer treatment in southwestern Nigeria. The cancer

registry of Federal medical teaching hospital is a hospital-based registry, thus the data are not population-based and may not be representative of all patients with breast cancer treated in western Nigeria. One of the first thing we noticed was that the results obtained from both techniques are identical and this result is supported by previous studies that compared both techniques (Ojo et al., 2017). The key significant factors in this section is age group, histologic type, breast cancer type and unit they placed their patients depending on the type of breast cancer being diagnosed. From the previous studies, age has been identified as an important risk factor for breast cancer (Chen et al., 2016; Alieldin et al., 2014). The study found that the type of breast cancer a patients is diagnosed also determine the type of treatment such patients were given. This was expected as patients breast cancer may not have gotten to malignant nature which may not require serious treatment like surgery. Patients with at least high school education were found to be positively associated with treatment modality. The most important biological factor associated with treatment modality is histologic type: patients with this type of histologic were 1.2 times more likely to be given surgery treatment than their counterpart. Patients age 35-49 years constituted the majority of patients (52%) in our study, indicating that women age 35-49 have a higher risk of developing breast cancer than their other counterpart in the group. Therefore, more attention on breast cancer treatment are necessary for this age group. Considering the effects of age in the treatment modality of breast cancer patient, Bayesian findings highlight that age bracket 35-49 years are 0.67 times more likely to receive surgery treatment than other age categories. This suggests that younger patients were more likely to be treated with surgery, compared to the older patients, findings consistent with those of many studies (Chen et al., 2016; Cluze et al., 2009; Ibrahim et al., 2014). Some studies reported that most of the available data on breast cancer indicate that young age is associated with a poor prognosis as a result of more invasive disease (Ibrahim et al., 2014; Chen et al., 2016; Alieldin et al., 2014; Brandt et al., 2015) and our results indicated that malignant breast lesion was more common in this region of Nigeria more than benign lesion. We can attribute this reason to say that this may be the more reason why Southwestern Nigeria considered age to be a determinant before subjecting their breast cancer patients to any form of treatment.

From the descriptive statistics, we found that the mean age of breast cancer patients in southwest is 42 years; this is similar to those of several Nigerian institutions which have been studies in the literature. Among these studies, Adebamowo reported 43 years (Adebamowo & Adekunle, 1999), Ikpat et.al 42.7 years (Ikpat et al., 2002) and 44.9 years by Ebughe.et al (Ebughe et al., 2013a). In model 1, age 35-49 and hospital appears to be a significant determinants for breast cancer treatment in Western

Nigeria. Similarly, results from model 2 and model 3 are identical in the sense that only hospital and malignant breast lesion were associated with the modality of treatment given to patients. Also, results of model 2 indicate significance for hospital but its influence passes through model 3 and hospital loses its significance as soon as type of breast cancer diagnosed (malignant) is introduced into the model. This could mean that hospital a patient attended in this part of Nigeria will determine the type of treatment offered for patients. It could also mean that the treatment offered depends on whether such patient have malignant breast lesion or not before placing their patients on a particular form of treatment.

Although our study did not investigate the risk factors for breast cancer, a recent study conducted by (Chen et al., 2016) evaluated the effect of age in breast cancer prognosis using Surveillance, Epidemiology and End Results (SEER) database. Their findings showed that younger breast cancer patients exhibit more aggressive disease than the older patients. Additional studies have provided a possible explanation for these findings, that women with high socioeconomic status (SES) are more likely to obtain routine breast cancer screening due to better access to preventive healthcare based on increasing age, hence, increasing the detection of breast cancer (Akinyemiju et al., 2015). Past studies also reported that elderly women sometimes experience poorer outcomes than younger patients (Yancik et al., 2001; Schonberg et al., 2010). In Bayesian, once the results of the model are computed, it is necessary to check for the convergence of Markov Chain Monte Carlo. Fig 4.7 shows the history plot for major independent variables. The mixing looks good and the three chains do not appear to be randomly fluctuating about the same general region of the parameter space. The chain for the independent parameter exhibits no autocorrelation. This plot suggests that the MCMC chain has converged. The plot of Gelman-Rubin in Figure 4.6 suggests that the MCMC sequence has converged in the posterior density as red lines fall towards one for all parameters. In addition, the blue and green lines which represent variance within and between chains suggest that the algorithm converges. Our findings are similar to the result obtained by Salameh et al. (Salameh et al., 2014; Jackman, 2000).

In case of multilevel datasets, classical regression models may fail to account for the dependency structure in the data sets while multilevel models do lead to valid inferences. However, multilevel regression model is more complicated than the standard single-level multiple regression model. One difference is the number of parameters, which is much larger in the multilevel model. This poses problems when models are fitted that have many parameters, and in model exploration. Another difference is that multilevel models often contain interaction effects in the form of cross-level

interactions. Interaction effects are tricky, and analysts should deal with them carefully. Finally, the multilevel model contains several different residual variances, and no single number can be interpreted as the amount of explained variance. In this section two types of the multilevel models have been implemented, which are classical and the Bayesian multilevel models and then their results were compared. One the advantage of multilevel models is that, it have a very good ability to accommodate hierarchical structure. It provide a useful framework for thinking about problems with samples which have hierarchical structure. One advantage of the multilevel modeling approach is that it can deal with data in which the times of the measurements vary from subject to subject. In multilevel modeling, variable selections can be complicated due to predictors. The biggest challenge in multilevel modeling is how to specify the covariance structure. To the best of our understanding, there are weaknesses in the estimation procedure for multilevel. There is no single agreed on the methods of estimation of multilevel parameters. Many methods of estimation can be applied, such as, maximum likelihood estimation (MLE), Restricted Maximum Likelihood Estimation (REML), and Bayesian estimation. One the main issue with the Bayesian technique is the time spent to run the model when using the Markov Chain Monte Carlo (MCMC) techniques, particularly when the data sets are large or probably the variables of the model are too many. It also requires a longer period to draw a final decision from Bayesian models, which implies that models should be run with different priors on the model parameters. In addition, in order to ensure that the prior chosen does not affect the final results, it was mentioned in the literature that sensitivity analysis should be carried out. In the case of this study, it is not done because the parameter estimates obtained from the Bayesian method were similar to that of the classical method.

We have demonstrated a way to analyze small sample binary data with covariates, and have applied two approaches on breast cancer data. The purpose of this study was therefore to introduce Bayesian multilevel as an alternative approach and demonstrate its application to parameter estimation in the setting of generalized linear mixed model for for comparative analysis with the classical multilevel. Findings of this study shows that the Bayesian multilevel via Markov Chain Monte Carlo (MCMC) algorithm proffers an alternative framework for modeling breast cancer data. Both the classical approach and Bayesian approach suggest that age, breast cancer type is associated with modality of treatment given to breast cancer patients in Western Nigeria. A comparison of the classical and Bayesian approach to modeling breast cancer reveals lower standard errors of the estimated coefficients in the Bayesian approach in the setting of generalized linear mixed model. Thus, the Bayesian approach is more stable.

4.6 MULTILEVEL MULTINOMIAL LOGIT REGRESSION MODEL WITH RANDOM EFFECTS

Unlike the model for multilevel discussed in Chapter 3, which assumes the same effect beta for each cumulative logit, the model considered in this paper deals with outcome variable with categories more than two; hence, different independent variables have different slopes, β_j for different logits. Because of the nature of the outcome data, the multilevel multinomial logit model with random effects is needed in such a case to interpret unobserved within-cluster heterogeneity. Generalized linear mixed model is a feasible modeling approach for multilevel ordinal response data by incorporating random effects into the multinomial logistic model, leading to the multilevel multinomial logit regression model with random effects. This article discusses the applicability of Bayesian multinomial logistic regression model on the prediction of histological type to breast cancer patient in two different hospitals in western Nigeria. It presents two approximate methods on multinomial logistic regression estimation; classical method and Bayesian method, to obtain the marginal posterior density for the parameters under consideration. A comparison of these two approaches is carried out to determine the usefulness of Bayesian technique on multinomial logistic regression estimation. R, SAS and WinBUGS (Bayesian Inference using Gibbs Sampling) programs have been used to fit the model. As both of these two methods have suggested; histological type in breast cancer decreases with educational level and increases with age. In addition, it is also shown that Bayesian Multinomial logistic regression is useful in direct computations and it produces very accurate approximations to the posterior density.

4.6.1 Introduction

Breast cancer remains an important caused of death among women both in the developed and less developed world Adebamowo & Ajayi (1999); Ojewusi & Arulogun (2016); Oladimeji et al. (2015) and it is a leading malignancy among western women of Nigeria Ojewusi & Arulogun (2016). Globally, patients diagnosed of breast cancer were estimated to be 1.7 million in 2012 while the prevalence stood at 6.3 million women. According to an estimate, breast cancer is reported to be responsible for close to 508,000 deaths in 2011 and increased to 522,000 in 2012 [WHO] was also the most frequently diagnosed cancer among women in 140 of 184 countries worldwide Organization (2013). It is estimated that breast cancer constituted 22.4% of

new cases of cancer registered in 5 years and this accounted for 35.4% of all cancers among Nigerian women Afolayan et al. (2012). As a result, breast cancer tumors continues to be a serious threat to women particularly Africa. Breast cancer risk factors remains controversial and the data regarding this issue are conflicting as reported by Chen et al. (2016). Some studies have documented that age is an important prognostic factor for breast cancer Ikpat et al. (2002) but many of these studies have not been able to explore the role of histological type and histological grade tumor in relation with socio economic status as an established prognostic factor for breast cancer.

Moreover, breast cancer patients diagnosed less than 2 years after birth often have a poor prognosis (Sotiriou et al., 2003; Rakha et al., 2010) and the tumors have also been found to be higher at time of diagnosis (Rakha et al., 2010; Blamey et al., 2010). Hormone receptor status as well as other breast cancer clinical tumors have also been found to differ by histological type (Weigelt et al., 2010). It has been suggested by past studies on breast cancer that reproductive factors affect the risk of histological types of breast cancer differently (Rakha et al., 2010; Ellis et al., 2005; Sundquist et al., 1999; Simpson et al., 2000). Lobular tumors which are an histological type have shown a strong correlation with age at first birth than other histological types Rakha et al. (2010). Few studies have examined whether breast cancer tumors and age at diagnosis tend to be of a particular histological type or not. Lack of understanding of the risk of prognostic factors associated with breast cancer discourages many people from seeking early intervention or even to admit that symptoms they may be experiencing are related to breast cancer. Deeper knowledge about this as well as the underlying prognostic factors may provide more valuable information for improved early diagnosis of breast cancer as well as treatment.

Therefore, to bridge the paucity of knowledge on the wider influencing role of socio economic status (SES) in relation with histological type, this study employed multilevel multinomial modeling. Multilevel modeling approach allows simultaneous performance of the role each socio economic has on histological type. Highlighting such would contribute to a greater understanding of early intervention among women particularly western Nigeria.

4.6.2 Methods

4.6.3 Study participants and methods

Data pertaining to 237 patients who were diagnosed with breast cancer were extracted from the cancer registry of federal teaching hospital. Extensive variable selection procedures were performed on the 20 variables, and the records of patients

aged 20 years and above were selected for the analysis. The information collected included age, marital status, educational level, religion, race, type of breast cancer, occupation, Lab number, case number, site of the cancer, type of diagnosis, and histological type. With respect to the quality of the data obtained, the main concern was the proportion of hospital records in which some of the relevant variables were absent. Information relating to types of patients were eligible for inclusion in this study: gender, tribe, histological type: infiltrating duct carcinoma or lobular carcinoma, type of treatment received and histological grade I, II or III. For input variable selection, we tried to limit the number of variables and select only the clinically relevant ones.

4.6.4 Ethical considerations

This study is based on secondary analysis of existing breast cancer data, with all personal identifying information removed. The data extracted received ethical permission from the ethical review committee of Federal teaching hospital, Ekiti State, Nigeria.

Multinomial logistic regression model

Multinomial logistic regression model is a technique of analysis which is applicable when the dependent variable under study event consists of more than two categories. The multinomial response could be ordinal (ordered categories) or nominal (unordered categories). Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases. Sample size guidelines for multinomial logistic regression indicate a minimum of 10 cases per independent variable (Chan, 2005).

In contrast to binary logistic regression model that compares one dichotomy, multinomial logistic regression model compares a number of dichotomies. Multinomial approach outputs a number of logistic regression models that make specific comparisons of the response categories. Considering a situation that we have j categories of the response variable, the model consists of $j-1$ logit equations which are fit simultaneously. The probability of a categorical variable in a multinomial model is estimated using maximum likelihood estimation (Bayaga, 2010).

Results

4.6.5 Descriptive statistics

Firstly, an analysis using classical multilevel multinomial logistic mixed effects was performed. Table 3.2 summarizes the results from multilevel mixed model for all the three models set up, some parameters estimates are highly significant ($P < 0.00968$). A total of 237 breast cancer patients were enrolled in the current study. The mean age at diagnosis of breast cancer in Nigeria was 42.2 ± 16.6 years. The 20-34 year age group comprised the most patients among the three groups ($n = 97, 40.9\%$), and the ≥ 70 year age group comprised the fewest patients ($n = 17, 7.2\%$), irrespective of tribe. Younger patients were more likely to have grade II disease than elderly patients ($P < 0.00968$).

4.6.6 Result of classical multilevel multinomial model

Table 4.17 presents the results of fitted models with the estimated effects. In Table 4.17, we set up three models in which model 1 is nested under model 2, and model 2 is nested under models. The essence is that, it makes it possible to compare the three models based on -2LL. Result of model 3 in Table 4.17 shows that there is a positive/negative, statistically significant relationship between patient's socio economic status/ breast cancer stages and their likelihood of having a particular histological type. Specifically, as patient's educational status increases, their likelihood of having infiltrating breast tumors decreases. Model 3 contains three significant covariates (educational status, age group and breast cancer stage). We find significant effect of educational status ($\hat{\beta} = 0.81, p = 0.01$), ($\hat{\beta} = 0.07, p = 0.01$) and breast cancer stages ($\hat{\beta} = -4.01, p = 0.0001$). The estimates for educational status and age group has positive effect on the log odds, whereas breast cancer stages has a negative effect. This implies if the educational status increases by one unit, the corresponding change in the log odds is 0.81. Table 4.17 depicts the odd ratio (OR) from the classical multinomial multilevel models. A significant association between histological type and educational status, age group, type of treatment patients received as well as breast cancer type in western Nigeria (see model M2 and M3). The results indicates that, age group (35-49) and those with at least high school education had 7% and 81% more likelihood of histological type (model M3) while the likelihood is higher in model M2 for patients with at least high school education 1 than model M3. From model M3, patients aged 35 to 49 years are 1.1 times more likely of having infiltrating duct carcinoma when diagnosed with breast cancer than their counterpart. Women who are not from this part of geopolitical zone of Nigeria were seen to be 2.4 times

more likely of developing histologic type more than the yoruba people.

Furthermore, findings reveal that patient with tertiary education are 13% (OR= 0.87)

Table 4.17: Multilevel multinomial model estimates of histological type

Fixed Effects	Model 1	Model 2	Model 3
Intercept1 (infiltrating duct)	-0.23(-8.050, 7.585)	0.24(-6.312, 6.797)	3.05(-11.578, 17.682)
Intercept2 (lobular)	0.90(-6.608, 9.163)	1.88(-4.805, 8.557)	4.75(-9.970, 19.471)
Social characteristics			
Educational level			
at least high school		0.87(0.334, 1.403)*	0.81(0.087, 1.539)*
none (ref)			
Age group			
(20-34)			-0.15(-1.318, 1.011)
(35-49)			0.07(1.019, 1.151)*
(50-69)			-0.20(-1.240, 0.835)
none (ref)			
Demographic characteristics			
Race			
efik/igbo			0.86(-1.059, 2.777)
none (ref)			
Biological characteristics			
Treatment type			
surgery		-1.29(-1.955, -0.628)*	-0.15(-0.939, 0.633)
none (ref)			
Type of breast cancer			
malignant			-4.01(-6.172, -1.853)*
none (ref)			
Histological grade			
well (I)			-0.305(-1.105, 0.496)
moderate (II)			0.230(-0.500, 0.960)
none (ref)			
Random effects			
Error variance intercept	0.716(0.753)	0.323(0.365)	0.115(0.164)
Model Fit			
-2LL	476.71	447.11	417.36

* indicates $p < 0.05$ while ** implies significant likelihood ratio test; ICC= 0.1788

less likely to have infiltrating breast carcinoma tumors compared to those who have less than tertiary education. In the case of breast cancer stages, results show that patient who are diagnosed of benign breast tumor are 79% (OR= 0.208) less likely to suffer infiltrating breast carcinoma tumors compared to those patient who their breast cancer are malignant in a nature. Based on the result of estimated intercept for hospital, there exists statistically significant variation between the histological type among the two hospitals patients attended in Western Nigeria. The covariance pa-

parameter estimate was used for the computation of intraclass correlation coefficient.

$$ICC = \frac{\tau_{00}}{\tau_{00} + 3.29} = 0.1788$$

The result of intraclass correlation coefficient indicates how much of the total variation in the likelihood of patient having a particular form of histological type between the two hospitals. The intraclass correlation coefficient is calculated as 0.1788 representing about 18% of the total variation in the outcome variable is accounted for by the hospitals.

4.6.7 Bayesian Estimation for Multinomial Logistic Regression

In this section, we present the Bayesian multinomial logistic regression in establishing the relationship of histologic type with socio-demographic and biological factors. The Bayesian multinomial logistic regression was fitted by using WinBUGS14 software (Bayesian Inference using Gibbs Sampling). We ran three chains for 1,500,000 iterations with the first 500,000 discarded as a burn-in period. The WinBUGS re-

Table 4.18: Posterior means, posterior standard deviations and 95% credible intervals

Parameter	Mean	SD	MC error	2.5%	97.50%
alpha[2]	3.774	0.151	0.00104	3.464	4.057
alpha[3]	3.725	0.156	0.00122	3.408	4.019
alpha[2,2]	-0.424	0.239	0.00152	-0.899	0.043
alpha[2,3]	-0.155	0.229	0.00166	-0.605	0.296
alpha[3,2]	-0.617	0.255	0.00153	-1.125	-0.129
alpha[3,3]	-0.411	0.247	0.00168	-0.898	0.069

sult in Table 4.18 above shows the Bayesian output of multinomial logit model in which the response variable is histologic type with three categorical level mastocytosis, invasive duct carcinoma and lobular carcinoma, with mastocytosis category of histologic type as the baseline category. The table also shows Node which is the name of the unknown quantity equivalent to the model parameter; the approximation of the mean of the average of the posterior distribution of the unknown quantity which is the coefficient of model parameter and the credible intervals. The Markov chain (MC) error is purely technical like round off error. It quantifies the variability in the estimate that is due to markov chain variability. It also means the estimate of the difference between mean of the sampled values and the true posterior mean. Therefore, it should be as small and also less than 5% of the posterior standard deviation for a parameter. In addition, the table also reveals that MC errors are low in comparison to the corresponding estimate posterior standard deviation. This is

an indication that markov chain has converged. The important aspect of Bayesian estimation is that posterior standard deviation is used as the standard error of the parameter estimate while the range between 2.5% and 97.5% percentiles represents 95% Bayesian credible interval.

Figure 4.8 shows posterior density plot, which provides a graphical representation of the posterior density estimate for each parameter (node). Hence, density plot shows that posterior distribution of each parameter is almost normally distributed.

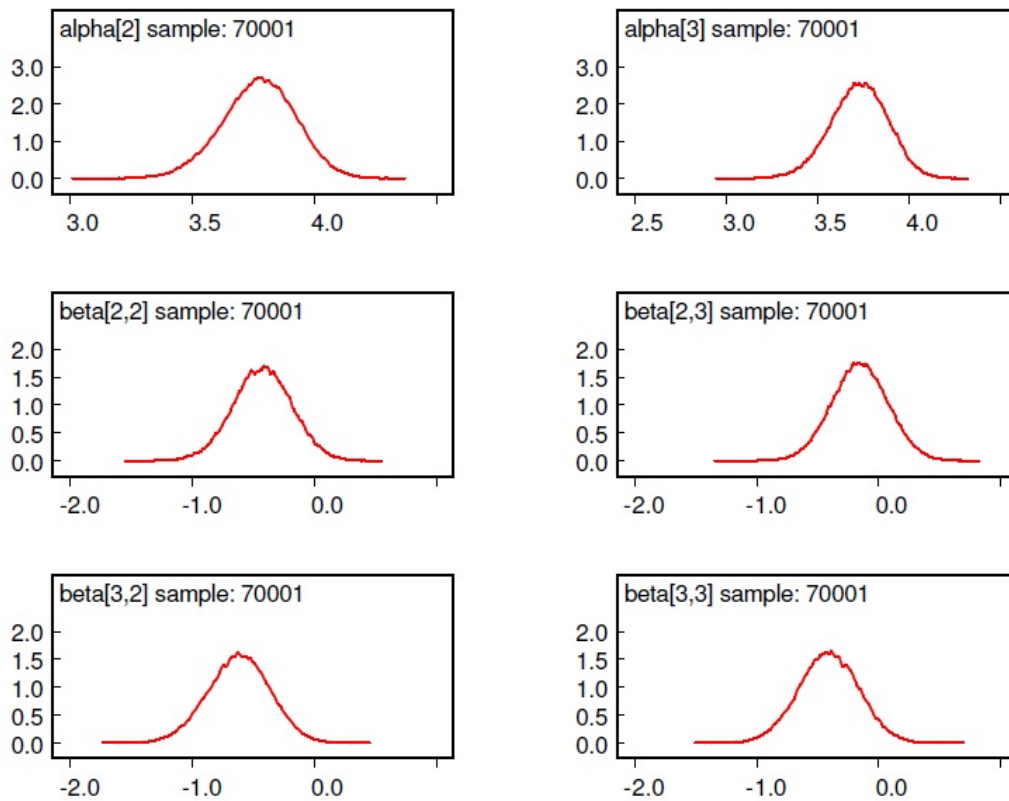


Figure 4.8. Posterior probability density plot

Figure 4.9 shows the autocorrelation plots. These plots appear to dampen quickly; therefore, this provides an evidence of the convergence of the Markov chain and suggests that it may be appropriate to average Markov chain output.

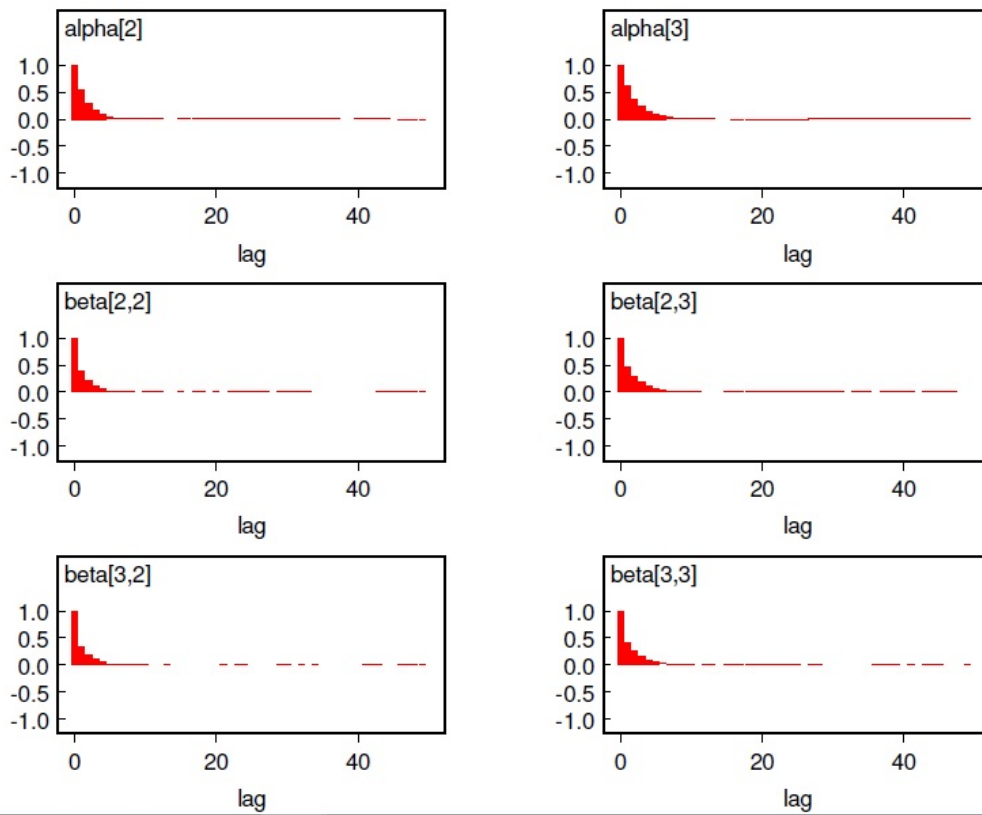


Figure 4.9. Sample autocorrelation plots

Figure 4.10 shows quantile plot. Quantile plot shows the median and the 2.5% and 97.5% percentiles for each iteration. Hence, it is a plot running mean with 95% confidence interval against iteration number. Therefore, it is clear that the requested quantiles have been stabilized implying that Markov chain has converged in terms of parameters.

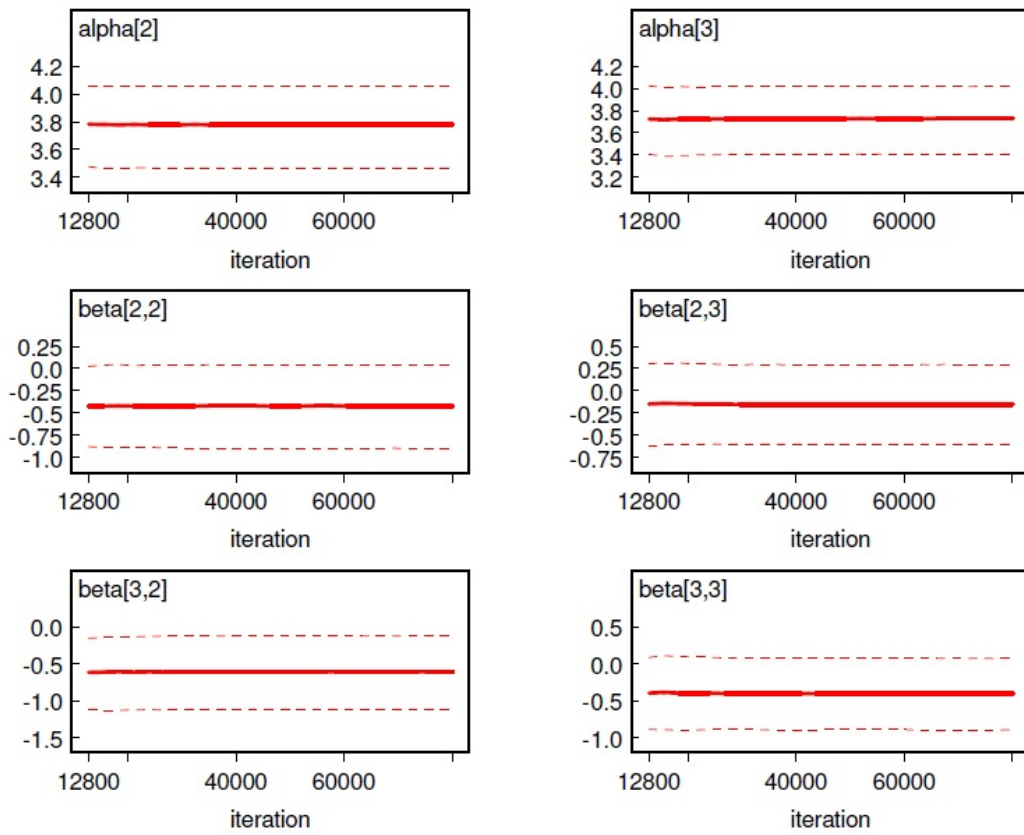


Figure 4.10. Quantile plots

Figure 4.11 shows the visual inspection of the time series plot produced by History. These plots appear to stabilize in. Hence, they suggest that the Markov chain have converged.

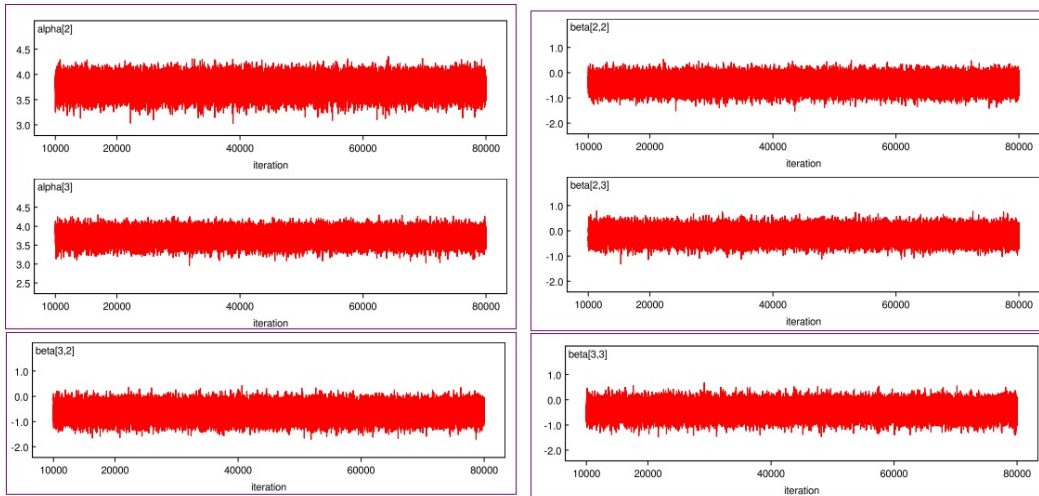


Figure 4.11. Time series plot of MCMC output

4.6.8 Bayesian multilevel multinomial results

In this section we presents the Bayesian approach results we obtained from WinBUGS for the multilevel multinomial for two different models, namely BLOCKING and WITHOUT BLOCKING model. For each model we obtained deviance information criteria (DIC) and the effective number of parameters so as to select the best model. After that we computed the mean estimate, standard error, and the 2.5% and 97.5% quantile credible intervals for the two models. Besides, we also present in this section some of the convergence diagnostics plots.

Model assessment and comparison

Table 4.19: WinBUGS output for the evaluation of logistic regression multilevel using pD and DIC

model	Dbar	Dhat	pD	DIC
M1	411.893	391.435	20.458	432.351
M2	411.889	391.440	20.449	432.339

Table4.19 presents model diagnostics for all the fitted models in Bayesian paradigm and it demonstrates that model M2 has the lowest DIC (432.339) which means it is the best models.

Table 4.20: Posterior means, posterior standard deviations and 95% credible intervals

Node	Mean	SD	MC error	2.5%	Median	97.50%	start	Sample
alpha[2]	3.7745	0.151	0.00101	3.464	3.778	4.057	10000	7001
alpha[3]	3.725	0.156	0.00122	3.408	3.729	4.019	10000	7001
beta[2,2]	-0.424	0.239	0.00152	-0.899	-0.422	0.043	10000	7001
beta[2,3]	-0.155	0.229	0.00166	-0.605	-0.156	0.296	10000	7001
beta[3,2]	-0.617	0.255	0.00153	-1.125	-0.614	-0.129	10000	7001
beta[3,3]	-0.411	0.247	0.00168	-0.898	-0.410	0.069	10000	7001

The WinBUGS result in the Table 4.20 indicates the output of multinomial logistic model which the outcome variable has three categorical levels and the first category is set as the baseline category and the explanatory variables of histological grade has three level of categories with first levels of histological grade set as the baseline categories. The table shows Node which is the name of the unknown quantity equivalent to the model parameter; the approximation of the average of the posterior distribution of the unknown quantity which is the model parameter coefficient, an approximation of the standard deviation of the posterior distribution and computational accuracy of the mean. Furthermore it shows selected percentiles which are median or 50th percentile of the posterior distribution, the 97.5th percentile or an approximation of the upper endpoint of the 95% credible interval and the 2.5 percentile or an approximation of the lower end point of the 95% credible interval. It also shows the starting simulation after discarding the start-up and the number of simulations used to approximate the posterior distribution. The MC error is purely technical like round-off error. It quantifies the variability in the estimate that is due to Markov chain variability, in other words it is an estimate of the difference between the mean of the sampled values and the true posterior mean. Hence it should be small, less than 5% of the posterior standard deviation for a parameter. Hence it can be made as small as possible by increasing the number of simulations. It also suggests that MC errors are low in comparison to the corresponding estimate posterior standard deviation. This is evidence that Markov chain has converged. The posterior standard deviation is used as the standard error of the parameter estimate. The range between 2.5th and 97.5th percentiles represents 95% Bayesian confidence interval and is called a credible interval.

4.6.9 Discussion and partial conclusion

In this section we present the results obtained from the analysis of multilevel multinomial model where we fitted three different models in classical approach and the best model was selected for further analysis in Bayesian approach using WinBUGS software. The essence is to be able to compare the result of the classical approach

with that of Bayesian approach. In Bayesian paradigm, we set up two different models, namely model with blocking and over-relaxation and model (M1) without blocking and over-relaxation (denoted M2). It was mentioned in the literature that model with blocking sometimes aids convergence (Lesaffre & Lawson, 2012), in this thesis we incorporated blocking with over-relaxation in order to see its effect on blocking. For each model we obtained deviance information criteria (DIC), so as to select the best model, the model without blocking (M2) appears to be the best model as it has the lower DIC. These findings are in line with those contained in the literature.

In this study two types of the multilevel models were implemented, namely the frequentist and the Bayesian multilevel models and their results were compared. The results obtained from both approaches are identical. At the end of this study, the results showed no significant difference between Bayesian multilevel and classical multilevel approach. One thing that is unique is that it is difficult to compare the result from Bayesian and classical because the former make use of credible interval while the latter uses confidence interval. The non-informative prior models utilized in the Bayesian approach could have accounted for the similarities between the two approaches. The central aim of the current study is to investigate the association of SES and biological characteristics for histological type of breast tumor among patients diagnosed of breast cancer in western Nigeria using multilevel multinomial regression analysis. The most significant observation we found in this study was that age of occurrence of breast cancer in this environment, the mean age of the patients was 42.2 years, this is similar to those of several Nigerian institutions which have been studied in the literature. Among these studies, Adebamowo reported 43 years (Adebamowo & Adekunle, 1999), Ikpat et al. 42.7 years (Ikpat et al., 2002) and 44.9 years by Ebughe et al. (Ebughe et al., 2013a). But the situation is different in countries with substantially mixed blacks and Caucasians like United States and South Africa, variation is noticed in both incidence and mean age of breast cancer occurrence (Fregene & Newman, 2005; Anderson et al., 2006). Previous studies have mentioned the peculiarities of breast cancer patients among women of African such as genetic factors and reproductive factors. In our study, there is a correlation between the histological type and histologic grade, breast cancer type as well as educational status. This simply means that reproductive and biological factors may determine histologic type of tumors. Other studies have also shown similar result (Ursin et al., 2005; Okobia et al., 2005; Kotsopoulos et al., 2010). But our result is contrary to what is obtainable in other part of Nigeria as reported by (Ebughe et al., 2013b) that reproductive factors may not determine histologic types and biological behavior. Hence, we can attribute this fact that environmental factors may have brought about these changes in the context of breast cancer in Nigeria. The pro-

portions of patients with grade II disease decreased with age, and younger patients were more likely to be diagnosed with a higher grade in this environment. This may be attributed to poor breast cancer screening in young women, as the incidence is high in this age group. In addition, younger patients were more likely to be prone to infiltrating duct carcinoma than the elderly patients, an observations that is supported and consistent with other studies (Chen et al., 2016). It was also found that malignant breast tumor was significant and it happen to be the most prominent type of breast cancer in this environment. This explain why majority of breast cancer patients in this environment are subjected to surgery treatment since the cancer has gotten to a higher level which can only be managed by surgical operation.

The result of Bayesian analysis showed that age group 35-49 years and education were significantly associated with histologic types. We found that patients with at least high school are 9.4 times more likely to have one histologic type of tumor disease than their counterpart. Histologic type of tumor might have resulted from their exposure in advancement in life without observing caution to health management. Although, our study did not investigate the influence of education on breast cancer, a study conducted by (Hussain et al., 2008) evaluated the effect of education on in situ and invasive breast cancer risk using Sweden Family-Cancer Database. Their findings revealed that significant increased risk for in-situ and invasive breast cancer associated with high educational levels. Additionally, previous studies have provided possible explanation for these findings, that highly educated women or women with high socio economic status (SES) are likely to obtain routine breast cancer screening as a result of having access to preventive healthcare (Akinyemiju et al., 2015).

In addition, previous studies also found that a high level of education was significantly associated with decreased incidence of high risk ductal breast cancer among postmenopausal women only. Combining this with our study, these finding indicates heterogeneity in the association between education and breast cancer risk factors exists not only by histologic type and age at diagnosis, but also tumor characteristics (Dalton et al., 2006).

The current study was not without limitations. A few limitations should be considered while interpreting the results from this study, as data collected does not contain information regarding regarding adjuvant chemotherapy or endocrine therapy. Hence, this may have affected our results. In summary, we found that age, histologic grade, breast cancer type were associated with histological type of tumor. Also, consistent with previous findings, our results indicate that the associations between biological factors and the risk of breast cancer differ by histological types of

the tumor. Certain histological types occurred with a significantly higher proportion shortly after women of this geopolitical zone have given birth, which may be as a result of exposure to hazard like chemical and radiation. More interesting was that women with at least high school were more likely to be diagnosed with histologic type, which is a new and unexpected findings in this part of Nigeria. One explanation for these result may be that women with at least high school education present their breast cancer cases to medical practitioners than the less educated women having it in mind that this study uses hospital data. In addition, Bayesian multilevel multinomial regression model helps in selecting the most significant factors between histological type and socio economic status (SES) and biological characteristics of breast cancer tumors as compared to classical multilevel multinomial.

4.7 SOCIO-ECONOMIC DETERMINANTS OF BREAST CANCER RISK FACTORS IN WESTERN NIGERIA: A MULTINOMIAL MODEL

This subsection is a stand alone research article of multinomial modeling of breast cancer in Western Nigeria. The aim is to is to better apprehend determinants of breast cancer by comparing results from three statistics approaches namely; classical, Bayesian and bootstrapping techniques and compared the findings.

Abstract Breast cancer is the most common malignancy among women globally and in Nigeria. Advanced late presentation at diagnosis is a common feature of breast cancer in Nigeria which resulted to poor survival rates. Establishing the prognostic factors is key to preventing death from this type of cancer among women of western Nigeria. Fundamentally, this study presents a comparison of the Bayesian approach with bootstrapping and classical technique in modeling breast cancer. This study used the breast cancer data extracted from cancer registry of Federal Medical Teaching Hospital and a simulated data set. The objective of the study is to better apprehend the socio-economic determinants of breast cancer by comparing results from 3 statistics approaches. Three approaches of multinomial models was applied, namely Bayesian, classical and multinomial technique. Findings highlight that both the classical and the Bayesian model suggest that histologic type is associated with age and educational status. The results also show that breast cancer patient aged 35-49 years had the highest risk of lobular carcinoma compared to mastocytosis. Patient with at least high school education have been found at higher risk of lobular carcinoma and mastocytosis. In addition, the results also highlight that Bayesian approach presented better result followed by bootstrapping and classical. Our find-

ings established the factors that are associated with histologic type of breast cancer patient among women of western Nigeria are age and educational status. It is also concluded that Bayesian Markov chain Monte Carlo algorithm performs better in modeling female breast cancer data in western Nigeria.

4.8 Introduction

Breast cancer [BC] is the most common type of cancer among Nigerian women (Oladimeji et al., 2015; Adebamowo et al., 2003; Adebamowo & Ajayi, 1999; Adebamowo & Adekunle, 1999; Ebughe et al., 2013a) with an overall age standardized rate [ASR] of 52.2 per 100,000 as reported by (Jedy-Agba et al., 2017). Previous studies documented that breast cancer among Nigerian women differs across the states of the federation. For instance, in Northern part of Nigeria breast cancer is the second most common diagnosed in women while western part has breast cancer as the most diagnosed cancer among their women (Ibrahim et al., 2015). However, several studies have assessed the risk factors associated with breast cancer in western Nigeria by using logistic regression model but the current study step further by incorporating Bayesian and bootstrapping and compared their result with classical approach using real life data and bootstrapping data. Although some studies have also utilized the classical multinomial logistic regression in other fields, but no empirical study has explored the application of the classical logistic regression model and compared it to the bootstrapping approach using breast cancer data in western Nigeria.

To the best of our knowledge, this is the first study to investigate socio-demographic and medical factors of breast cancer in two different hospital in western Nigeria using the three different techniques to established the best method in analyzing cancer data. Bootstrapping technique is one of the most popular statistical tools used for assessing uncertainty of unknown quantities in situations where analytical solutions are not available (Davison & Hinkley, 1997; Rubinstein & Kroese, 2016; Efron, 1979). This statistical technique is appealing as a result of its simplicity and it is sometimes regarded as a free distribution technique belonging to a class of non-parametric bootstrap. In this study our focus is to simulate data set using bootstrapping, and describes its implementation for statistics. The approach used for bootstrapping in this study is based on multinomial sampling. The aim of this study is to model breast cancer data via two approaches namely, the classical, Bayesian and the bootstrapping approach. Fundamentally, the results were compared and shown which approach performs better. This paper is organized as follows: classical multinomial logistic regression model is summarized in the next section. Bootstrapping technique and Bayesian inference is introduced in Section 3. The data set and the main

findings are carefully described in Section 4. The diagnostic checking is presented in Section 5. Finally, the discussion section presents some final remarks, points out that Bayesian approach performs better than the classical and bootstrapping technique in analyzing cancer data using a hospital based record.

4.8.1 Ethical Approval

Our study was approved by Ethic Committee of Federal Medical Teaching Hospital, Ekiti State, Nigeria. All the personal information collected was considered confidential.

4.8.2 Methods

Basic concept and notation of the multinomial regression model analysis and bootstrapping are discussed briefly in this section.

4.8.3 Model and Parameters Estimation in Multinomial

This subsection introduces multinomial logistic regression which is applicable when the response variable under consideration consists of more than two categories. Supposing Z is a categorical response variable with three categories, represented as 1, 2, or 3. Since the outcome variable has three categories, we need two logit models as the logistic regression model uses for a binary outcome variable which parameterizes in terms of the logit $z=1$ against $z=0$. We assume there are k explanatory variables, $x = (x_1, \dots, x_k)$ in our model. The logit models for nominal responses pair each response category to a baseline category and the choice is arbitrary. If we set the last category as the baseline, then the baseline category logits are represented as follows

$$\ln \left\{ \frac{p(z=1|x)}{p(z=3|x)} \right\} = \lambda_{10} + \lambda_{11}x_1 + \dots + \lambda_{1k}x_k = \lambda'_1 x \quad (4.2)$$

$$\ln \left\{ \frac{p(z=2|x)}{p(z=3|x)} \right\} = \lambda_{20} + \lambda_{21}x_1 + \dots + \lambda_{2k}x_k = \lambda'_2 x \quad (4.3)$$

Considering the above model, the response probabilities are set up as follows

$$p(z=1|x) = \frac{\exp(\lambda'_1 x)}{1 + \exp(\lambda'_1 x) + \exp(\lambda'_2 x)}$$

$$p(z = 2|x) = \frac{\exp(\lambda'_2 x)}{1 + \exp(\lambda'_1 x) + \exp(\lambda'_2 x)}$$

and

$$p(z = 3|x) = \frac{1}{1 + \exp(\lambda'_1 x) + \exp(\lambda'_2 x)}$$

with parameters $\lambda = (\lambda_1, \lambda_2)$ as the unknown. In the context of this study, the outcome variables is recoded as follows

$$Z_1 = 1, Z_2 = 0, Z_3 = 0 \quad : Z = 1,$$

$$Z_1 = 0, Z_2 = 1, Z_3 = 0 \quad : Z = 2$$

and

$$Z_1 = 0, Z_2 = 0, Z_3 = 1 \quad : Z = 3.$$

Recall that no matter what value Z takes on, the sum of these outcome variables is $\sum_{i=1}^k z_i = 1$. The conditional likelihood function given the covariates for independent observations of sample n is expressed as

$$L(\lambda) = \prod_{j=1}^n \left\{ \left(\frac{e^{\lambda'_1 x_j}}{1 + e^{\lambda'_1 x_j} + e^{\lambda'_2 x_j}} \right)^{z_{1j}} \left(\frac{e^{\lambda'_2 x_j}}{1 + e^{\lambda'_1 x_j} + e^{\lambda'_2 x_j}} \right)^{z_{2j}} \left(\frac{1}{1 + e^{\lambda'_1 x_j} + e^{\lambda'_2 x_j}} \right)^{z_{3j}} \right\}. \quad (4.4)$$

If we take the log on both sides, the above expression reduces to

$$\ell(\lambda) = \sum_{j=1}^n \left\{ z_{1j} \lambda'_1 x_j + z_{2j} \lambda'_2 x_j - \ln(1 + e^{\lambda'_1 x_j} + e^{\lambda'_2 x_j}) \right\}. \quad (4.5)$$

as $\sum_{i=1}^3 z_{ji} = 1$ for each j .

The maximum likelihood estimators are obtained by taking the first partial derivatives of ℓ with respect to each of the unknown parameters and setting these equations equal to zero. In addition, the estimates of the parameters and variance covariance

matrix can be obtained by any standard statistical computer packages like SAS, and R (nnet package).

4.8.4 Bootstrapping technique

In statistics, the use of resampling methods plays a major role especially when the estimators under consideration do not possess an explicit formula. The bootstrap was introduced in 1979 as a computer based method for estimating variance. According to (Boos et al., 2003), bootstrap method is also used to obtain standard errors for estimators, confidence intervals for unknown parameters. Because of modern technological break throughs, the bootstrap has been improved because the modern computer power it requires to simplify intricate calculations (Efron, 1979). The general idea of the bootstrap is to create artificial data-sets with the same structure and sample size as the original data. To create these artificial data-sets, simple random samples are taken from the original with replacement, so that the same PSU may be chosen multiple times and included in the same artificial or pseudo sample. Once the artificial data-sets are chosen, an estimate, θ_b^* of the parameter of interest, θ is calculated from each pseudo sample. Then an estimate of the variance of the parameter of interest is calculated as follows for the bootstrap:

$$Var_{BS}(\theta) = \frac{1}{B-1} \sum_{b=1}^B (\theta_b^* - \theta^*)^2 \quad (4.6)$$

where B is the number of replicated samples and

$$\theta^* = \frac{1}{B} \sum_{b=1}^B \theta_b^*.$$

The issue of how many replicates is required to provide an acceptable variance estimate arises. This problem is not trivial since the precision of the variance estimator continues to increase as the number of replicates increases, but the resources needed to carry out the bootstrap method obviously increases as well (Rust & Rao, 1996). It has been suggested that the number of replicate samples needs to be large. (Efron & Tibshirani, 1993) states that a large B would be 200 replicates, however if confidence intervals will be calculated then it has been suggested that B needs to be 1000 (Efron & Tibshirani, 1993). While most literature when describing the appropriate number of replicates reference, (Efron & Tibshirani, 1993) who says for just variance estimation B = 200 is efficient, several studies have been done showing that perhaps this standard is low. (Booth & Sarkar, 1998) published an article that argued that the

number of replicates should be determined by the conditional coefficient of variation. (Efron & Tibshirani, 1993) suggestion is based on the unconditional coefficient of variation which involves both sampling and resampling variability. (Booth & Sarkar, 1998) argues that only the resampling variability needs to be considered and provides the following simple formula for B,

$$B = \frac{2|\Phi^{-1}(\frac{\alpha}{2})|^2}{\delta^2} \quad (4.7)$$

where α and δ values are obtained by

$$1 - \alpha = P\left(1 - \delta < \frac{\hat{\delta}_B^2}{\delta^2} < 1 + \delta\right) \quad (4.8)$$

where the term in the middle of the above expression is called relative error due to resampling. In this case, δ is a user defined positive constant, $\hat{\delta}_B^2$ is the bootstrap approximation of the variance, δ^2 is the true variance and Φ is the distribution function for the standard normal distribution.

This formula requires approximately 800 replicates to achieve a relative error less than 10% with probability .95 (Booth & Sarkar, 1998). However when considering confidence intervals, (Booth & Sarkar, 1998) article calculates a required B similar to (Efron, 1979) B = 1000. This study used B=1000 bootstrap replicates.

Bootstrapping is related with simulation, but with one major difference. In simulation, the data are obtained completely artificially, while bootstrapping obtains a description of the properties of estimators by using the sample data points themselves, and involves sampling repeatedly with replacement from the actual data. There are two major advantages of bootstrap approach over analytical results of traditional statistical methods.

- Bootstrapping allow individuals to make inferences without making strong distributional assumptions. The bootstrap involves empirically estimating the sampling distribution by looking at the variation of the statistic within sample. Hence, this procedure treating the sample as a population from which samples can be drawn.
- The bootstrap are more robust than the classical approach which makes it effective with relatively small samples and preserved the estimator stability during the periods of unexpected volatility shifts.

Bayesian multinomial model

Suppose y_i is a sequence of categorical data ($i = 1, 2, \dots, N$) and y_i is assumed to take a value in one of the ordered categories given as $0, 1, \dots, c$. For each category of J , with $1 \leq j \leq c$, the probability that y_i takes the value J depends on covariates x_i . Hence, the multinomial logit model is represented as

$$p(y_i = j | \lambda_1, \dots, \lambda_c) = \frac{\exp(x_i \lambda_j)}{1 + \sum_{k=1}^c \exp(x_i \lambda_k)} \quad (4.9)$$

Clearly, equation above is complex which require numerical techniques in order to obtain the marginal posterior distribution for each of the model parameters. The alternative approach is the use of Markov chain Monte Carlo (MCMC) techniques. Markov chain Monte Carlo (MCMC) is a general approach that is employed to generate samples from a distribution π , the equilibrium distribution, which is known up to proportionality constant. The main purpose is to construct a Markov chain that has π as its stationary distribution and then use the chain to estimate functions of the target distribution. Within the Bayesian paradigm the target is the posterior distribution of the model parameters $\pi(\theta|y)$. MCMC encompasses a broad range of algorithms; but this study presents the ones that are relevant to our work. For a more detailed on the MCMC, as well as theoretical results (Gilks et al., 1995; Brooks et al., 2011).

4.8.5 Implementations of Markov chain Monte Carlo (MCMC)

The theory of MCMC guarantees convergence to the correct target distribution but the rate of convergence cannot be typically known in advance. Therefore, it is advisable to examine the output of MCMC in order to check whether the chains have reached their stationarity. Stationarity can be assessed by visual inspection of trace plots or using existing formal diagnostic tests such as (Geweke et al., 1991; Gelman & Rubin, 1992). Moreover, in order to ensure that the samples taken are representative of the target posterior, the early values in the chain are usually discarded as a burn-in period. The length of the burn-in generally depends on the starting values since it will take more iteration to reach stationarity when the initial state of the algorithm is far from the posterior mode. In addition, a further issue concerning MCMC implementation is the autocorrelation within chains. If the output exhibits strong autocorrelation then the samples contain less information regarding the desired distribution compared to when being independent. Also, chains with high autocorrelation may require more iterations to sufficiently explore the parameter space. When dealing with highly correlated chains, one common practice is to do thinning that is save the output every k -th iteration. However, it must be noted that there should be

a balance between the amount of thinning and the cost of sampling.

4.8.6 Bayesian implementation in WinBUGS

We implemented all models in WinBUGS (Spiegelhalter et al., 2003) where the estimates were obtained via MCMC simulation using 1,000,000 iterations (including 500,000 burn-in). We checked convergence by visually assessing the history, chains, and autocorrelation using graphical tools in WinBUGS and using the Geweke method in the BOA package (Smith et al., 2007). We present all posterior estimates as means with the 95% highest probability density intervals (HPDIs)

4.9 Results and Discussion

This section introduces the results obtained from an analysis for both classical, Bayesian and bootstrapping approaches. The aim of this study is to determine significant predictors of breast cancer in Western Nigeria. Fundamentally, this study presents a comparison of the classical, bootstrapping and Bayesian approach. To achieve this, we set up a multinomial logit model for the three approaches and the final outcome is to find predictors of individuals risks of having breast cancer in western Nigeria. To the best of our knowledge, this may be the first study to investigate socio-demographic determinants of breast cancer in this part of Nigeria using three different approaches. Various predictors including medical factors and socio-economic factors are included in the model. The predictors (excluding the reference category) introduced in the model are: Intercept (ψ_0), Age group: 35-49 (ψ_1), 50-69 (ψ_2), 70+ (ψ_3), Treatment received (ψ_4) and Educational status: at least high school (ψ_5).

From descriptive statistics, we found that the percentage of patients with highest breast cancer is among those who had at least high school education. Also, our results indicated that age and educational status is a significant predictors of breast cancer patients with histologic type. In addition, treatment given to breast cancer patient with mixed histologic type is also a significant predictor.

4.9.1 Classical multinomial model

The proc logistic in SAS is used to fit the classical model. From result in Table 4.21, the category histologic type = 1 is chosen as the reference category. The results have two parts, labelled with the categories of response variable. For breast cancer patients having lobular carcinoma, age and treatment (surgery) are significant while educational status is not significant. Findings from this model shows that educational status and age are significant predictors of breast cancer patient with lobular

carcinoma. Furthermore, findings highlight that breast cancer patient aged 35-49 years are 3.8 times more likely to suffer lobular carcinoma compared with patients aged 20-35 years. In addition, the odds ratio of patient aged 35-49 years with lobular carcinoma is 0.62. This simply means that the expected risk of having lobular carcinoma is lower for breast cancer patient in this age group compared to other age groups. The interesting thing about the result of this multinomial regression model concerns with treatment modality. This variable appears to be a significant explanatory factor of breast cancer patient with mastocytosis. The odd ratio associated to the treatment modality highlight that women with mastocytosis but aged 50+ are 75% less at risk of histologic type. The findings of this study support that patient with at least high school education are prone to mastocytosis compared to those with less education.

Table 4.21: Comparative results of the estimated parameters and their standard errors based on the classical, Bayesian and bootstrapping multinomial regression models

Parameter	Classical		Bayesian		Bootstrapping		
	Estimate	S.E	Estimate	S.E	Estimate	S.E	
MORF=2	Intercept	-2.99	0.53	0.97	9.99	0.98	10.01
	35-49	1.35	0.50	0.03	0.36	0.02	0.18
	50-69	1.43	0.70	1.13	0.41	0.12	0.64
	70+	1.44	0.88	-0.19	0.38		
	treatment	1.90	0.49	-0.39	0.36	-0.19	0.39
	education	0.36	0.54	0.16	0.34	-0.21	0.18
MORF=3	Intercept	-3.78	0.77				
	50-69	-0.24	0.89	-0.08	1.45	3.14	3.91
	70+	-13.82	676.0	-0.57	0.33		
	treatment	0.95	0.64	-0.62	0.23	-0.01	1.36
	education	2.34	0.80	-0.41	0.25	-0.57	0.34

4.9.2 Bayesian multinomial model

The WinBUGS software is used to fit the Bayesian model and the same covariates used in classical model are included in the Bayesian model. The Bayesian models were fitted used non informative prior. The results for the Bayesian model with non informative prior are given in Table 4.22. The findings of this study show that the result of the Bayesian model with non informative prior is similar to that of classical model. From the literature, it was established that non informative priors should not have effect on the posterior distribution. The most significant part of the Bayesian inference is that the credible interval is quiet different from the confidence interval

(CI) of classical statistics.

The findings of the Bayesian model with non-informative prior indicate that the variables age, educational status and treatment modality are the only significant predictors of the risk of breast cancer patient with histologic type. Considering the effects of age on the risk of breast cancer patients with histologic type, finding highlight that age group 35-49 years are 42% (OR=0.58) less likely to have mastocytosis compared to other age group. Our result also indicate that patient treated with surgery and having mastocytosis are 98.9% (1.10) less at risk of breast cancer than those that didn't received surgery treatment. In addition, patient with at least high school education are at higher risk of breast cancer in this part of Nigeria.

As part of the requirement for the Bayesian statistics, it is important to check for the convergence of Markov chain Monte Carlo (MCMC). Fig4.12 illustrates the convergence of the Bayesian model with non-informative prior using the quantiles and autocorrelation. The algorithm converged after 1,500,000 iterations. In order to remove autocorrelation and burn-in period, a lag of 40 was considered which requires an iterations up to 1,500,000 iterations the first 500, 000 iterations removed to cater for the burn-in period.

Using the point estimates, Table 4.21 compares the result of classical and Bayesian with bootstrapping respectively. In a comparison of Bayesian and classical model, we observed a lower standard errors of the estimated coefficients in the Bayesian compared to the classical model. An observation that is supported by previous studies (Acquah, 2013; Gordóvil-Merino et al., 2010).

Similarly, in Table 4.21, we found that the estimated means and standard errors

Table 4.22: Posterior Distribution Summaries of parameters from MCMC Multinomial model

Parameter	MorF2			Parameter	MorF3		
	Estimate	MC error	95% Cred.I		Estimate	MC error	95% Cred.I
Intercept	0.97	0.014	(-18.65, 20.54)	Intercept	0.014	0.003	(-18.65, 20.54)
ψ_1	2.70	0.004	(0.69, 4.75)	ψ_1	-0.86	0.004	(-3.11, 1.32)
ψ_2	2.35	0.007	(0.68, 4.75)	ψ_2	-0.08	0.007	(-2.94, 2.74)
ψ_3	3.15	0.014	(-3.23, 9.81)	ψ_3	-1.29	0.013	(-7.74, 5.34)
ψ_4	-3.74	0.001	(-5.65, -1.94)	ψ_4	-4.57	0.002	(-6.74, -2.61)
ψ_5	1.72	0.004	(-0.22, 3.68)	ψ_5	2.72	0.005	(0.70, 4.79)

differs between the Bayesian and bootstrapping. The results show a reduction of standard errors associated with the coefficients obtained from the Bayesian model, thus bringing higher stability to the coefficients.

Fig 4.13 provides a graphical representation of the posterior density estimate for each parameter. This plot indicate normality for the posterior distribution of each

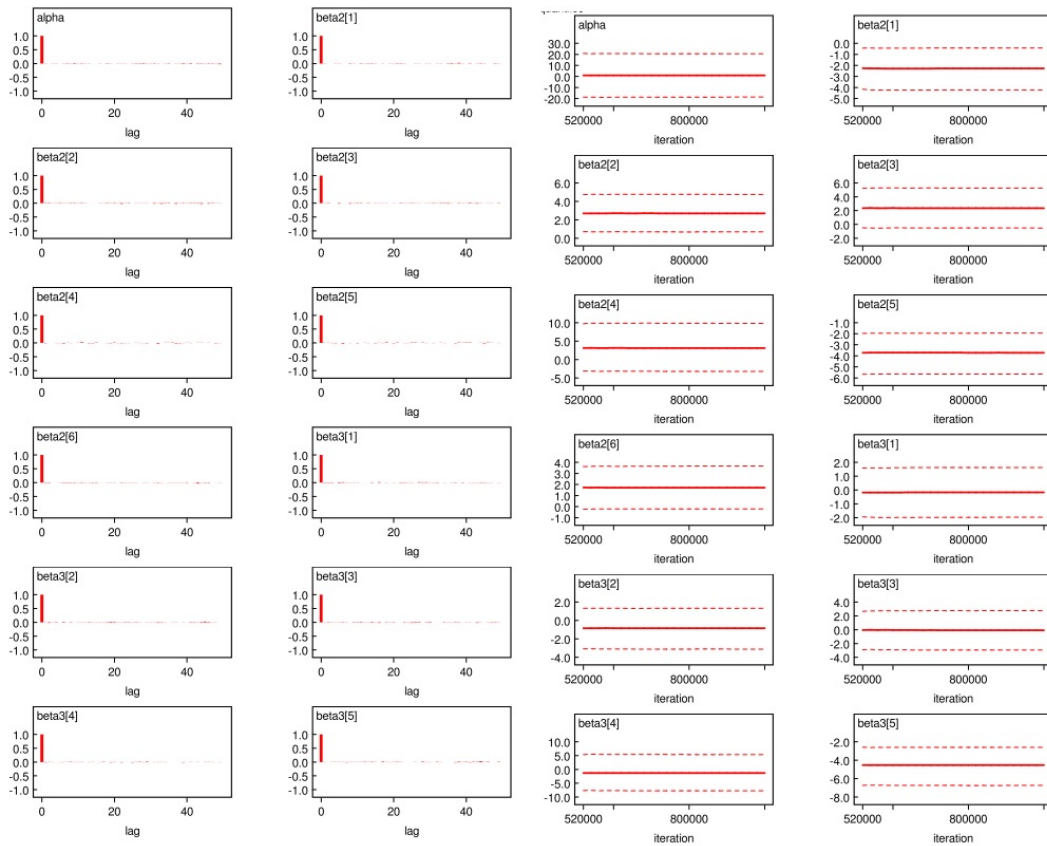


Figure 4.12. WinBUGS' output autocorrelation

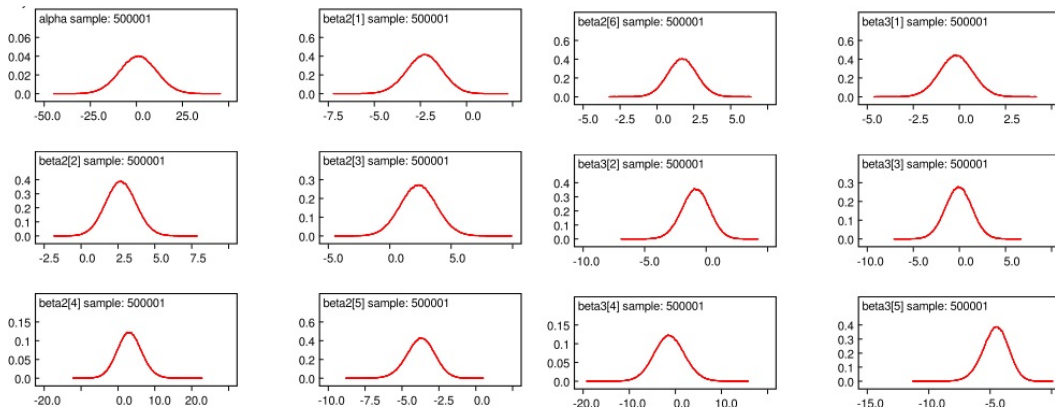


Figure 4.13. WinBUGS' output autocorrelation

parameter.

4.9.3 Discussion and partial conclusion

At the end of this study, our results suggest that there is a significant difference between the three approaches. The model with histologic type as the outcome variable and age, educational status and treatment modality as the covariates was estimated for both the Bayesian, classical and bootstrapping multinomial logistic regression model. A comparison between the three approaches to better apprehend the socio-economic determinants of breast cancer highlights lower standard errors of the estimated coefficients in the Bayesian multinomial logistic regression model. Thus, the Bayesian multinomial logistic regression is more stable. On the other hand, the results from Bayesian approach and classical statistics are difficult to compare because both utilized different tools for decision-making. Moreover, when both approaches produce similar results, findings from Bayesian model are given preference because the technique is more robust and precise than the classical statistics. Our results also give some support to previous findings (Acquah, 2013; Gordóvil-Merino et al., 2010). Addition of priors will actually reduce the variance of the model and thereby lead to a better model in the Bayesian approach. Based on the prior definition and the result from our analysis we concluded that Bayesian approach give better result.

Conclusion

Findings of this paper highlight that age, educational status and treatment modality are the major socio-economic determinants in this study. On the other, the study reveals that Bayesian approach performs better than the bootstrapping and classical method in apprehending the socio-economic determinants of breast cancer in Western Nigeria.

Chapter 5

DISCUSSION AND RECOMMENDATIONS

5.1 Summary of Discussion

In this chapter, an overall outline of our results, their implications in terms of previous findings, and meaning in the context our study is presented. We also explore their implications in the light of the extant literature. We begin this chapter with the determinants of breast cancer types through data analysis. Thereafter, we proceed to a look at a diagnostic determinant of preferred cancer treatment. The comparison between bootstrapping techniques and classical statistics in assessing breast cancer prognostic factors is also discussed in this chapter.

This study has focused on the comparison between two major statistical methods for analyzing breast cancer data. Specifically, we have been concerned with statistical techniques for binary response data, which are gaining more popularity in biostatistics and epidemiological studies. Logistic regression models are most commonly used, but ignore the hierarchical structure typical of such data. This usually leads to biased estimates of parameters, standard errors, wide condence intervals and misleading statistical tests. The main hindrance to using logistic regression analysis is how to incorporate the hierarchical levels into the estimation procedure. In this study, we endeavor to give deep insight into statistical techniques for analyzing data when the response variable is binary. These techniques have been presented with in-depth analysis of a real-life dataset with a binary outcome. This study used cancer data with a binary outcome variable and explanatory variables classied as socio-demographic and medical/biological factors of breast cancer. In order to analyze the dataset, three models have been applied, namely: classical multilevel model, Bayesian multilevel models and bootstrapping methods. Both classical and Bayesian multilevel models were used to model the binary response variable, and for each pa-

parameter in the classical and Bayesian models, we calculate the odd ratio (OR) and 95% confidence interval/credible interval, as the case may be. The results from our findings agreed with previous studies to some extent. The modeling starts with a classical multilevel model because it takes into account the hierarchical structure in the dataset as is the case of our data. For instance, individual patients are nested within the hospital. We then applied the Bayesian approach with the aim of comparing the results of the classical analysis with those of the Bayesian in order to see the best approach between the two methods. We finally proceeded to compare both approaches with bootstrapping methods. The models fitted were Generalized Linear Models (GLM) and Generalized Linear Mixed Models (GLMM). In the case of the Bayesian multilevel multinomial, we used Markov Chain Monte Carlo (MCMC) to estimate the parameters of the Bayesian models in the presence of model with blocking and without blocking. In order to select the best model, we used Deviance Information Criteria (DIC), and established that the model without blocking is the best model because it has the smallest DIC.

The overall mean age of the respondents was 42.2 ± 16.6 years for the entire study. For more detailed analysis on the modeling of breast cancer, MCMC diagnosis were also computed and presented in the second section of chapter 4. The study established that out of 237 breast cancer patients' records extracted for analysis, 192 cases, accounting for 81.01%, were malignant in nature, while 45 cases, representing 18.99%, were benign breast lesions. From this, we established that malignant breast lesions were more prominent in this region of Nigeria compared to other regions within Nigeria.

The results show that breast cancer risk factors in western Nigeria are dependent on age at diagnosis, educational status, and breast cancer type. The outstanding medical factor associated with breast cancer diagnosis among women in western Nigeria was grade tumor. The findings indicated that women with lobular carcinoma were at higher risk, compared to those who had mastocytosis and mixed histologic types. Our findings add to the larger volume of previous studies which report that grade tumor has an effect on breast cancer diagnosis.

This study also showed that the classical logistic regression estimation model has some important limitations, which can be overcome through an alternative technique. Fundamentally, this study introduces the performance of Bayesian analysis as an alternative technique and shows its practicality for the parameter estimation of logistic regression models for comparative analysis with the classical statistics. The study found that the Bayesian Markov Chain Monte Carlo algorithm proffers an alternative framework for estimating the logistic regression model using breast cancer data. Both the classical and Bayesian model suggest that age at diagnosis as well as educational status are risk factors for breast cancer, but a comparison of the

two techniques reveals that the result from Bayesian is more robust than the classical statistics.

Another finding of this study is that the likelihood of breast cancer is higher among women with at least high school education. This is an unexpected finding in this part of Nigeria. This finding may be attributed to two reasons: Foremost, those who are educated may be interested in presenting their health problems to physicians rather than consulting quack medical health practitioners, and as such, be entered into the hospital's patient records. Secondly, the higher proportion of breast cancer in this class of people may be because of the advancement in lifestyle without the observation of proper health management practices. In another development, the Bayesian analysis results indicated that those who had at least high school education were 1.3 times more likely to have malignant breast lesions compared to those with less education. It was also found that a significant increased risk for in-situ and invasive breast cancer is associated with a high level of education. Other studies have also shown similar results (Akinyemiju et al., 2015). The possible explanation for this is that highly-educated women or women with high socio-economic status (SES) are likely to obtain routine breast cancer screening, as a result of having access to preventive health care.

The second approach was a multilevel logistic regression model that assumed that the data structure in the population was hierarchical. We considered a two-level model for our analysis. This second approach considered the results from classical and Bayesian multilevel models. Bayesian techniques provide a better accurate estimate of the parameters and the uncertainty associated with them than the classical approach. The third approach used was a multilevel multinomial model as well as multinomial model approach where model parameters were estimated under the classical and Bayesian approach paradigms. The results of this approach was compared with the bootstrapping technique.

Bayesian inference was carried out via Markov Chain Monte Carlo (MCMC) simulation. Markov Chain Monte Carlo was implemented through the Metropolis-Hastings algorithm using the WinBUGS14 statistical software, which is considered as a very powerful tool for Bayesian computation. In Bayesian analyses, the standard problems in MCMC simulation are the assessment of convergence and the determination of the length of the burn-in period. We employed several graphical tools and the Brooks-Gelman-Rubin (BGR) convergence statistic from WinBUGS and the CODA/BOA package in R to assess convergence and determine the length of the burn-in period. In addition, we present all posterior estimates as means with the 95% highest probability density intervals (HPDIs).

In conclusion, this study provides a closer look at the prevalence of breast cancer in Nigeria, specifically with regard to the Western region. The findings are useful in

terms of identifying risk factors of breast cancer, and for strengthening prevention, sensitization, policy making, and treatment strategies.

5.2 Limitations of the study

The limitation of this thesis can be broken down in two major ways: data and methodological limitations.

5.2.1 Data Limitations

This study uses data extracted from breast cancer patient records in an hospital. These types of data are otherwise known as hospital-based cancer registries. Hospital-based cancer registries are mainly concerned with the recording of information on the cancer patients seen in a particular hospital. Therefore, these types of data cannot provide measures of the occurrence of cancer in a defined population, which put a limit on the level of analysis we could perform in the current study. With respect to the quality of the data extracted from the patient records, the major challenge was the proportion of records in which some of the relevant variables were not recorded or missing. Information regarding variables like age at menarche, weight, height, or age at first term pregnancy were missing. In spite of these limitations however, the data had several merits which made it useful.

5.2.2 Methodological Limitations

The focus of this thesis is on the performance of the Bayesian approach in analyzing breast cancer data. The major limitation of Bayesian analysis is the time needed to run models via Markov Chain Monte Carlo (MCMC) methods of estimation, particularly when the datasets are very large or when many variables are included in the model. Bayesian multilevel model fitting via MCMC simulation is computer-intensive. However, Integrated Nested Laplace Approximation (INLA) can be employed instead. All the Bayesian analyses in this thesis were carried out in WinBUGS via MCMC, resulting in models which took many days to fit. Also, another limitation is that models should be run with different priors on the parameters, which may increase the time taken to draw final conclusions from Bayesian models. Nevertheless, one can extend this work to compare the methods used to analyze cross sectional data by computer simulation to show which model is the effective. One can also fit the models and estimates to the parameters using INLA and MCMC and compare both approaches. In addition, multilevel method can be applied to meta analysis, because meta analysis can be viewed as a special case of multilevel analysis. Another future application in multilevel models is applying bootstrap technique

in the framework of hierarchical linear modeling. Also, small area estimation from unit level in a data with generalized linear mixed model can be used considered as possible direction for future work.

5.3 Scientific papers and articles

From this thesis, four articles have been sent for publication. These are;

- Ogunsakin RE. and Siaka Lougue (2017): Bayesian Inference on Malignant Breast Cancer in Nigeria: A Diagnosis of MCMC Convergence. Asian Pacific Journal of Cancer Prevention (Published)
- Ogunsakin RE. and Siaka Lougue (2017): Bayesian generalized linear mixed modeling of Breast Cancer in Nigeria. Iranian Journal of Public Health, Paper accepted for publication (In press).
- Ogunsakin RE. and Siaka Lougue (2017): Multilevel Multinomial Logit Model for the Bayesian analysis of Breast Cancer data. Sent for publication (Under review).
- Ogunsakin RE. and Siaka Lougue (2017): Socio-Economic Determinants of Breast Cancer risk factors in Western Nigeria: A multinomial model

ARTICLE 1 Published

RESEARCH ARTICLE

Bayesian Inference on Malignant Breast Cancer in Nigeria: A Diagnosis of MCMC Convergence

Ropo Ebenezer Ogunsakin*, Lougue Siaka

Abstract

Background: There has been no previous study to classify malignant breast tumor in details based on Markov Chain Monte Carlo (MCMC) convergence in Western, Nigeria. This study therefore aims to profile patients living with benign and malignant breast tumor in two different hospitals among women of Western Nigeria, with a focus on prognostic factors and MCMC convergence. **Materials and Methods:** A hospital-based record was used to identify prognostic factors for malignant breast cancer among women of Western Nigeria. This paper describes Bayesian inference and demonstrates its usage to estimation of parameters of the logistic regression via Markov Chain Monte Carlo (MCMC) algorithm. The result of the Bayesian approach is compared with the classical statistics. **Results:** The mean age of the respondents was 42.2 ± 16.6 years with 52% of the women aged between 35-49 years. The results of both techniques suggest that age and women with at least high school education have a significantly higher risk of being diagnosed with malignant breast tumors than benign breast tumors. The results also indicate a reduction of standard errors is associated with the coefficients obtained from the Bayesian approach. In addition, simulation result reveal that women with at least high school are 1.3 times more at risk of having malignant breast lesion in western Nigeria compared to benign breast lesion. **Conclusion:** We concluded that more efforts are required towards creating awareness and advocacy campaigns on how the prevalence of malignant breast lesions can be reduced, especially among women. The application of Bayesian produces precise estimates for modeling malignant breast cancer.

Keywords: Bayesian - malignant breast cancer – MCMC

Asian Pac J Cancer Prev, **18 (10)**, 2709-2716

Introduction

Breast cancer is the commonest cause of mortality and morbidity among women worldwide, and currently the most common cancer among Nigerian women (Adebamowo and Ajayi, 1999; Ebughe et al., 2013; Ojewusi and Arulogun, 2016; Oladimeji et al., 2015; Banjo, 2004). The human breast is a pair of mammary glands composed of specialized epithelium and stroma in which both benign and malignant lesions can occur (Dauda et al., 2011). Benign breast constitutes the larger of the breast lesions but much concern is given to malignant lesions of the breast since breast cancer is the most frequent malignancy in the majority of the women (Uwaezuoke and Udoye, 2014). Globally, breast cancer accounts for 18.4% of cancers associated with women. In 2012, (Jedy-Agba et al., 2012) reported that the incidence of breast cancer in Nigeria has risen significantly with the incidence in 2009-2010 reported to be at 54.3 per 100 000, thereby representing a 100% increase within the last decade. The report about patients diagnosed with breast cancer in eastern Nigeria suggested that every 1 out of 5, representing 23%, are malignant in nature (Yusufu et al., 2003). From literature, we found that previous studies

only focused on benign breast cancer (Abudu et al., 2007; Adesunkanmi and Agbakwuru, 2000; Forae et al., 2014; Guray and Sahin, 2006; Kumar et al., 2014; Anyikam et al., 2008; Godwins et al., 2011; Ochicha et al., 2002).

The application of the Bayesian technique and its usage to analyze cancer data has proliferated in recent years. Several researchers such as (Acquah, 2013) studied the comparison of Bayesian and classical and found that Bayesian gave a better result than the classical statistics. Other studies have also shown similar result (Yu and Wang, 2011; Mila and Michailides, 2006; Albert, 1996; Congdon, 2014; Marrelec et al., 2003; Daíz and Batanero, 2016; Lozano-Fernández, 2008; Gordóvil-Merino et al., 2010). In general, studies comparing both methods find that Bayesian technique proffers a better solution compared to classical statistics. The Bayesian technique assumes model parameters as random variables and not as constants, while the probability of the unascertained parameters can be obtained via Bayes theorem (Congdon, 2005; O'Neill, 2002; O'Neill et al., 2000; Wong and Ismail, 2016). This provides information regarding parameter uncertainty that might be very difficult to obtain using the classical technique. Classical technique fits the logistic regression by means of an iterative approach

Statistics Department, University of Kwa Zulu Natal, Westville Campus, Durban, South Africa. *For Correspondence: 215082165@stu.ukzn.ac.za

and in some cases, as a result of this iterative approach, convergence may be difficult to achieve. The robustness and accuracy of the results produced by Bayesian approach makes its gain popularity in data analysis. As such, this paper investigates the significant predictors as well as characterizing patients diagnosed of benign and malignant breast cancer lesion using both classical approach and Bayesian approach.

Materials and Methods

Data Collection

Ethical approval was obtained from the ethics committee of the Federal Medical Teaching Hospital, Ekiti State, Nigeria. This data was extracted from cancer registry of the Federal Medical Teaching Hospital. We accessed 237 records and 20 variables of breast cancer data. Some of these variables describe socio-demographic and cancer-specific information of an incidence of breast cancer. Extensive variable selection procedures were performed on the 20 variables. The records of patients aged 20 years and above were sorted out for this analysis. Information collected includes age, marital status, educational level, religion, race, type of breast cancer, occupation, Lab number, case number, site of the female breast cancer, type of diagnosis and histological type. Other information recorded was the modality of treatment received: surgery, chemotherapy, hormonal therapy, radiotherapy. The software R was used for the classical statistical analysis and the software WinBUGS14 for the Bayesian analysis. As a requirement of the Bayesian approach, several diagnostics tests were performed to answer convergence of the Markov chain Monte Carlo (MCMC) algorithm and the true reflection of the posterior distribution.

Bayesian Binary Logistic Regression

Bayesian logistic regression, which applies Bayesian inference, has the formulation of a logistic equation and includes both continuous and categorical explanatory variables. Binary regression model is used to describe the probability of a binary response variable as function of some covariates. The logistic regression model belongs to the class of Generalized Linear Models. Generalized linear models generalize the standard linear model:

$$\Phi(\eta) = \tau_i^T \lambda \tag{1}$$

Binary logistic regression model is represented as:

$$\pi_i | \eta_i \sim \text{Bin}(\eta_i) \tag{2}$$

If the response under consideration is observed, we have $\pi_i = 1$ for the i th individual and zero otherwise. And π_i is the probability that the i th individual presents the response under consideration, λ is the j vector of unknown parameters, $\tau_i = (\tau_{i1}, \dots, \tau_{ij})$ represent vector of known covariates associated to the i th individual. Therefore, G is now define as any transformation assuming values between 1 and 0. Since function G can be any arbitrary

cumulative distribution, this study consider only the logistic part as earlier mentioned. Hence,

$$G(\tau_i^T \lambda) = \frac{\exp(\tau_i^T \lambda)}{1 + \exp(\tau_i^T \lambda)}$$

The link function defines the linear predictor as expressed below

$$\varphi_i = G^{-1}(\eta_i) = \lambda_{-1} \tau_{i1} + \dots + \lambda_j \tau_{ij} \tag{3}$$

Suppose η_i denotes the probability of having malignant or benign breast lesion, the logit transformation is expressed as

$$\text{logit}(\eta_i) = \log\left(\frac{\eta_i}{1 - \eta_i}\right) \tag{4}$$

The likelihood function for data $\pi = (\pi_1, \dots, \pi_k)^T$ is expressed as

$$P(\pi | \lambda) = \prod_{i=1}^n [G(\tau_i^T \lambda)]^{\pi_i} [1 - G(\tau_i^T \lambda)]^{1 - \pi_i} \tag{5}$$

For the Bayesian analysis, it is important to provide a joint prior distribution over the parameter space. The preferred prior for logistic regression parameters is a multivariate normal distribution and is given by (Ojo et al., 2017; Ntzoufras, 2011):

$$\lambda_k \sim N(b_0, \Sigma_0^2) \tag{6}$$

where $\Sigma_1^{-1} = \Sigma_0^{-1} + \Sigma_{ML}^{-1}$ and

$$b_1 = \Sigma_1 | \Sigma_{ML}^{-1} | \Sigma_{ML}^{-1} + \Sigma_1 | \Sigma_0^{-1} | b_0$$

Note ML is the maximum likelihood estimate. Hence we have

$$p(\lambda_k) \propto \prod_{i=1}^k \exp\left[-\frac{1}{2}(\lambda_k - b_0)' \Sigma_k^{-1} (\lambda_k - b_0)\right] \tag{7}$$

Therefore, the posterior distribution is represented as follows

$$P(\lambda | \pi) \propto p(\lambda_k) \prod_{i=1}^n [G(\tau_i^T \lambda)]^{\pi_i} [1 - G(\tau_i^T \lambda)]^{1 - \pi_i} \tag{8}$$

The latter part of expression (8) can be regarded as normal distribution for parameters λ and it has no closed form. Posterior distribution is usually of high dimension and analytically intractable which sometimes required knowledge of powerful integration. In order to overcome these difficulties, Markov chain Monte Carlo (MCMC) algorithm is needed (Ngesa et al., 2014). MCMC technique is among of the technique employed to generates the estimates of unknown parameters θ and corrects the values generated in order to have a better estimate of the desired posterior distribution, $p(\theta | \eta)$ (Ojo et al., 2017; Ntzoufras, 2011). When MCMC is employed to generate a sample of $p(\theta | \eta)$, there is need to check that the MCMC

Table 1. Heidelberger and Welch Stationarity and Half-Width Tests for the Bayesian Chains Used in the Diagnosis of MCMC

Param.	Stationarity Test	P-Value			Half-width Test	Mean			Half width		
		C1	C2	C3		C1	C2	C3	C1	C2	C3
λ_0	passed	0.927	0.888	0.308	passed	0.117	0.132	0.129	0.054	0.054	0.055
λ_1	passed	0.591	-0.219	0.364	passed	-1.148	-1.144	-1.148	0.01	0.009	0.01
λ_2	passed	0.0572	0.818	0.204	passed	1.348	1.343	1.35	0.01	0.011	0.011
λ_3	passed	0.394	0.51	0.994	passed	0.836	0.838	0.839	0.013	0.012	0.012
λ_4	passed	0.915	0.987	0.112	passed	1.22	1.216	1.22	0.016	0.016	0.016
λ_5	passed	0.893	0.815	0.313	passed	-0.216	-0.226	-0.228	0.044	0.044	0.044
λ_6	passed	0.808	0.914	0.237	passed	-0.977	-0.977	-0.983	0.038	0.037	0.038
λ_7	passed	0.824	0.94	0.507	passed	-1.536	-1.55	-1.551	0.044	0.044	0.044
λ_8	passed	0.64	0.954	0.163	passed	0.6168	0.613	0.612	0.02	0.042	0.019
λ_9	passed	0.4	0.896	0.092	passed	1.054	1.058	1.053	0.009	0.009	0.009

algorithm converges to the desired posterior distribution (Ojo et al., 2017).

Assessing Bayesian Markov Chain Monte Carlo and Convergence

In this study, non-informative prior were assumed in order not to influence the posterior distribution and it was assumed that $\lambda_k \sim N(1, 0.0001)$. All the Bayesian analysis was carried out using WinBUGS 14 (Ntzoufras, 2011). We ran 1,500,000 Markov chain Monte Carlo (MCMC) iterations, with the initial 200,000 discarded to cater for the burn-in period. The 5,000 iterations left were used for assessing convergence of the MCMC. We assessed MCMC convergence of our model parameters by checking Heidelberger-Welch diagnostic, autocorrelation plot, Gelman-Rubin plots (Gelman et al., 2014a), and running quantiles of the MCMC output.

Gelman-Rubin

The diagnostic of Gelman and Rubin requires two or more chains from over-dispersed starting points by computing the within and between chains variability respectively. Large deviation between two variances implies non-convergence of the chain. If all the chains have converged as expected, the posterior marginal variance estimate is expected to be very close within the chain variance. The test statistics for the Gelman-Rubin diagnostic test can be estimated as follows (Lesaffre and Lawson, 2012):

$$W = \frac{1}{M} \sum_{m=1}^M S_m^2$$

$$B = \frac{n}{M-1} \sum_{m=1}^M (\bar{\theta}_m - \bar{\theta})^2$$

$$\hat{v} = \frac{n-1}{n} W + \frac{1}{n} B$$

where k is the number of iterations of the chains.
 $\hat{R} = \frac{\hat{v}}{W}$

Convergence is monitored when $\hat{R} \rightarrow 1$. \hat{R} is called the estimated potential scale reduction factor (PSRF). Brooks and Gelman (Gelman et al., 2014a) proposed an alternative approach that generalizes the initial method to consider more than one parameter concurrently. The estimate of the posterior variance covariance is now computed as:

$$\hat{v} = \frac{n-1}{n} W + \left(1 + \frac{1}{m}\right) B/n$$

where

$$W = \frac{1}{m(n-1)} \sum_{m=1}^M \sum_{t=n+1}^{2n} (\theta'_m - \bar{\theta}_m)(\theta'_m - \bar{\theta}_m)'$$

and

$$B/n = \frac{1}{M-1} \sum_{m=1}^M (\bar{\theta}_m - \bar{\theta})^2$$

denote the q -dimensional within and between covariance matrix estimates of the q -variate. It then imply, if λ_m is the highest eigen value of, hence

$$\hat{R}^c = \frac{n-1}{n} + \left\{\frac{m+1}{m}\right\} \lambda_m \quad (9)$$

Where \hat{R}^c is the multivariate potential scale reduction factor (MPSRF). Convergence is attained when multivariate shrink factor converges to 1.

Heidelberger and Welch's: In order to test the hypothesis of stationarity, we first propose that we have a sequence $\{X^t : t = 1, 2, \dots, k\}$ from a covariance stationary process with unknown spectral density, $S(\omega)$. Therefore, for, $k \geq 1$

$$R_0 = 0, \quad R_k = \sum_{t=1}^k \theta^t$$

and

$$\bar{\theta} = \frac{1}{k} R_k$$

$$\hat{B}_k(r) = \frac{R_k(r) - [kr]\bar{\theta}}{\{kS(0)\}^{1/2}}, \quad 0 \leq r \leq 1 \quad (10)$$

Table 2. WinBUGS Posterior Summaries for Breast Cancer Patients

	Mean	SD	MC error	2.50%	Median	97.50%	start	Sample
λ_0	0.126	1.973	0.01568	-3.514	0.049	4.251	5,000	49,749
λ_1	-1.147	0.669	0.003012	-2.453	-1.146	0.176	5,000	49,749
λ_2	1.347	0.637	0.002987	0.18	1.316	2.695	5,000	49,749
λ_3	0.838	0.815	0.003789	-0.637	0.796	2.561	5,000	49,749
λ_4	1.219	0.92	0.004805	-0.633	1.224	3.012	5,000	49,749
λ_5	-0.223	1.672	0.001283	3.91	-0.102	4.685	5,000	49,749
λ_6	-0.979	1.523	0.01116	-4.418	-0.836	1.603	5,000	49,749
λ_7	-1.545	1.686	0.01269	-5.25	-1.415	1.371	5,000	49,749
λ_8	0.614	1.133	0.005761	-1.454	0.554	3.002	5,000	49,749
λ_9	1.055	0.59	0.00283	-0.093	1.047	2.241	5,000	49,749

$\ddot{B}_k(r)$ is approximately distributed as a Brownian bridge for large k , where

$$\ddot{B}_k(r) = \{\ddot{B}_k(r) : 0 \leq r \leq 1\}$$

Hence, the null hypothesis for stationarity is now tested using Cramer-von Mises statistic.

Results

Socio-demographic profile of participants

The main goal of this paper is to investigate the significant predictors as well as characterizing patients diagnosed of benign and malignant breast cancer lesion and presents diagnosis of MCMC convergence in western Nigeria, comparing the classical approach and Bayesian approach. Various prognostic factors are considered which include: intercept (λ_0), marital status: separated (λ_1), level of education: at least high school (λ_2), religion: Christian (λ_3), tribe: yoruba (λ_4), age: 35-49 (λ_5), 50-69 (λ_6), 70+ (λ_7), occupation: retired (λ_8), self employed (λ_9). A total of 237 breast cancer patients' data was extracted for analysis in the current study. Of these, 192 cases accounting for (81.01%) were malignant breast lesions, while 45 cases (18.99%) were benign giving a ratio of 4.3:1 for malignant to benign breast lesion. The mean age of the respondents was 42.2 ± 16.6 years with 52% of the women aged between 35-49 years. Table1 shows the Heidelberger and Welch stationarity tests for the Bayesian Markov chain Monte Carlo. It shows the stationarity and convergence during the burn-in period.

Table 2 present the result of MCMC diagnostics for the patients diagnosed with benign and malignant breast cancer. The posterior means were obtained after a burn-in period of 5,000 with Monte Carlo error less than 2%. The posterior means and medians of the coefficient and indicate significance. The results of the posterior provide some evidence about the important variable to be selected while profiling patients diagnosed with malignant breast cancer. For , Table 2 shows that those with at least high school education are 1.3 times more likely than others to have benign breast cancer. The results indicate that women with age ≥ 35 years were at a higher risk of been diagnosed with malignant breast cancer than those with

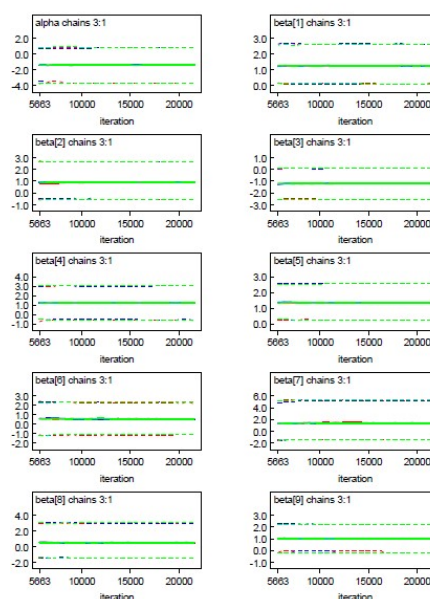


Figure 1. Running Quantiles for the Posterior Parameters in the Case of Female Benign and Malignant Breast Cancer Patients

age < 35 years.

Table 3 shows the result of a classical logistic analysis of the malignant breast cancer. The results indicate that malignant was observed to be strongly associated with age and educational status. This indicates that women with

Table 3. Result of Classical Logistic Regression for Patients Diagnosed of Benign and Malignant

	Est	Std Error	z value	Pr(> z)
λ_0	-2.421	1.2308	-1.967	0.0492
λ_1	1.2479	0.5976	2.088	0.4459
λ_2	0.5926	0.7774	0.762	0.0368
λ_3	1.0782	0.6392	1.687	0.0916
λ_4	1.2048	0.8515	1.415	0.1571
λ_5	1.1952	0.5732	2.085	0.0371
λ_6	0.5034	0.8188	0.615	0.5387
λ_7	1.0534	1.4439	0.73	0.4656
λ_8	0.4823	1.0432	0.462	0.6439
λ_9	0.9898	0.5658	1.749	0.0802

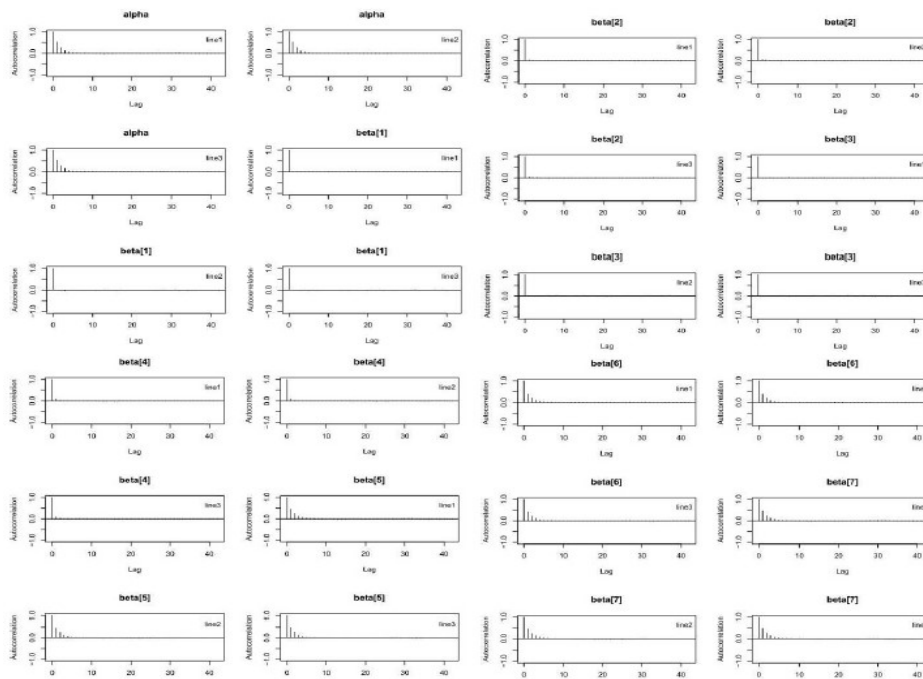


Figure 2. Auto-Correlation Plots for the Female Benign and Malignant Breast Cancer Patients

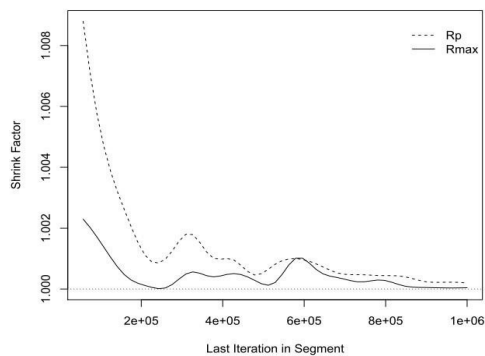


Figure 3. The Plot of the Brooks-Gelman MPSRF for Three Chains of 49,749 Iterations

at least high school education have a significantly higher risk of being diagnosed with malignant breast tumors. *Assessing the performance of Markov Chain Monte Carlo (MCMC) chains in WinBUGS*

When the results of the model are computed, it is necessary to check for the stationarity of Markov chain Monte Carlo algorithm. Both Figure 1 and Figure 2 were presented to demonstrate that there is no problem of autocorrelation among the MCMC chain. The blue and red lines in Figure 4 denote the variance within and between chains. To support that the chain is converged, the ratio must converge to one and the blue and red lines must converge to a stable value. It also displays the red lines representing the potential scale reduction factor

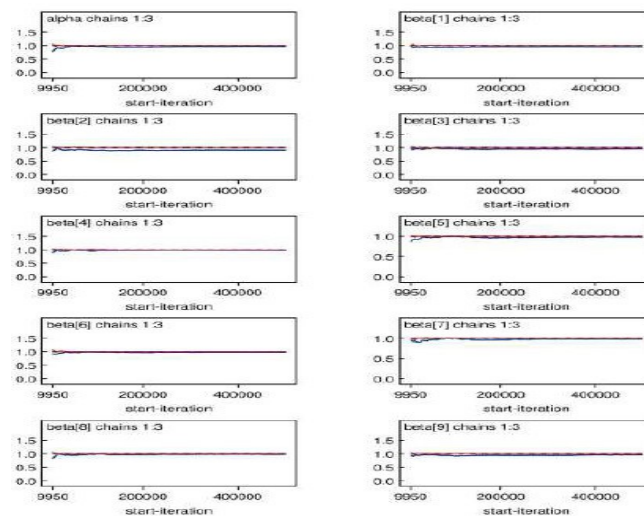


Figure 4. Gelman Rubin Convergence Diagnosis for Independent Variables

denoted by $\hat{\theta}$. Hence, Figure 4 indicates that all the $\hat{\theta}_i \rightarrow \theta_i$ which suggests that the algorithm converges. Both Figure 3 and Figure 4 explain the same thing but one is obtained through CODA/BOA and the other through WinBUGS.

Discussion

The present confirm findings from studies conducted in Nigeria over the past years, on Breast cancer among women in the western Nigeria (Olugbenga et al., 2012; Abudu et al., 2007). All these studies showed that age could be a risk factor for malignant breast lesion. Similar studies have been documented in other parts of Africa and the rest of the world (Arora and Simmons, 2009). From the results of analysis, patient's age 35-49 years constituted the majority of patients (52%) in our study, indicating that women age 35-49 have a higher risk of developing malignant breast cancer than their other counterpart in the group. Therefore, more attention on breast cancer treatment are necessary for this age group. This agrees with breast cancer facts and figures released between 2011-2012. However, this corresponds to the working class population and it is also the child bearing age for many women. This may be as a result of the use of contraceptive and hormonal imbalance which common among the women (Onyeanus, 2015; Olugbenga et al., 2012). From the study, malignant breast lesions appeared to have higher distribution among those who had at least high school education, an observation which supports previous studies (Yuksele et al., 2017; Ibrahim et al., 2015; Yusufu et al., 2003; Ntekim et al., 2009). This result was supported by the descriptive statistics which shows that 52.3% of those diagnosed had at least high school education meaning that those who are educated were more interested in presenting their health problems to rather than consulting fake medical doctors. The high proportion of malignant breast lesion might also be attributed with lifestyle changes among those educated. In addition, this may also be as a result of their exposure to advancement in life like the nature of occupation, diet, without observing caution to health management. We found that the mean age of breast cancer patients in western is 42 years; this is similar to several Nigerian institution based studies, Adebamowo reported 43 years (Adebamowo and Adekunle, 1999), Ikpat et al. 42.7 years (Ikpat et al., 2002) and 44.9 years by Ebughe et al (Ebughe et al., 2013). Although our variables' interactions did not categorize age and educational status as components of socioeconomic status (SES), our findings are similar to those of some studies which showed that higher socioeconomic status (SES) is associated with higher breast cancer incidence (Pudrovska and Anikputa, 2012; Krieger et al., 2010; Vainshtein, 2008). Additional studies have provided a possible explanation for these findings that women with high SES are more likely to obtain routine breast cancer screening due to better access to preventive healthcare based on their level of education and increasing age, hence, increasing the detection of breast cancer (Akinyemiju et al., 2015).

Although our study did not investigate the risk factors for breast cancer in association with its sub molecular

types, a recent study conducted by (Akinyemiju et al., 2015) evaluated the association between SES and breast cancer subtypes using a valid measure of SES and the Surveillance, Epidemiology and End Results (SEER) database. Socioeconomic status based on measures of income, occupational class, education and house value, were categorized into quintiles and explored. Their findings showed that a positive association between SES and breast cancer incidence is primarily driven by hormone receptor positive lesion. Malignant breast lesions which can be subdivided into non-invasive and invasive tumors are documented to be more commonly diagnosed in postmenopausal women (Lehmann et al., 2013). A molecular classification of breast cancer, with more than five reproducible subtypes (basal-like, ERBB2, normal-like, luminal A, and luminal B) has been defined through gene expression profiling and microarray analysis (Lønning et al., 2007). In addition, performing the gene set enrichment analysis (GSEA), a gene set linked to the growth factor (GF) signaling was observed to be significantly enriched in the luminal B tumors (Loi et al., 2009). Another study states that multiple pathways were identified by mapping gene sets defined in Gene Ontology Biological Process (GOBP) for estrogen receptor positive (ER+) or estrogen receptor negative (ER-); and among them, in a separate set, pathways related to apoptosis and cell division or G-protein coupled receptor signal transduction were associated with the metastatic capability of ER+ or ER- tumours, respectively (Jack et al., 2007).

The plot of Gelman Rubin convergence in Fig 4 suggesting that the MCMC sequence has converged on the posterior density as red line fall towards one for all parameters. Our findings are similar to the result obtained by Salameh et al (Salameh et al., 2014), Jackman et al (Jackman, 2000). Figure 3 shows a plot of the Brooks-Gelman MPSRF (denoted \hat{R}) along with the maximum PSRF (denoted \hat{P}) for successively larger segments of the chains. This plot suggests that although the chains differs significantly for the first few thousand iterations, they mix together after that and three chains of 1,500,000 iterations each is probably sufficient to ensure convergence of the chains. It also suggests using a burn-in of about 200,000 each. The result in Table 2 shows that each parameter passes the stationarity and half-width test respectively. This suggests that for the current study, the stationarity of the Markov chain and the sample size obtained is adequate for the estimation of mean values of the three iterations.

Findings from Bayesian and classical inference are not significantly different which could be due to the non-informative prior utilized in the Bayesian model. When both techniques produced similar results, findings from Bayesian are given more attention because it is more robust compared to the classical. The model used in this paper updates quickly and adding complexity will also improve the required time for updating. This diagnostic are necessary to ensure that we are actually sampling from a chain that has converged after a desirable burn-in. Using the posterior mean as a point estimate, Table 3 compares the classical statistics estimates with the simulation (MCMC) result. The estimated means and standard errors appear quite close with minimum results

show a reduction of standard errors associated with the coefficients obtained from the Bayesian approach, hence resulting in higher stability to the coefficients. Other studies have also shown similar result (Gordóvil-Merino et al., 2010; Acquah, 2013).

Findings of this study shows that age of the patients and those with at least high school education are at higher risk of being diagnosed with malignant breast lesion than benign breast lesion in Western Nigeria. The higher proportion of those affected by malignant breast lesion is found among the educated and younger women. Therefore, this shows that non-educated women do not patronize these services based on our findings. More efforts are required towards creating awareness and advocacy campaigns on how the prevalence of malignant breast lesions can be reduced, especially among women.

We recommend that governments, non-governmental organizations and other sectors involved in policy making to put in place policies, strategies and sensitization that target non-educated women to enhance their patronization of breast cancer screening in the health facilities, so as to access the appropriate management health assessment as well as providing financially supported treatments for breast cancer patients.

Statement conflict of Interest

The authors have declared no conflict of interest.

References

- Abudu E, Banjo A, Izegbu M, et al (2007). Malignant breast lesions at olabisi onabanjo university teaching hospital (oouth), sagamu-a histopathological review. *Niger Postgrad Med J*, **14**, 57–9.
- Acquah HDG (2013). Bayesian logistic regression modelling via markov chain monte carlo algorithm. *J Soc Dev Sci*, **4**, 193–7.
- Adebamowo C, Adekunle O (1999). Case-controlled study of the epidemiological risk factors for breast cancer in nigeria. *Br J Surg*, **86**, 665–8.
- Adebamowo CA, Ajayi O (1999). Breast cancer in Nigeria. *West Afr J Med*, **19**, 179–91.
- Adesunkanmi A, Agbakwuro E (2000). Benign breast disease at wesley guild hospital, ilesha, nigeria. *West Afr J Med*, **20**, 146–51.
- Akinyemiju TF, Pisu M, Waterbor JW, Altekruze SF (2015). Socioeconomic status and incidence of breast cancer by hormone receptor subtype. *Springerplus*, **4**, 508.
- Albert JH (1996). Bayesian computation using Minitab. Duxbury Press, pp 117-24.
- Anyikam A, Nzegwu MA, Ozumba BC, Okoye I, Olusina DB (2008). Benign breast lesions in eastern Nigeria. *Saudi Med J*, **29**, 241–4.
- Arora N, Simmons RM (2009). Malignant breast disease: Diagnosis and assessment. *General Surgery*, pp 1481–94.
- Banjo A (2004). Overview of breast and cervical cancers in Nigeria: are there regional variations. In Paper presentation at the International workshop on new trends in Management of breast and cervical cancers, Lagos, Nigeria, pp 1-12.
- Congdon P (2005). Bayesian models for categorical data. John Wiley and Sons, pp 1-10.
- Congdon P (2014). Applied bayesian modelling. *Comput Stat Data Anal*, volume 595. John Wiley and Sons, pp 34-97.
- Dauda A, Misauno M, Ojo E (2011). Histopathological types of breast cancer in gombe, north eastern Nigeria: A seven-year review. *Afr J Reprod Health*, **15**, 107-9.
- Díaz C, Batanero C (2016). ¿cómo puede el método bayesiano contribuir a la investigación en psicología y educación?. *Paradigma*, **27**, 35–53.
- Ebughe G, Ekanem I, Omoronyia O, et al (2013). Age specific incidence of breast cancer in calabar, nigeria. *Int J Trop Dis*, **16**, 1-12.
- Forae GD, Nwachokor FN, Igbe AP, et al (2014). Benign breast diseases in warri southern Nigeria: A spectrum of histopathological analysis. *Ann Nigerian Med*, **8**, 28.
- Gelman A, Carlin JB, Stern HS, Rubin DB (2014a). Bayesian data analysis, volume 2. Chapman and Hall/CRC Boca Raton, FL, USA, pp 261-78.
- Gelman A, Hwang J, Vehtari A (2014b). Understanding predictive information criteria for bayesian models. *Comput Stat*, **24**, 997–1016.
- Godwins E, David D, Akeem J (2011). Histopathologic analysis of benign breast diseases in makurdi, north central nigeria. *Int J Med Sci*, **3**, 125–8.
- Gordóvil-Merino A, Guàrdia-Olmos J, Peró-Cebollero M, de la Fuente-Solanas EI (2010). Classical and bayesian estimation in the logistic regression model applied to diagnosis of child attention deficit hyperactivity disorder. *Psychol Rep*, **106**, 519–33.
- Guray M, Sahin AA (2006). Benign breast diseases: classification, diagnosis, and management. *Oncologist*, **11**, 435–9.
- Ibrahim IM, Iliyasu Y, Mohammed AZ, et al (2015). Histopathological review of breast tumors in kano, northern nigeria. *Sub-Saharan Afr J Med*, **2**, 47.
- Ikpatt O, Ndoma-Egba R, Collan Y (2002). Influence of age and prognosis of breast cancer in Nigeria. *East Afr Med J*, **79**, 651–7.
- Jack XY, Sieuwerts AM, Zhang Y, et al (2007). Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*, **7**, 182.
- Jackman S (2000). Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo. *Am J Political Sci*, **44**, 375–404.
- Jedy-Agba E, Curado MP, Ogunbiyi, et al (2012). Cancer incidence in Nigeria: a report from population-based cancer registries. *Cancer Epidemiol*, **36**, 271–8.
- Krieger N, Chen JT, Waterman PD (2010). Decline in us breast cancer rates after the women's health initiative: socioeconomic and racial/ethnic differentials. *Am J Public Health*, **100**, 132–9.
- Kumar V, Abbas AK, Fausto N, Aster JC (2014). Robbins and cotran pathologic basis of disease. Elsevier Health Sciences, pp 1043-52.
- Lehmann-Che J, Hamy AS, Porcher R, et al (2013). Molecular apocrine breast cancers are aggressive estrogen receptor negative tumors overexpressing either her2 or gdcfp15. *Breast Cancer Res*, **15**, R37.
- Lesaffre E, Lawson AB (2012). Bayesian biostatistics. John Wiley and Sons, pp 176-86.
- Loi S, Sotiriou C, Haibe-Kains B, et al (2009). Gene expression profiling identifies activated growth factor signaling in poor prognosis (luminal-b) estrogen receptor positive breast cancer. *BMC Med Genomics*, **2**, 37.
- Lønning PE, Chrisanthar R, Staalesen V, Knappskog S, Lillehaug J (2007). Adjuvant treatment: the contribution of expression microarrays. *Breast Cancer Res*, **9**, S14.
- Lozano-Fernández LM (2008). Bayesian inference for binomial populations. bayesian estimation for child depression prevalence. *Adv Appl Stat*, **9**, 13–35.
- Marrelec G, Benali H, Ciuciu P, Pelgrini-Issac M, Poline JB (2003). Robust bayesian estimation of the hemodynamic response function in event-related bold fmri using basic

- physiological information. *Hum Brain Mapp*, **19**, 1–17.
- Mila AL, Michailides TJ (2006). Use of bayesian methods to improve prediction of panicle and shoot blight severity of pistachio in california. *Phytopathology*, **96**, 1142–7.
- Ngesa O, Mwambi H, Achia T (2014). Bayesian spatial semi-parametric modeling of hiv variation in kenya. *PLoS One*, **9**, e103299.
- Ntekim A, Nufu F, Campbell O (2009). Breast cancer in young women in ibadan, nigeria. *Afr Health Sci*, **9**, 242-6.
- Ntzoufras I (2009). Bayesian modeling in winbugs. hoboken. **698**, pp 447-60.
- Ntzoufras I (2011). Bayesian modeling using WinBUGS, volume 698. John Wiley and Sons.
- Ochicha O, Edino S, Mohammed A, Amin S (2002). Benign breast lesions in kano. *Niger J Surg Sci*, **4**, 1–5.
- Ojewusi AA, Arulogun OS (2016). Breast cancer knowledge and screening practices among female secondary schools teachers in an urban local government area, ibadan, nigeria. *J Public Health Epidemiol*, **8**, 72–81.
- Ojo OB, Lougue S, Woldegerima WA (2017). Bayesian generalized linear mixed modeling of tuberculosis using informative priors. *PLoS One*, **12**, e0172580.
- Oladimeji KE, Tsoka-Gwegweni JM, Igbodekwe FC, et al (2015). Knowledge and beliefs of breast self-examination and breast cancer among market women in ibadan, south west, nigeria. *PLoS One*, **10**, e0140904.
- Olugbenga AM, Olanrewaju MJ, Kayode OM (2012). Profile of cancer patients attending tertiary health institutions in southwestern nigeria. *Asian J Pharm Clin Res*, **5**, 34–7.
- O'Neill PD, Balding DJ, Becker NG, Eerola M, Mollison D (2000). Analyses of infectious disease data from household outbreaks by markov chain montecarlo methods. *J R Stat Soc Ser C Appl Stat*, **49**, 517–42.
- Onyeanusu CG (2015). Effect of family support on medication adherence and glycemic control of type 2 diabetes outpatients in a tertiary hospital in south-eastern Nigeria. PhD thesis, pp 51-67.
- O'Neill PD (2002). A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain montecarlo methods. *Math. Biosci*, **180**, 103–14.
- Pudrovska T, Anikputa B (2012). The role of early-life socioeconomic status in breast cancer incidence and mortality: unraveling life course mechanisms. *J Aging Health*, **24**, 323–44.
- Salameh P, Waked M, Khayat G, Dramaix M (2014). Bayesian and frequentist comparison for epidemiologists: A non mathematical application on logistic regressions. *Open Epidemiol J*, **7**, 17-26.
- Uwaezuoke SC, Udoye EP (2014). Benign breast lesions in bayelsa state, niger delta nigeria: a 5 year multicentre histopathological audit. *Pan Afr Med J*, **19**, 395.5717
- Vainshtein J (2008). Disparities in breast cancer incidence across racial/ethnic strata and socioeconomic status: a systematic review. *J Natl Med Assoc*, **100**, 833–9.
- Wilhelmsen M, Dimakos XK, Husebø T, Fiskaaen M (2009). Bayesian modelling of credit risk using integrated nested laplace approximations, NR publication, pp 1-25.
- Wong RSY, Ismail NA (2016). An application of bayesian approach in modeling risk of death in an intensive care unit. *PLoS One*, **11**, e0151949.
- Yu L, Wang L (2011). Bayesian and classical estimation of mixed logit model for simulated experimental data. *IEEE J Biomed Health Inform*, **28**, 255–8.
- Yüksel S, Ugras GA, Çavdar, I, et al (2017). A risk assessment comparison of breast cancer and factors affected to risk perception of women in turkey: A cross-sectional study. *Iran J Public Health*, **46**, 308.
- Yusufu L, Odigie V, Mohammed A (2003). Breast masses in zaria, nigeria. *Ann Afr Med*, **2**, 13–6.

ARTICLE 2 Accepted for publication (in process)



Iranian Journal of Public Health
Official Publication of the Iranian Public Health Association

Date: 2017-11-05

Dear Mr. **Ogunsakin R.Ebenezer**

It is my pleasure to inform you that after peering review of your manuscript, entitled:

Bayesian Generalized Linear Mixed Modeling of Breast Cancer

With cooperate:

Ogunsakin R.Ebenezer, Siaka Lougue

is accepted as **Original Article**, for publication in **Iranian J Public Health**.

Please note that the manuscript will go through the editing process, as English style, structural setup, and any misconduct, by the journal. Therefore, the editor reserves the right to reject the paper, in any stage of the process. **Due to the overload of articles in the line of lay outing please note that your article may take *more than one year* to be published, so please let us know for any comment or desire to withdraw the article.**

The Editorial Board will express its sincere appreciation to your collaboration with this journal and wish to receive your valuable upcoming papers.

Best Regards

Prof Dariush D. Farhud, MD, Ph.D., MG

Editor-in-Chief
Iranian J Public Health
Member of WHO Expert Panel
<http://ijph.tums.ac.ir>

P.O. Box: 14155-6446. Tehran, Iran, Tel/Fax: + 98-21-88950184
Email: ijph@tums.ac.ir
URL: <http://ijph.tums.ac.ir>

ARTICLE 3 Still with the reviewer

Multilevel Multinomial Logit Model for the Bayesian analysis of Breast Cancer data

R.E Ogunsakin ^{a,1}, Siaka Lougue ^b

^a*Discipline of Statistics, University of KwaZulu Natal, Westville Campus, Durban, South Africa*

^b*Discipline of Statistics, University of KwaZulu Natal, Westville Campus, Durban, South Africa*

Abstract

Background: Multinomial regression analysis has seen rapid usage in many fields, especially medical and social sciences. Many of these studies make use of multinomial regression, this study step further by incorporating the hierarchical structure as related to multinomial regression via Bayesian techniques. In this study, we develop a multilevel multinomial regression model in WinBUGS and the model is applied to breast cancer data in western Nigeria.

Methods: Multilevel multinomial logistic regression analysis was applied to breast cancer data using two approaches, namely the frequentist multilevel and the Bayesian approach and compared their results. This study used data extracted two different hospitals in western Nigeria to identify key socio-demographic, and biological factors associated with histological types. Several multilevel models have been compared, and a final model was decided based on deviance information criteria.

Results: The mean age at diagnosis of breast cancer in Nigeria was 42.2 years. Bayesian multilevel multinomial model indicated that grade tumor, age group (35-49), patient with at least high school and breast cancer type were all associated with histological types. The results of the two approaches were compared, and the results are similar but preference are given to Bayesian because the approach is more robust than the frequentist. Findings of models from Bayesian revealed that women having invasive duct carcinoma are 2.9 times more at risk of breast cancer than others.

Conclusions: This study identifies key factors associated with histological types in Nigeria. The findings suggest that differences in biological factors like grade tumor and socio-demographic factors such as age at diagnosis and education are more important in assessing risk for breast cancer in western Nigeria. However, as the life expectancy of breast cancer women over the age group 35-49 years increases, treatment should also be based on histological prognostic features of the grade tumor rather than age alone.

¹Corresponding Author: Discipline of Statistics, University of KwaZulu Natal, Westville Campus, Durban, South Africa; E-mail: 215082165@stu.ukzn.ac.za

Introduction

Breast cancer remains an important cause of death among women both in the developed and less developed world (Adebamowo and Ajayi, 1999; Ojewusi and Arulogun, 2016; Oladimeji et al., 2015) and it is a leading malignancy among western women of Nigeria (Ojewusi and Arulogun, 2016). Globally, patients diagnosed of breast cancer were estimated to be 1.7 million in 2012 while the prevalence stood at 6.3 million women. According to an estimate, breast cancer is reported to be responsible for close to 508,000 deaths in 2011 and increases to 522,000 in 2012 [WHO] was also the most frequently diagnosed cancer among women in 140 of 184 countries worldwide (Organization, 2013). It is estimated that breast cancer constituted 22.4% of new cases of cancer registered in 5 years and this accounted for 35.4% of all cancers among Nigerian women (Afolayan et al., 2012). As a result, breast cancer tumors continues to be a serious threat to women particularly Africa. Breast cancer risk factors remains controversial and the data regarding this issue are conflicting (Chen et al., 2016). Some studies have documented that age is an important prognostic factor for breast cancer (Ikpat et al., 2002) but many of these studies have not been able to explore the role of histological type and histological grade tumor in relation with socio economic status as an established prognostic factor for breast cancer.

Moreover, breast cancer patients diagnosed less than 2 years after birth often have a poor prognosis (Sotiriou et al., 2003; Rakha et al., 2010) and the tumors have also been found to be higher at time of diagnosis (Rakha et al., 2010; Blamey et al., 2010). Hormone receptor status as well as other breast cancer clinical features tumors have also been found differed by histological type (Weigelt et al., 2010). It has been suggested by past studies on breast cancer that reproductive factors affect the risk of histological types of breast cancer differently (Rakha et al., 2010; Ellis et al., 2005; Sundquist et al., 1999; Simpson et al., 2000). Lobular tumors which is an histological type have shown a strong correlation with age at first birth than other histological types (Rakha et al., 2010). Few studies have examined whether breast cancer tumors and age at diagnosis tend to be of a particular histological type or not. Lack of understanding of the risk of prognostic factors associated with breast cancer discourages many people from seeking early intervention or even to admit that symptoms they may be experiencing are related to breast cancer. Deeper knowledge about this as well as the underlying prognostic factors may provide more valuable information for improved early diagnosis of breast cancer as well as treatment.

Therefore, to bridge the paucity of knowledge on the wider influencing role of socio economic status (SES) in relation with histological type, this study employed multilevel multinomial modeling. Multilevel modeling approach allows simultaneous performance of the role each socio economic has on histological type. Highlighting such would contribute to a greater understanding of early intervention among women particularly western Nigeria.

Methods

Study participants and methods

Data pertaining to 237 patients who were diagnosed with breast cancer were extracted from the cancer registry of federal teaching hospital. Extensive variable selection procedures were performed on the 20 variables, and the records of patients aged 20 years and above were selected for the analysis. The information collected included age, marital status, educational level, religion, race, type of breast cancer, occupation, Lab number, case number, site of the cancer, type of diagnosis, and histological type. With respect to the quality of the data obtained, the main concern was the proportion of hospital records in which some of the relevant variables were absent. Information relating to types of patients were eligible for inclusion in this study: gender, tribe, histological type: invasive duct carcinoma or lobular carcinoma, type of treatment received and histological grade I, II or III. For input variable selection, we tried to limit the number of variables and select only the clinically relevant ones.

Ethical considerations

This study is based on secondary analysis of existing breast cancer data, with all personal identifying information removed. The data extracted received ethical permission from the ethical review committee of Federal teaching hospital, Ekiti State, Nigeria.

Statistical Analysis

In the present study we employ both classical and Bayesian techniques to identify the socio-economic factors that are associated with histological types in breast cancer patients. Therefore, the hierarchical structure of our dataset where individual patients are nested within hospitals warrants the use of multilevel multinomial modeling approach. Full details of the statistical models employed are discussed extensively. All analysis were performed using SAS 9.4 (Inc, 2014) and WinBUGS 14 Ntzoufras (2011). In this study, the intra clusters correlation (ICC) was used as a measure of random effects.

Multinomial logistic regression model

Multinomial logistic regression model is a technique of analysis which is applicable when the dependent variable under study event consists of more than two categories. The multinomial response could be ordinal (ordered categories) or nominal (unordered categories) but in the current study, the object of interest is not mainly in the ordered categories. Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases. Sample size guidelines for multinomial logistic regression indicate a minimum of 10 cases per independent variable (Chan, 2005).

In case of binary logistic regression model that compares one dichotomy, multinomial logistic regression model compares a number of dichotomies. Multinomial approach out-

puts a number of logistic regression models that make specific comparisons of the response categories. Considering a situation that we have j categories of the response variable, the model consists of $j-1$ logit equations which are fit simultaneously. The probability of categorical variable in a multinomial model is estimated using maximum likelihood estimation (Bayaga, 2010).

Model formulation

Multinomial logistic regression models pairs each outcome category with a reference category. The last category or probably the most common category is assumed to be picked as reference Agresti and Kateri (2011). Suppose $x_i = (x_{i0}, \dots, x_{im})^T$ denote the explanatory variables for individual $1 \leq i \leq n$ and $\beta_j = (\beta_{j0}, \beta_{j1} \dots \beta_{jm})$, ($1 \leq j \leq J-1$), a row vector, represent the regression parameters for the j th reference category. Suppose $y_i = (y_{i1}, y_{i2} \dots y_{iJ})$ denote a multinomial trial for individual $1 \leq i \leq n$. The trial y_{ij} is equal to one whenever a trial occurs in category j . Let $\pi_j(x_i) = Pr((y_{ij} = 1)|x_i)$ be the probability that the i th trial occurs in category j given a set of covariates x_i . Then the multinomial logistic regression model

$$\ln \frac{\pi_j(x_i)}{\pi_J(x_i)} = \beta_j^T x_i \quad j = 1, \dots, J-1 \quad (1)$$

With the logit link we can interpret the coefficients. The exponential of the coefficient, $\exp(\beta_j)$, represents the odds of a trial falling into the category j against category J , all other things equal. A odds greater than one represents that a trial is more likely to occur in category j than J and by symmetry if it is less than one it is more likely to occur in category J than j . If the odds is equal to one, there is independence between y and covariates. Using the logit link we have response probabilities

$$\pi_j(x) = \frac{\exp(\beta_j^T x)}{1 + \sum_{c=1}^{J-1} \exp(\beta_c^T x)} \quad (2)$$

In order to fit the multinomial logit regression model we need to derive the log likelihood for regression parameters. For the log likelihood of the regression parameters we use notation from (Agresti and Kateri, 2011). The likelihood of the regression coefficients is derived from the multinomial likelihood function. For n independent observations

$$L(\beta_i|y_i) = \prod_{i=1}^n \prod_{j=1}^J \pi_j(x_i)^{y_{ij}} \quad (3)$$

while the log likelihood of the above expression is represented as

$$\log L(\beta_i|y_i) = \sum_{i=1}^n \sum_{j=1}^J y_{ij} \log \pi_k(x_i) \quad (4)$$

$$= \sum_{i=1}^n \sum_{j=1}^{J-1} y_{ij} \log \left(\frac{\pi_j(x_i)}{1 - \sum_{k=1}^{J-1} \pi_k(x_i)} \right) + \log \left(1 - \sum_{k=1}^{J-1} \pi_k(x_i) \right) \quad (5)$$

putting the response probabilities in expression (3) into expression (4) results in the log likelihood for our regression parameters as represented below:

$$= \sum_{j=1}^{J-1} \left(\beta_{j0} \left(\sum_{i=1}^n x_{i0} y_{ij} \right) + \sum_{k=1}^m \beta_{jk} \left(\sum_{i=1}^n x_{ik} y_{ij} \right) \right) - \sum_{i=1}^n \log \left(1 + \sum_{j=1}^{J-1} \exp \left(\beta_j x_i \right) \right) \quad (6)$$

Multilevel representation

We consider multilevel multinomial logistic regression model to established prognostic factors for breast cancer in Western Nigeria. Attempts were made to predict the likelihood of patient having some particular histological type. The preliminary analysis started with classical statistics where we set up three models. We consider random intercept model where the model is fitted with only the intercept (model 1). Model 1 is also called unconditional model (i.e model containing no predictors). We used it to calculate the intraclass correlation coefficient (ICC) between hospitals. In model 2, we add educational status and treatment while model 3 include more variables. Based on the values of -2LL, model 3 appears to fits significantly than other models. We therefore resort to model 3 for further analysis in Bayesian analysis technique. The PROC GLIMMIX in SAS is used to fit the classical model and the result is given in table 1. The Bayesian models are fitted using WinBUGS software.

Multilevel multinomial logistic regression model

Suppose η_i denote an ordinal categorical response of a patient i in an hospital j . We then assume that the cumulative probability of η_i taking on a particular value is structured as following

$$\text{logit}(\eta_{kij}) = \beta_{0j} + \beta_{1j} X_{ij} + b_k \quad (7)$$

where η_{kij} is an ordinal categorical outcome for individual i in hospital j , estimated by the overall intercept β_{0j} plus an individual-level error term, b_k . Multilevel models are simple extensions of ordinary regression models, which account for the nesting of data within higher order units. The need for multilevel statistical models is firmly rooted in both theoretical and methodological rationales.

When dealing with multinomial outcomes, multiple logits are simultaneously estimated (i.e M-1 logits, where M is the number of response categories). In our own case where we have three categories of response, there will be three logits and their corresponding intercepts simultaneously estimated, each of them indicating the probability of responding at a particular category. The model for level-1 may be written as:

$$\eta_{lij} = \log \left(\frac{p(\varphi_{ij} \leq k)}{1 - p(\varphi_{ij} \leq k)} \right) = \beta_{0j} + \beta_{1j} X_{ij}, \quad k = 1, 2, \dots, k-1 \quad (8)$$

A random intercept multinomial multilevel logistic model was used to determine factors that were associated with histological tumor in breast cancer patients by considering a 2-level analyses (i.e. hospital level). Multilevel multinomial logistic regression analysis

was performed with histological type as the outcome variable, and considering a 2-level analyses was included for the multilevel modeling analysis. Suppose η_{ijk} represent the status of patient j in hospital i and the associated covariate. The model for level- k may then be expressed as

$$\eta_{ijk} = \log\left(\frac{p(\varphi_{ij} \leq k)}{1 - p(\varphi_{ij} \leq k)}\right) = \beta_{0j} + \beta_{1j}X_{ij} + b_k \quad (9)$$

We assumed that the intercept β_{0j} is random and β_{1j} is the slope coefficient corresponding to X_{ij} that measures the increase in the log odds of patients being at a given age per unit change in the level-1 predictor. And $p(\varphi_{ij} \leq k)$ represents the probability of patient being at or below the k th level of the outcome variable, and b_k represents the difference between the k th category and the preceding one. Normally, for $k=1$, $b_1=0$. Therefore, we consider intercept to be random having the form of model

$$\beta_{0j} = \gamma_{00} + \omega_{0j}$$

Where γ_{00} is the average log odds of a particular patient being in the hospital, and ω_{0j} is the random error term expressed as $\omega_{0j} \sim N(0, \sigma_{00})$. Hence, the model in (9) becomes

$$\eta_{ijk} = \log\left(\frac{p(\varphi_{ij} \leq k)}{1 - p(\varphi_{ij} \leq k)}\right) = \gamma_{00} + \beta_{1j}X_{ij} + b_k + \omega_{0j} \quad (10)$$

Furthermore, model (10) gives log odds when fitted to data and the probability of the event of interest can be expressed as Ene et al. (2015):

$$\phi_{ij} = \frac{\exp(\eta_{kij})}{1 + \exp(\eta_{kij})}$$

Model Building Strategies and Diagnostics

The multilevel multinomial modeling analysis was performed using WinBUGS 14 software. A random intercept multinomial multilevel logistic model was used to determine factors that were associated with histological type tumor in breast cancer patients by considering a 2-level analyses. Different models were constructed and compared in WinBUGS 14. A model with blocking and over-relaxation was specified to aid convergence of the Markov chain. In the next model, blocking and over-relaxation was not included. We consider multilevel multinomial logistic regression model to evaluate the effect of age on breast cancer prognosis. The steps of the model building process are done using Bayesian approach and the result are presented in TABLE 3. The multilevel multinomial are fitted using the same predictors. Model with blocking is specified as model M1 while the one without blocking is called model M2.

The ability to fit complex multilevel models using Markov Chain Monte Carlo (MCMC) techniques presents a need for methods to compare alternative models. The standard model comparison techniques such as AIC, and BIC, require the specification of the number of parameters in each model. For multilevel models which contain random effects, the number of parameters is not generally obvious and as such an alternative technique of comparison is demanded. The most widely used of such alternative technique is the Deviance information criteria (DIC) as suggested by (Spiegelhalter et al., 2002).

Deviance information criteria (DIC) statistic is a generalization of the AIC, and is based on the posterior mean of the deviance, which is also a measure of model complexity and fit. The deviance is defined as

$$D(\theta) = -2\log f(y|\theta)$$

since DIC is a measure of model complexity, it considers a measure of the effective number of parameters in a model, and is defined by

$$pD = \bar{D}(\theta) - (\check{\theta})$$

where $\bar{D}(\theta)$ is the posterior expectation of the deviance, given by

$$\bar{D}(\theta) = -2E \left[\log f(y|\theta) | y \right]$$

and $(\check{\theta})$ is the deviance evaluated at some estimate $\check{\theta}$ of θ . Therefore, we now define the deviance information criteria (DIC) by

$$DIC = \bar{D}(\theta) + pD \quad (11)$$

Where \bar{D} is the posterior mean of the deviance that measures the goodness of fit and pD represent the effective number of parameters in the model. In the case of Bayesian and bootstrapping model, low values of \bar{D} implies a better fit while small values of pD implies a parsimonious model. pD is higher for a more complex model and DIC appears to select the correct model. The best fitting model is one with the smallest Deviance Information Criterion as suggested by (Spiegelhalter et al., 2002; Lesaffre and Lawson, 2012).

Bayesian parameters estimation

A full classical and Bayesian approach in estimation were used in the current study. In the case of Bayesian approach, prior distributions were assigned to all the parameters as discussed below:

Prior distribution

For the prior distribution, this study uses a non-informative. Non informative priors are employed since we want prior information to play a very little role in our analysis which makes the data to remain influential in the analysis. For the purpose of this study, we use a multivariate normal prior on β .

$$\beta_0 \sim N(b_0, \Sigma_0^2) \quad (12)$$

The variance (σ^2) is needed to be transformed before introduced into the model. Hence, we now use $\tau = 0.001$ as the transformed variance. In the case of Bayesian multilevel regression model, each random effect uses a gamma distribution with $\alpha = 0.1$ and $\beta = 0.01$. This study utilizes multivariate normal (b_0, Σ_0) prior density for the parameter

vector β . We also assumed that the prior for the i th component be normal (b_1, S_1^2) while the prior for each component is independent of each other. Therefore,

$$\Sigma_0 = \begin{bmatrix} s_0^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & s_1^2 \end{bmatrix}$$

and

$$b_0 = \begin{bmatrix} b_0 \\ \vdots \\ b_1 \end{bmatrix}$$

Hence, we can therefore write the general formula for prior distribution as follows:

$$p(\beta) \propto \exp \left[-\frac{1}{2} \left(\beta - b_0 \right)' \Sigma_0^{-1} \left(\beta - b_0 \right) \right] \quad (13)$$

Since multivariate normal prior does not have to be made up from independent components, hence, posterior distribution will be multivariate normal (b_1, Σ_1) where

$$\Sigma_1^{-1} = \Sigma_0^{-1} + \Sigma_{ML}^{-1}$$

and

$$b_1 = \Sigma_1 | \Sigma_{ML}^{-1} | \hat{\beta}_{ML} + \Sigma_1 | \Sigma_0^{-1} | b_0$$

where Σ_{ML} is the covariance matrix of the maximum likelihood estimate (MLE) vector and its inverse is represented as

$$\Sigma_{ML}^{-1} = X \Sigma^{-1} X$$

while $\hat{\beta}_{ML}$ is the maximum likelihood vector and Σ_0^{-1} is the prior precision matrix. Posterior distribution refers to the distribution of the parameters after observing the data. The estimates of Bayesian inference is obtained by sampling from the posterior distribution. In terms of Bayesian approach, we can write posterior distribution as

$$p(\beta|y_i) \propto p(y_i|\beta)p(\beta) \quad (14)$$

where $p(y_i|\beta)p(\beta)$ is the likelihood function and is expressed as:

$$p(y|\beta) = \exp \left[-\frac{1}{2} \left(\beta - b_{LS} \right)' \Sigma_{LS}^{-1} \left(\beta - b_{LS} \right) \right]$$

Considering the fixed effects alone, the posterior distribution for a multinomial logistic regression is expressed as:

$$p(\beta|y_i) \propto \exp\left[-\frac{1}{2}\left(\beta - b_1\right)' \Sigma_1^{-1}\left(\beta - b_1\right)\right] \quad (15)$$

Because of the complexity of the posterior distribution which makes it difficult to directly sample from and it becomes complicated when random effects is included in the model. Clearly expression (14) is a complex function which required numerical techniques to obtain the marginal posterior distribution for each of the model parameters. We now employed Markov Chain Monte Carlo (MCMC) approach for simulation of the random numbers from the posterior distribution. In reality, posterior distribution is often of higher dimension and analytically intractable and when the posterior distribution comes from a distribution that is complex, MCMC approach offers a better alternative to summarize the posterior distribution. The quality of MCMC sample depends on how quickly the sampling procedure explore the posterior distribution.

MCMC approach

To estimate the parameters in the models described above, we used a Bayesian approach. In Bayesian approach, the prior knowledge about the parameters is combined with the observed data (likelihood) to yield the posterior distribution. We obtained the posterior summary measures of the parameters by using the MCMC sampling approach (e.g., (Lesaffre and Lawson, 2012; Best, 1996)). We performed the MCMC calculations in WinBUGS14 (Lunn et al., 2000). We used non-informative priors expressing that we do not have prior information on the parameters. For the regression coefficients, we assumed vague independent priors to follow a normal distribution with mean 0 and large variance, that is, $\beta \sim N(0, 10000)$. There were 1500000 iterations, discarding the first 500000 iterations keeping every 100-th iteration of the remaining one million iterations. We assessed MCMC convergence of all models parameters by checking trace plot, kernel density, Gelman Rubin plot (Salameh et al., 2014; Gelman et al., 2003). The scale reduction factor, also known as Gelman-Rubin convergence diagnostic (Clèries et al., 2012) was calculated for model parameter to assess convergence and adequate mixing of the chains. For the model parameter, the scale reduction factor should be close to 1 when convergence is achieved (Clèries et al., 2012). We checked the convergence of these MCMC by using the CODA package in R (Plummer et al., 2010). In particular, we used the Gelman and Rubin's diagnostics measure \hat{R} (Plummer et al., 2010), and this value was close to 1 for all the parameters, which means there was no evidence against convergence.

Results

Descriptive statistics

Firstly, an analysis using classical multilevel multinomial logistic mixed effects was performed. Table1 summarizes the results from multilevel mixed model for all the three models set up, some parameters estimates are highly significant ($P < 0.00968$). A total of 237 breast cancer patients were enrolled in the current study. The mean age at diagnosis of breast cancer in Nigeria was 42.2 ± 16.6 years. The 20-34 year age group comprised

the most patients among the three groups (n= 97, 40.9%), and the ≥ 70 year age group comprised the fewest patients (n= 17, 7.2%), irrespective of tribe. Younger patients were more likely to have grade II disease than elderly patients ($P < 0.00968$).

Result of classical multilevel multinomial model

Table1 presents the results of fitted models with the estimated effects. In Table1, we set up three models in which model 1 is nested under model 2, and model 2 is nested under models. The essence is that, it makes it possible to compare the three models based on -2LL. Result of model 3 in Table1 shows that there is a positive/negative, statistically significant relationship between patient's socio economic status/ breast cancer stages and their likelihood of having a particular histological type. Specifically, as patient's educational status increases, their likelihood of having invasive breast tumors decreases. Model 3 contains three significant covariates (educational status, age group and breast cancer stage). We find significant effect of educational status ($\hat{\beta} = 0.81, p = 0.01$), ($\hat{\beta} = 0.07, p = 0.01$) and breast cancer stages ($\hat{\beta} = -4.01, p = 0.0001$). The estimates for educational status and age group has positive effect on the log odds, whereas breast cancer stages has a negative effect. This implies if the educational status increases by one unit, the corresponding change in the log odds is 0.81. Table1 depicts the odd ratio (OR) from the classical multinomial multilevel models. A significant association between histological type and educational status, age group, type of treatment patients received as well as breast cancer type in western Nigeria (see model 2 and model 3). The results indicates that, age group (35-49) and those with at least high school education had 7% and 81% more likelihood of histological type (model 3) while the likelihood is higher in model 2 for patients with at least high school education 1 than model 3. Furthermore, findings reveal that patient with at least high school education are 13% (OR= 0.87) less likely to have invasive breast carcinoma tumors compared to others. In the case of breast cancer stages, results show that patient who are diagnosed of benign breast tumor are 79% (OR= 0.208) less likely to suffer invasive breast carcinoma tumors compared to those patient who their breast cancer are malignant in nature. Based on the result of estimated intercept for hospital, there exists statistically significant variation between the histological type among the two hospitals patients attended in Western Nigeria. The covariance parameter estimate was used for the computation of intraclass correlation coefficient.

$$ICC = \frac{\tau_{00}}{\tau_{00} + 3.29} = 0.1788$$

The result of intraclass correlation coefficient indicates how much of the total variation in the likelihood of patient having a particular form of histological type between the two hospitals. The intraclass correlation coefficient is calculated as 0.1788 representing about 18% of the total variation in the outcome variable is accounted for by the hospitals.

Table 1. Multilevel multinomial model estimates of histological type

Fixed Effects	Model 1	Model 2	Model 3
Social characteristics			
Educational level			
at least high school		0.87(0.334, 1.403)*	0.81(0.087, 1.539)*
none (ref)			
Age group			
(20-34)			-0.15(-1.318, 1.011)
(35-49)			0.07(1.019, 1.151)*
(50-69)			-0.20(-1.240, 0.835)
none (ref)			
Demographic characteristics			
Race			
efik/igbo			0.86(-1.059, 2.777)
none (ref)			
Biological characteristics			
Treatment type			
surgery		-1.29(-1.955, -0.628)*	-0.15(-0.939, 0.633)
none (ref)			
Type of breast cancer			
malignant			-4.01(-6.172, -1.853)*
none (ref)			
Histological grade			
well (I)			-0.305(-1.105, 0.496)
moderate (II)			0.230(-0.500, 0.960)
none (ref)			
Random effects			
Error variance intercept	0.716(0.753)	0.323(0.365)	0.115(0.164)
Model Fit			
-2LL	476.71	447.11	417.36

* indicates $p < 0.05$ while ** implies significant likelihood ratio test; ICC = 0.1788

Bayesian approach

For the Bayesian method, WinBUGS (Bayesian Inference using Gibbs Sampling) is used to fit the model. The WinBUGS result in Table 2 indicates the output of multilevel multinomial logit model. Considering the credible interval, result of Bayesian multilevel multinomial model indicate that histological grade, age at diagnosis, educational status and breast cancer type are the only significant predictors associated with histologic type of tumor. Patients aged 35-49 years are 3.32 times more at risk of histological type than their counterpart. Regarding the association of grade tumor on patient having histologic

Table 2. Multilevel-multinomial model estimates of histological type of breast cancer

Parameter	Model M_1		Model M_2	
	Estimate	95% Cred.I	Estimate	95% Cred.I
$\psi_2(\text{grade}2)$	-1.215	[-3.110, 0.624]	-1.223	[-3.112, 0.617]
$\psi_2(\text{grade}3)$	-1.503	[-4.628, 1.591]	-1.510	[-4.641, 1.580]
$\psi_2(35 - 49)$	2.934	[0.737, 5.209]	2.932	[0.722, 5.20]
$\psi_2(50 - 69)$	3.125	[-0.037, 6.350]	3.120	[-0.053, 6.344]
$\psi_2(70+)$	4.213	[-2.624, 11.450]	4.197	[-2.648, 11.43]
$\psi_2(\text{treatment})$	0.690	[-1.866, 3.314]	0.684	[-1.878, 3.307]
$\psi_2(\text{education})$	1.097	[-1.038, 3.258]	1.101	[-1.02, 3.258]
$\psi_2(\text{tribe})7$	8.360	[-7.821, 27.940]	8.337	[-7.831, 27.80]
$\psi_2(\text{malignant})$	-27.12	[-59.560, -9.464]	-27.110	[-59.51, -9.468]
$\psi_3(\text{grade}2)$	-1.832	[-3.702, -0.018]	-1.8406	[-3.71, -0.023]
$\psi_3(\text{grade}3)$	-1.309	[-4.339, 1.695]	-1.314	[-4.352, 1.68]
$\psi_3(35 - 49)$	-0.789	[-3.199, 1.560]	-0.794	[-3.219, 1.565]
$\psi_3(50 - 69)$	0.547	[-2.544, 3.633]	0.537	[-2.579, 3.630]
$\psi_3(70+)$	-0.553	[-7.404, 6.576]	-0.582	[-7.444, 6.517]
$\psi_3(\text{treatment})$	-0.537	[-3.342, 2.211]	-0.541	[-3.34, 2.187]
$\psi_3(\text{education})$	2.231	[0.069, 4.457]	2.237	[0.076, 4.478]
$\psi_3(\text{tribe})$	7.295	[-8.90, 27.79]	7.474	[-9.23, 27.88]
$\psi_3(\text{malignant})$	-10.01	[-17.95, -4.203]	-10.01	[-17.970, -4.197]
$\tau.$ Hosp	3.385	[0.011, 17.430]	3.41	[0.014, 17.54]

type, findings highlight that women with grade 2 tumor are 17% (OR=0.83) less likely to have histologic type (i.e. other form) than others with grade 1 or 3. Results also show that patient with at least high school are 9.02 times more at risk to have invasive duct than others. Furthermore, patient with grade 2 tumor are 16.5% (OR=0.835) less likely of having other forms of histologic type of tumor than their counterpart with grade 3 tumor. One of the key features of Bayesian over classical statistics is its ability to check for the convergence of each model. In this study, we checked convergence of the Markov Chain Monte Carlo (MCMC) chain using the Brooks-Gelman-Rubin (BGR) approach in WinBUGS. This approach compares within-chain and between-chain variability for multiple chains starting at over dispersed initial values. Convergence of the chain is monitored by a ratio close to unity (i.e. one).

Multilevel multinomial regression analysis

In this section the multilevel multinomial model with only random intercept was considered. We assessed convergence to ensure how closer we are to the true posterior distribution. Diagnostics tools such as autocorrelation plots, running mean and trace plot and formal test like Gelman-Rubin and Raftery-Lewis were examined. All diagnostic plots for the regression parameter estimated are checked. They indicated that the sampling was done in almost independent manner. The chains did not depend on their initial values and

stationarity was achieved. In general all plots showed good convergence as well as good mixing rate. Potential scale reduction factor for all parameters were less than one, indicating convergence to posterior was achieved. Based on informal and formal tests, we can conclude that the burn-in of 1,500,000 was enough to forget the initial values, there were no dependence of iterations, stationarity and higher mixing rate were achieved. This means that the estimates were derived from the true posterior distribution.

Model assessment and comparison

Table 3. WinBUGS output for the evaluation of logistic regression multilevel using pD and DIC

Model	Dbar	Dhat	pD	DIC
M1	411.893	391.435	20.458	432.351
M2	411.889	391.440	20.449	432.339

Table3 presents model diagnostics for all the fitted models in Bayesian paradigm. Model with a small DIC value provides a better fit. Comparison of the goodness of fit and complexity of the models, model M2 is the preferred model.

Discussion

In this study two types of multilevel models were implemented, namely the frequentist and the Bayesian multilevel models and their results were compared. The results obtained from both approaches are identical. At the end of this study, the results showed no much significant difference between Bayesian multilevel and classical multilevel approach. One thing that is unique is that it is difficult to compare the result from Bayesian and classical because the former make use of credible interval while the latter uses confidence interval. The non-informative utilize in the Bayesian approach could have accounted for the similarities between the two approaches. The central aim of the current study is to investigate the association of SES and biological characteristics for histological type of breast tumor among patients diagnosed of breast cancer in western Nigeria using multilevel multinomial regression analysis. The most significant observation we found in this study was that age of occurrence of breast cancer in this environment, the mean age of the patients was 42.2 years, this is similar to those of several Nigerian institutions which have been studies in the literature. Among these studies, Adebamowo reported 43 years (Adebamowo and Adekunle, 1999), Ikpatt et.al 42.7 years (Ikpat et al., 2002) and 44.9 years by Ebughe.et al (Ebughe et al., 2013a). But the situation is different in countries with substantially mixed blacks and Caucasians like United States and South Africa, variation is noticed in both incidence and mean age of breast cancer occurrence (Fregene and Newman, 2005; Anderson et al., 2006). Previous studies have mentioned the peculiarities of breast cancer patients among women of African such as genetic factors and reproductive factors. In our study, there is a correlation between the histological type and histologic grade, breast cancer type as well as educational status. This simply means that reproductive and biological factors may determine histologic type of tumors. Other studies have also shown similar result (Ursin et al., 2005; Okobia

et al., 2005; Kotsopoulos et al., 2010). But our result is contrary to what is obtainable in other part of Nigeria as reported by (Ebughe et al., 2013b) that reproductive factors may not determine histologic types and biological behavior. Hence, we can attribute this fact that environmental factors may have brought about this changes in the context of breast cancer in Nigeria. The proportions of patients with grade II disease decreased with age, and younger patients were more likely to be diagnosed with a higher grade in this environment. This may be attributed to poor breast cancer screening in young women, as the incidence is high in this age group. In addition, younger patients were more likely to be prone to invasive duct carcinoma than the elderly patients, an observations that is supported and consistent with other studies (Chen et al., 2016). It was also found that malignant breast tumor was significant and it happen to be the most prominent type of breast cancer in this environment. This explain why majority of breast cancer patients in this environment are subjected to surgery treatment since the cancer has gotten to a higher level which can only be managed by surgical operation.

The result of Bayesian analysis showed that age group 35-49 years, educational status and grade tumor were significantly associated with histologic types. We found that patient with at least high school are 9.4 times more likely to have one histologic type of tumor disease than their counterpart. Histologic type of tumor might have resulted from their exposure in advancement in life without observing caution to health management. Although, our study did not investigate the influence of education on breast cancer, a study conducted by (Hussain et al., 2008) evaluated the effect of education on in situ and invasive breast cancer risk using Sweden Family-Cancer Database. Their findings revealed that significant increased risk for in-situ and invasive breast cancer associated with high educational levels. Additionally, previous studies have provided possible explanation for these findings, that highly educated women or women with high socio economic status (SES) are likely to obtain routine breast cancer screening as a result of having access to preventive healthcare (Akinyemiju et al., 2015). In addition, previous studies also found that a high level of education was significantly associated with decreased incidence of high risk ductal breast cancer among postmenopausal women only. Combining this with our study, these finding indicates heterogeneity in the association between education and breast cancer risk factors exists not only by histologic type and age at diagnosis, but also tumor characteristics (Dalton et al., 2006). The grade tumor of disease was also seen to be significant with histologic types. Majority of the tumors were seen to be of high grade tumors as reported from our analysis. An observation that is supported by previous studies (Ikpatt and Ndoma-Egba, 2003; Ebughe et al., 2013b). Moreover, most of the breast cancers occur in the age group less than 50 years, according to previous studies (Henson et al., 2003; Rosen et al., 1984) this is age group where the tumors are expected to be more predominant. Hence, this observation needs proper monitoring and studied in this part of Nigeria in order to unravel the scientific reason.

The current study was not without limitations. A few limitations should be considered while interpreting the results from this study, as data collected does not contain information regarding regarding adjuvant chemotherapy or endocrine therapy. Hence, this may have affected our results. In summary, we found that age, grade tumor, educational status and breast cancer type were associated with histological type of tumor. Also, consistent with previous findings, our results indicate that the associations between biological

factors and the risk of breast cancer differ by histological types of the tumor. Certain histological types occurred with a significantly higher proportion shortly after women of this geopolitical zone have given birth, which may be as a result of exposure to hazard like chemical and radiation. More interesting was that women with at least high school were more likely to be diagnosed with histologic type, which is a new and unexpected findings in this part of Nigeria. One explanation for these result may be that women with at least high school education present their breast cancer cases to medical practitioners than the less educated women having it in mind that this study uses hospital data. However, this study suggests that as the life expectancy of breast cancer women over the age group 35-49 years increases, treatment should also be based on histological prognostic features of the grade tumor rather than age alone. In addition, Bayesian multilevel multinomial regression model helps in selecting the most significant factors between histological type and socio economic status (SES) and biological characteristics of breast cancer tumors as compared to classical multilevel multinomial.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SL and RE equally contributed in the conception of the study, statistical analysis and other aspect of the work. All authors read and approved the final manuscript prior to submission.

Acknowledgements

We acknowledge the efforts of the ethics and research committee of Federal Medical Teaching Hospital [Cancer Registry Unit], Ekiti State, Nigeria and University of Kwazulu Natal, Westville Campus, Durban, South Africa.

References

- Adebamowo, C. and Adekunle, O. (1999). Case-controlled study of the epidemiological risk factors for breast cancer in nigeria. *British Journal of Surgery*, 86(5):665–668.
- Adebamowo, C. A. and Ajayi, O. (1999). Breast cancer in nigeria. *West African journal of medicine*, 19(3):179–191.
- Afolayan, E., Ibrahim, O., and Ayilara, G. (2012). Cancer patterns in ilorin: An analysis of ilorin cancer registry statistics. *Tropical Journal of Health Sciences*, 19(1).
- Agresti, A. and Kateri, M. (2011). *Categorical data analysis*. Springer.
- Akinyemiju, T. F., Pisu, M., Waterbor, J. W., and Altekruise, S. F. (2015). Socioeconomic status and incidence of breast cancer by hormone receptor subtype. *SpringerPlus*, 4(1):508.

- Anderson, W. F., Pfeiffer, R. M., Dores, G. M., and Sherman, M. E. (2006). Comparison of age distribution patterns for different histopathologic types of breast carcinoma. *Cancer Epidemiology and Prevention Biomarkers*, 15(10):1899–1905.
- Bayaga, A. (2010). Multinomial logistic regression: usage and application in risk analysis. *Journal of applied quantitative methods*, 5(2):288–297.
- Best, D. S. A. T. N. (1996). Bugs 0.5* bayesian inference using gibbs sampling manual (version ii) david spiegelhalter andrew thomas nicky best wally gilks.
- Blamey, R., Hornmark-Stenstam, B., Ball, G., Blichert-Toft, M., Cataliotti, L., Fourquet, A., Gee, J., Holli, K., Jakesz, R., Kerin, M., et al. (2010). Oncopool—a european database for 16,944 cases of breast cancer. *European journal of cancer*, 46(1):56–71.
- Chan, Y. H. (2005). Biostatistics 305. multinomial logistic regression. *Singapore medical journal*, 46(6):259.
- Chen, H.-l., Zhou, M.-q., Tian, W., Meng, K.-x., and He, H.-f. (2016). Effect of age on breast cancer patient prognoses: A population-based study using the seer 18 database. *PloS one*, 11(10):e0165409.
- Clèries, R., Ribes, J., Buxo, M., Ameijide, A., Marcos-Gragera, R., Galceran, J., Martínez, J. M., and Yasui, Y. (2012). Bayesian approach to predicting cancer incidence for an area without cancer registration by using cancer incidence data from nearby areas. *Statistics in medicine*, 31(10):978–987.
- Dalton, S. O., Düring, M., Ross, L., Carlsen, K., Mortensen, P. B., Lynch, J., and Johansen, C. (2006). The relation between socioeconomic and demographic factors and tumour stage in women diagnosed with breast cancer in denmark, 1983–1999. *British journal of cancer*, 95(5):653.
- Ebughe, G., Ekanem, I., Omoronyia, O., Nnoli, M., Nwagbara, V., Udosen, J., Umoh, M., and Ugbem, T. (2013a). Age specific incidence of breast cancer in calabar, nigeria.
- Ebughe, G., Ugare, G. U., Nnoli, M. A., Basse, I.-A., Nwagbara, V. J., Udosen, J., Omoronyia, O. E., Chukwuegbo, C. C., Ugbem, T. I., Omotoso, A. J., et al. (2013b). Histological type and tumour grade in nigerian breast cancer: Relationship to menarche, family history of breast cancer, parity, age at first birth, and age at menopause. *Iosrjournal*, 7(5):58–63.
- Ellis, I. et al. (2005). Pathology reporting of breast disease: a joint document incorporating the third edition of the nhs breast screening programme's guidelines for pathology reporting in breast cancer screening and the second edition of the royal college of pathologists' minimum dataset for breast cancer histopathology. *NHS Cancer Screening Programmes, Royal College of Pathologists, London*.
- Ene, M., Leighton, E. A., Blue, G. L., and Bell, B. A. (2015). Multilevel models for categorical data using sas® proc glimmix: the basics. In *SAS Global Forum 2015 Proceedings*, pages 3430–2015.
- Fregene, A. and Newman, L. A. (2005). Breast cancer in sub-saharan africa: How does it relate to breast cancer in african-american women? *Cancer*, 103(8):1540–1550.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). Bayesian data analysis, (chapman & hall/crc texts in statistical science).
- Henson, D. E., Chu, K. C., and Levine, P. H. (2003). Histologic grade, stage, and survival in breast carcinoma: comparison of african american and caucasian women. *Cancer*, 98(5):908–917.

- Hussain, S. K., Altieri, A., Sundquist, J., and Hemminki, K. (2008). Influence of education level on breast cancer risk and survival in Sweden between 1990 and 2004. *International journal of cancer*, 122(1):165–169.
- Ikpatt, O., Ndoma-Egba, R., and Collan, Y. (2002). Influence of age and prognosis of breast cancer in Nigeria. *East African medical journal*, 79(12):651–657.
- Ikpatt, O. and Ndoma-Egba, R. (2003). Oestrogen and progesterone receptors in Nigerian breast cancer: relationship to tumour histopathology and survival of patients. *The Central African journal of medicine*, 49(11-12):122–126.
- Inc, S. (2014). Base sas® 9.4 procedures guide: statistical procedures. Cary: SAS Institute Inc.
- Kotsopoulos, J., Chen, W. Y., Gates, M. A., Tworoger, S. S., Hankinson, S. E., and Rosner, B. A. (2010). Risk factors for ductal and lobular breast cancer: results from the nurses' health study. *Breast Cancer Research*, 12(6):R106.
- Lesaffre, E. and Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.
- Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.
- Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, volume 698. John Wiley & Sons.
- Ojewusi, A. A. and Arulogun, O. S. (2016). Breast cancer knowledge and screening practices among female secondary schools teachers in an urban local government area, Ibadan, Nigeria. *Journal of Public Health and Epidemiology*, 8(5):72–81.
- Okobia, M. N., Bunker, C., Lee, L., Osime, U., and Uche, E. (2005). Case-control study of risk factors for breast cancer in Nigerian women: a pilot study. *East African medical journal*, 82(1).
- Oladimeji, K. E., Tsoka-Gwegweni, J. M., Igbodekwe, F. C., Twomey, M., Akolo, C., Balarabe, H. S., Atilola, O., Jegede, O., and Oladimeji, O. (2015). Knowledge and beliefs of breast self-examination and breast cancer among market women in Ibadan, south west, Nigeria. *PloS one*, 10(11):e0140904.
- Organization, W. H. (2013). *WHO report on the global tobacco epidemic, 2013: enforcing bans on tobacco advertising, promotion and sponsorship*. World Health Organization.
- Plummer, M., Best, N., Cowles, K., and Vines, K. (2010). coda: Output analysis and diagnostics for mcmc. R package version 0.14-2.
- Rakha, E. A., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V., Fox, S. B., Ichihara, S., Jacquemier, J., Lakhani, S. R., et al. (2010). Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, 12(4):207.
- Rosen, P. P., Lesser, M. L., Kinne, D. W., and Beattie, E. J. (1984). Breast carcinoma in women 35 years of age or younger. *Annals of surgery*, 199(2):133.
- Salameh, P., Waked, M., Khayat, G., and Dramaix, M. (2014). Bayesian and frequentist comparison for epidemiologists: A non mathematical application on logistic regressions. *The Open Epidemiology Journal*, 7(1).
- Simpson, J. F., Gray, R., Dressler, L. G., Cobau, C. D., Falkson, C. I., Gilchrist, K. W., Pandya, K. J., Page, D. L., and Robert, N. J. (2000). Prognostic value of histologic grade and proliferative activity in axillary node-positive breast cancer: results from the eastern cooperative oncology group companion study, est 4189. *Journal of Clinical Oncology*, 18(10):2059–2069.

- Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*, 100(18):10393–10398.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.
- Sundquist, M., Thorstenson, S., Brudin, L., and Nordenskjöld, B. (1999). Applying the nottingham prognostic index to a swedish breast cancer population. *Breast cancer research and treatment*, 53(1):1–8.
- Ursin, G., Bernstein, L., Lord, S. J., Karim, R., Deapen, D., Press, M. F., Daling, J. R., Norman, S. A., Liff, J. M., Marchbanks, P. A., et al. (2005). Reproductive factors and subtypes of breast cancer defined by hormone receptor and histology. *British journal of cancer*, 93(3):364.
- Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *The Journal of pathology*, 220(2):263–280.

References

- Abudu, E., Banjo, A., Izegbu, M., Agboola, A., Anunobi, C., & Musa, O. (2007). Malignant breast lesions at olabisi onabanjo university teaching hospital (oouth), sagamu-a histopathological review. *The Nigerian postgraduate medical journal*, 14(1), 57–59.
- Acquah, H. D.-G. (2013). Bayesian logistic regression modelling via markov chain monte carlo algorithm. *Journal of Social and Development Sciences*, 4(4), 193–197.
- Adebamowo, C., & Adekunle, O. (1999). Case-controlled study of the epidemiological risk factors for breast cancer in nigeria. *British Journal of Surgery*, 86(5), 665–668.
- Adebamowo, C. A., & Ajayi, O. (1999). Breast cancer in nigeria. *West African journal of medicine*, 19(3), 179–191.
- Adebamowo, C. A., Ogundiran, T. O., Adenipekun, A. A., Oyeseun, R. A., Campbell, O. B., Akang, E. U., Rotimi, C. N., & Olopade, O. I. (2003). Obesity and height in urban nigerian women with breast cancer. *Annals of epidemiology*, 13(6), 455–461.
- Adesunkanmi, A., & Agbakwuru, E. (2000). Benign breast disease at wesley guild hospital, ilesha, nigeria. *West African journal of medicine*, 20(2), 146–151.
- Afolayan, E. (2008). Cancer in north western region of nigeriaan update analysis of zaria cancer registry data. *Western Niger J Med Sci*, 1, 37–43.
- Afolayan, E., Ibrahim, O., & Ayilara, G. (2012). Cancer patterns in ilorin: An analysis of ilorin cancer registry statistics. *Tropical Journal of Health Sciences*, 19(1).
- Agboola, A., Musa, A., Wanangwa, N., Abdel-Fatah, T., Nolan, C., Ayoade, B., Oye-badejo, T., Banjo, A., Deji-Agboola, A., Rakha, E., et al. (2012). Molecular characteristics and prognostic features of breast cancer in nigerian compared with uk women. *Breast cancer research and treatment*, 135(2), 555–569.
- Agresti, A., & Kateri, M. (2011). *Categorical data analysis*. Springer.

- Akarolo-Anthony, S. N., Ogundiran, T. O., & Adebamowo, C. A. (2010). Emerging breast cancer epidemic: evidence from africa. *Breast Cancer Research*, 12(4), S8.
- Akinyemiju, T. F., Pisu, M., Waterbor, J. W., & Altekruse, S. F. (2015). Socioeconomic status and incidence of breast cancer by hormone receptor subtype. *SpringerPlus*, 4(1), 508.
- Albert, J., & Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, (pp. 747–759).
- Albert, J. H. (1996). *Bayesian computation using Minitab*. Duxbury Press.
- Alieldin, N. H., Abo-Elazm, O. M., Bilal, D., Salem, S. E., Gouda, E., Elmongy, M., & Ibrahim, A. S. (2014). Age at diagnosis in women with non-metastatic breast cancer: Is it related to prognosis? *Journal of the Egyptian National Cancer Institute*, 26(1), 23–30.
- Almeida, G., Almeida, L. A. L., Araujo, G. M. R., & Weller, M. (2015). Reproductive risk factors differ among breast cancer patients and controls in a public hospital of paraiba, northeast brazil. *Asian Pac J Cancer Prev*, 16(7), 2959–2965.
- Amadou, A., Ferrari, P., Muwonge, R., Moskal, A., Biessy, C., Romieu, I., & Hainaut, P. (2013). Overweight, obesity and risk of premenopausal breast cancer according to ethnicity: A systematic review and dose-response meta-analysis. *Obesity Reviews*, 14(8), 665–678.
- Anderson, B. O., & Jakesz, R. (2008). Breast cancer issues in developing countries: an overview of the breast health global initiative. *World journal of surgery*, 32(12), 2578–2585.
- Anderson, L. N., Cotterchio, M., Vieth, R., & Knight, J. A. (2010). Vitamin d and calcium intakes and breast cancer risk in pre-and postmenopausal women. *The American journal of clinical nutrition*, (pp. ajcn-28869).
- Anderson, W. F., Pfeiffer, R. M., Dores, G. M., & Sherman, M. E. (2006). Comparison of age distribution patterns for different histopathologic types of breast carcinoma. *Cancer Epidemiology and Prevention Biomarkers*, 15(10), 1899–1905.
- Andrieu, N., Easton, D. F., Chang-Claude, J., Rookus, M. A., Brohet, R., Cardis, E., Antoniou, A. C., Wagner, T., Simard, J., Evans, G., et al. (2006). Effect of chest x-rays on the risk of breast cancer among brca1/2 mutation carriers in the international brca1/2 carrier cohort study: a report from the embrace, genepso, gehobon, and ibccs collaborators group. *Journal of Clinical Oncology*, 24(21), 3361–3366.

- Antoniou, A., Pharoah, P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J., Loman, N., Olsson, H., Johannsson, O., Borg, Å., et al. (2003). Average risks of breast and ovarian cancer associated with *brca1* or *brca2* mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics*, 72(5), 1117–1130.
- Anyanwu, S. N. (2008). Temporal trends in breast cancer presentation in the third world. *Journal of Experimental & Clinical Cancer Research*, 27(1), 17.
- Anyikam, A., Nzegwu, M. A., Ozumba, B. C., Okoye, I., & Olusina, D. B. (2008). Benign breast lesions in eastern nigeria. *Saudi medical journal*, 29(2), 241–244.
- AO, P., Omodele, F., Oludara, M., NA, I., AI, I., & SBL, M. (2013). Prevalence and pattern of cancers among adults attending a tertiary health institution in lagos, nigeria. *Journal of Dental and Medical Sciences*, 6(3), 68–73.
- Arora, N., & Simmons, R. M. (2009). Malignant breast disease: Diagnosis and assessment. *General Surgery*, (pp. 1481–1494).
- Autier, P., Boniol, M., LaVecchia, C., Vatten, L., Gavin, A., Héry, C., & Heanue, M. (2010). Disparities in breast cancer mortality trends between 30 european countries: retrospective trend analysis of who mortality database. *Bmj*, 341, c3620.
- Baan, R., Straif, K., Grosse, Y., Secretan, B., El Ghissassi, F., Bouvard, V., Altieri, A., & Coglianò, V. (2007). Carcinogenicity of alcoholic beverages. *Lancet Oncology*, 8(4), 292.
- Banjo, A. (2004). Overview of breast and cervical cancers in nigeria: are there regional variations. In *Paper presentation at the International workshop on new trends in Management of breast and cervical cancers, Lagos, Nigeria*.
- Bayaga, A. (2010). Multinomial logistic regression: usage and application in risk analysis. *Journal of applied quantitative methods*, 5(2), 288–297.
- Bedrick, E. J., Christensen, R., & Johnson, W. (1997). Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician*, 51(3), 211–218.
- Bello, M. (2012). Awareness is the first step in battle against breast cancer.
- Best, D. S. A. T. N. (1996). Bugs 0.5* bayesian inference using gibbs sampling manual (version ii) david spiegelhalter andrew thomas nicky best wally gilks.
- Best, N., Cowles, M. K., & Vines, K. (1995). Coda* convergence diagnosis and output analysis software for gibbs sampling output version 0.30. *MRC Biostatistics Unit, Cambridge*, 52.

- Bhaskaran, K., Douglas, I., Forbes, H., dos Santos-Silva, I., Leon, D. A., & Smeeth, L. (2014). Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million uk adults. *The Lancet*, 384(9945), 755–765.
- Biganzoli, L., Wildiers, H., Oakman, C., Marotti, L., Loibl, S., Kunkler, I., Reed, M., Ciatto, S., Voogd, A. C., Brain, E., et al. (2012). Management of elderly patients with breast cancer: updated recommendations of the international society of geriatric oncology (siog) and european society of breast cancer specialists (eusoma). *The lancet oncology*, 13(4), e148–e160.
- Blamey, R., Hornmark-Stenstam, B., Ball, G., Blichert-Toft, M., Cataliotti, L., Fourquet, A., Gee, J., Holli, K., Jakesz, R., Kerin, M., et al. (2010). Oncopool—a european database for 16,944 cases of breast cancer. *European journal of cancer*, 46(1), 56–71.
- Bolstad, W. M., & Curran, J. M. (2007). *Introduction to Bayesian statistics*. John Wiley & Sons.
- Boos, D. D., et al. (2003). Introduction to the bootstrap world. *Statistical science*, 18(2), 168–174.
- Booth, J. G., & Sarkar, S. (1998). Monte carlo approximation of bootstrap variances. *The American Statistician*, 52(4), 354–357.
- Boyd, N. F., Dite, G. S., Stone, J., Gunasekara, A., English, D. R., McCredie, M. R., Giles, G. G., Trichler, D., Chiarelli, A., Yaffe, M. J., et al. (2002). Heritability of mammographic density, a risk factor for breast cancer. *New England Journal of Medicine*, 347(12), 886–894.
- Brandt, J., Garne, J. P., Tengrup, I., & Manjer, J. (2015). Age at diagnosis in relation to survival following breast cancer: a cohort study. *World journal of surgical oncology*, 13(1), 33.
- Bray, F., McCarron, P., & Parkin, D. M. (2004). The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Research*, 6(6), 229.
- Brooks, S., Gelman, A., Jones, G., & Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.
- Brooks, S. P., & Roberts, G. O. (1998). Convergence assessment techniques for markov chain monte carlo. *Statistics and Computing*, 8(4), 319–335.
- Brown, S. J., Donath, S., MacArthur, C., McDonald, E. A., & Krastev, A. H. (2010). Urinary incontinence in nulliparous women before and during pregnancy: prevalence, incidence, and associated risk factors. *International urogynecology journal*, 21(2), 193–202.

- Browne, W. J., Draper, D., et al. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian analysis*, 1(3), 473–514.
- Bryan, L., Conway, M., Keesmaat, T., McKenna, S., & Richardson, B. (2010). Strengthening sub-saharan african health systems: a practical approach. *McKinsey Quarterly*.
- Buchanan, N., Roland, K. B., Rodriguez, J. L., Miller, J. W., & Fairley, T. (2013). Opportunities for public health communication, intervention, and future research on breast cancer in younger women. *Journal of Women's Health*, 22(4), 293–298.
- Chan, M.-F., Dowsett, M., Folkard, E., Bingham, S., Wareham, N., Luben, R., Welch, A., & Khaw, K.-T. (2007). Usual physical activity and endogenous sex hormones in postmenopausal women: the european prospective investigation into cancer–norfolk population study. *Cancer Epidemiology and Prevention Biomarkers*, 16(5), 900–905.
- Chan, Y. H. (2005). Biostatistics 305. multinomial logistic regression. *Singapore medical journal*, 46(6), 259.
- Chen, H.-l., Zhou, M.-q., Tian, W., Meng, K.-x., & He, H.-f. (2016). Effect of age on breast cancer patient prognoses: A population-based study using the seer 18 database. *PloS one*, 11(10), e0165409.
- Chi, A. C., Day, T. A., & Neville, B. W. (2015). Oral cavity and oropharyngeal squamous cell carcinoma update. *CA: a cancer journal for clinicians*, 65(5), 401–421.
- Cibula, D., Gompel, A., Mueck, A., La Vecchia, C., Hannaford, P., Skouby, S., Zikan, M., & Dusek, L. (2010). Hormonal contraception and risk of cancer. *Human reproduction update*, 16(6), 631–650.
- Clèries, R., Ribes, J., Buxo, M., Ameijide, A., Marcos-Gragera, R., Galceran, J., Martínez, J. M., & Yasui, Y. (2012). Bayesian approach to predicting cancer incidence for an area without cancer registration by using cancer incidence data from nearby areas. *Statistics in medicine*, 31(10), 978–987.
- Cluze, C., Colonna, M., Remontet, L., Poncet, F., Sellier, E., Seigneurin, A., Delafosse, P., & Bossard, N. (2009). Analysis of the effect of age on the prognosis of breast cancer. *Breast cancer research and treatment*, 117(1), 121.
- Colditz, G. A., Hankinson, S. E., Hunter, D. J., Willett, W. C., Manson, J. E., Stampfer, M. J., Hennekens, C., Rosner, B., & Speizer, F. E. (1995). The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. *New England Journal of Medicine*, 332(24), 1589–1593.

- Colditz, G. A., Willett, W. C., Hunter, D. J., Stampfer, M. J., Manson, J. E., Hennekens, C. H., Rosner, B. A., & Speizer, F. E. (1993). Family history, age, and risk of breast cancer: prospective data from the nurses' health study. *Jama*, 270(3), 338–343.
- Coleman, M., Forman, D., Bryant, H., Butler, J., Rachet, B., Maringe, C., Nur, U., Tracey, E., Coory, M., Hatcher, J., et al. (2011). Cancer survival in australia, canada, denmark, norway, sweden, and the uk, 1995–2007 (the international cancer benchmarking partnership): an analysis of population-based cancer registry data. *The Lancet*, 377(9760), 127–138.
- Collaborators, M. W. S., et al. (2003). Breast cancer and hormone-replacement therapy in the million women study. *The Lancet*, 362(9382), 419–427.
- Congdon, P. (2005). *Bayesian models for categorical data*. John Wiley & Sons.
- Congdon, P. (2014). *Applied bayesian modelling*, vol. 595. John Wiley & Sons.
- Congdon, P. D. (2010). *Applied Bayesian hierarchical methods*. CRC Press.
- Coughlin, S. S., & Ekwueme, D. U. (2009). Breast cancer as a global health concern. *Cancer epidemiology*, 33(5), 315–318.
- Curado, M.-P., Edwards, B., Shin, H. R., Storm, H., Ferlay, J., Heanue, M., Boyle, P., et al. (2007). *Cancer incidence in five continents, Volume IX*. IARC Press, International Agency for Research on Cancer.
- Dalton, S. O., Düring, M., Ross, L., Carlsen, K., Mortensen, P. B., Lynch, J., & Johansen, C. (2006). The relation between socioeconomic and demographic factors and tumour stage in women diagnosed with breast cancer in denmark, 1983–1999. *British journal of cancer*, 95(5), 653.
- Dauda, A., Misauno, M., & Ojo, E. (2011). Histopathological types of breast cancer in gombe, north eastern nigeria: A seven-year review. *African journal of reproductive health*, 15(1).
- Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*, vol. 1. Cambridge university press.
- De Leeuw, J., Meijer, E., & Goldstein, H. (2008). *Handbook of multilevel analysis*.
- Deane, K. A., & Degner, L. F. (1998). Information needs, uncertainty, and anxiety in women who had a breast biopsy with benign outcome. *Cancer Nursing*, 21(2), 117–126.

- Dellaportas, P., & Smith, A. F. (1993). Bayesian inference for generalized linear and proportional hazards models via gibbs sampling. *Applied Statistics*, (pp. 443–459).
- DeSantis, C., Siegel, R., Bandi, P., & Jemal, A. (2011). Breast cancer statistics, 2011. *CA: a cancer journal for clinicians*, 61(6), 408–418.
- Díaz, C., & Batanero, C. (2016). ¿ cómo puede el método bayesiano contribuir a la investigación en psicología y educación? *Paradigma*, 27(2), 35–53.
- Diez-Roux, A. V. (2000). Multilevel analysis in public health research. *Annual review of public health*, 21(1), 171–192.
- Dite, G., Whittemore, A., Knight, J., John, E., Milne, R., Andrulis, I., Southey, M., McCredie, M., Giles, G., Miron, A., et al. (2010). Increased cancer risks for relatives of very early-onset breast cancer cases with and without brca1 and brca2 mutations. *British journal of cancer*, 103(7), 1103.
- Dunson, D. B., & Colombo, B. (2003). Bayesian modeling of markers of day-specific fertility. *Journal of the American Statistical Association*, 98(461), 28–37.
- Dvaladze, A. L. E. (2012). *Living with breast cancer: Experiences and Perceptions of women in Georgia*. University of Washington.
- Ebughe, G., Ekanem, I., Omoronyia, O., Nnoli, M., Nwagbara, V., Udosen, J., Umoh, M., & Ugbem, T. (2013a). Age specific incidence of breast cancer in calabar, nigeria.
- Ebughe, G., Ugare, G. U., Nnoli, M. A., Basse, I.-A., Nwagbara, V. J., Udosen, J., Omoronyia, O. E., Chukwuegbo, C. C., Ugbem, T. I., Omotoso, A. J., et al. (2013b). Histological type and tumour grade in nigerian breast cancer: Relationship to menarche, family history of breast cancer, parity, age at first birth, and age at menopause. *Iosrjournal*, 7(5), 58–63.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife annals of statistics 7: 1–26. *View Article PubMed/NCBI Google Scholar*.
- Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap: Monographs on statistics and applied probability, vol. 57. *New York and London: Chapman and Hall/CRC*.
- Ellis, I., et al. (2005). Pathology reporting of breast disease: a joint document incorporating the third edition of the nhs breast screening programmes guidelines for pathology reporting in breast cancer screening and the second edition of the royal college of pathologists minimum dataset for breast cancer histopathology. *NHS Cancer Screening Programmes, Royal College of Pathologists, London*.

- Ellison-Loschmann, L., McKenzie, F., Highnam, R., Cave, A., Walker, J., & Jeffreys, M. (2013). Age and ethnic differences in volumetric breast density in new zealand women: a cross-sectional study. *PloS one*, 8(7), e70217.
- Erkanli, A., Soyer, R., & Costello, E. J. (1999). Bayesian inference for prevalence in longitudinal two-phase studies. *Biometrics*, 55(4), 1145–1150.
- Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.
- Ferlay, J. (2004). (globocan 2002) cancer incidence, mortality and prevalence worldwide. iarc cancer base no. 5, version 2.0. <http://www-depdb.iarc.fr/globocan/GLOBOframe.htm>.
- Ferlay, J., Bray, F., Pisani, P., for Research on Cancer, W. I. A., et al. (2000). Iarc cancer epidemiology database, globocan 2000. *Cancer Incidence, Mortality and Prevalence Worldwide*.
- Ferlay, J., Shin, H.-R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: Globocan 2008. *International journal of cancer*, 127(12), 2893–2917.
- Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., & Bray, F. (2012). Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5).
- Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., & Bray, F. (2013a). Globocan 2012: Cancer incidence and mortality worldwide: Iarc cancer-base no. 11. Lyon, France: International agency for research on cancer.
- Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., & Bray, F. (2013b). Cancer incidence and mortality worldwide: Globocan 2012 v1. 0, iarc cancer base no. 11. *International Agency for Research on Cancer: Lyon, France*.
- Flenady, V., Koopmans, L., Middleton, P., Frøen, J. F., Smith, G. C., Gibbons, K., Coory, M., Gordon, A., Ellwood, D., McIntyre, H. D., et al. (2011). Major risk factors for stillbirth in high-income countries: a systematic review and meta-analysis. *The Lancet*, 377(9774), 1331–1340.
- for the Women's Health Initiative Investigators, W. G., et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal re-

- sults from the women's health initiative randomized controlled trial. *Jama*, 288(3), 321–333.
- Forae, G. D., Nwachokor, F. N., Igbe, A. P., Odokuma, E. I., Ijomone, E. A., et al. (2014). Benign breast diseases in warri southern nigeria: A spectrum of histopathological analysis. *Annals of Nigerian Medicine*, 8(1), 28.
- Forbes, J. F. (1997). The incidence of breast cancer: the global burden, public health considerations. In *Seminars in oncology*, vol. 24, (pp. S1–20).
- Forman, D., de Martel, C., Lacey, C. J., Soerjomataram, I., Lortet-Tieulent, J., Bruni, L., Vignat, J., Ferlay, J., Bray, F., Plummer, M., et al. (2012). Global burden of human papillomavirus and related diseases. *Vaccine*, 30, F12–F23.
- Forouzanfar, M. H., Foreman, K. J., Delossantos, A. M., Lozano, R., Lopez, A. D., Murray, C. J., & Naghavi, M. (2011). Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *The lancet*, 378(9801), 1461–1484.
- Fregene, A., & Newman, L. A. (2005). Breast cancer in sub-saharan africa: How does it relate to breast cancer in african-american women? *Cancer*, 103(8), 1540–1550.
- Frempong, M. A., Darko, E., & Addai, B. W. (2008). The use of carbohydrate antigen (ca) 15-3 as a tumor marker in detecting breast cancer. *Pakistan Journal of Biological Sciences*, 11(15), 1945–1948.
- Fritsch, G., Flamm, M., Hepner, D., Panisch, S., Seer, J., & Soennichsen, A. (2012). Abnormal pre-operative tests, pathologic findings of medical history, and their predictive value for perioperative complications. *Acta Anaesthesiologica Scandinavica*, 56(3), 339–350.
- Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453), 194–209.
- Gary-Webb, T. L., Baptiste-Roberts, K., Pham, L., Wesche-Thobaben, J., Patricio, J., Pi-Sunyer, F. X., Brown, A. F., Jones, L., & Brancati, F. L. (2010). Neighborhood and weight-related health behaviors in the look ahead (action for health in diabetes) study. *BMC Public Health*, 10(1), 312.
- Gathani, T., Ali, R., Balkwill, A., Green, J., Reeves, G., Beral, V., & Moser, K. (2014). Ethnic differences in breast cancer incidence in england are due to differences in known risk factors for the disease: prospective study. *British journal of cancer*, 110(1), 224.

- Gelfand, A. E., Sahu, S. K., & Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3), 479–488.
- Gelman, A. (2002). Prior distribution. *Encyclopedia of environmetrics*.
- Gelman, A., Carlin, J., Stern, H., & Savitz, D. (2004). *Bayesian data analysis*.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis*, (Chapman & Hall/CRC texts in statistical science).
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014a). *Bayesian data analysis*, vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel hierarchical models*, vol. 1. Cambridge University Press New York, NY, USA.
- Gelman, A., Hwang, J., & Vehtari, A. (2014b). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, (pp. 457–472).
- Geweke, J., et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, vol. 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.
- Godwins, E., David, D., & Akeem, J. (2011). Histopathologic analysis of benign breast diseases in Makurdi, north central Nigeria. *International Journal of Medicine and Medical Sciences*, 3(5), 125–128.
- Gordóvil-Merino, A., Guàrdia-Olmos, J., Peró-Cebollero, M., & de la Fuente-Solanas, E. I. (2010). Classical and Bayesian estimation in the logistic regression model applied to diagnosis of child attention deficit hyperactivity disorder. *Psychological reports*, 106(2), 519–533.
- Guray, M., & Sahin, A. A. (2006). Benign breast diseases: classification, diagnosis, and management. *The oncologist*, 11(5), 435–449.
- Harford, J. B., Otero, I. V., Anderson, B. O., Cazap, E., Gradishar, W. J., Gralow, J. R., Kane, G. M., Niëns, L. M., Porter, P. L., Reeler, A. V., et al. (2011). Problem solving for breast health care delivery in low and middle resource countries (Imcs): consensus statement from the breast health global initiative. *The Breast*, 20, S20–S29.

- Heidelberger, P., & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6), 1109–1144.
- Hirte, L., Nolte, E., Bain, C., & McKee, M. (2007). Breast cancer mortality in russia and ukraine 1963–2002: an age-period-cohort analysis. *International journal of epidemiology*, 36(4), 900–906.
- Hislop, T., & Elwood, J. (1981). Risk factors for benign breast disease: a 30-year cohort study. *Canadian Medical Association Journal*, 124(3), 283.
- Holmes, M. D., & Willett, W. C. (2004). Does diet affect breast cancer risk? *Breast Cancer Research*, 6(4), 170.
- Hosmer, D. W., & Lemeshow, S. (2000a). Interpretation of the fitted logistic regression model. *Applied Logistic Regression, Second Edition*, (pp. 47–90).
- Hosmer, D. W., & Lemeshow, S. (2000b). *Special topics*. Wiley Online Library.
- Hsieh, C.-C., Trichopoulos, D., Katsouyanni, K., & Yuasa, S. (1990). Age at menarche, age at menopause, height and obesity as risk factors for breast cancer: Associations and interactions in an international case-control study. *International journal of cancer*, 46(5), 796–800.
- Hulka, B. S., & Moorman, P. G. (2001). Breast cancer: hormones and other risk factors. *Maturitas*, 38(1), 103–113.
- Hussain, S. K., Altieri, A., Sundquist, J., & Hemminki, K. (2008). Influence of education level on breast cancer risk and survival in sweden between 1990 and 2004. *International journal of cancer*, 122(1), 165–169.
- Ibrahim, A., Salem, M., & Hassan, R. (2014). Outcome of young age at diagnosis of breast cancer in south egypt. *The Gulf journal of oncology*, 1(15), 76–83.
- Ibrahim, I. M., Iliyasu, Y., Mohammed, A. Z., et al. (2015). Histopathological review of breast tumors in kano, northern nigeria. *Sub-Saharan African Journal of Medicine*, 2(1), 47.
- Ikpat, O., Ndoma-Egba, R., & Collan, Y. (2002). Influence of age and prognosis of breast cancer in nigeria. *East African medical journal*, 79(12), 651–657.
- Institute, S., et al. (2008). *SAS/STAT 9.1 User's Guide the Reg Procedure:(Book Excerpt)*. SAS Institute.
- Institute Inc, S. (2008). *Sas/stat® 9.2. users guide*.

- Jack, X. Y., Sieuwerts, A. M., Zhang, Y., Martens, J. W., Smid, M., Klijn, J. G., Wang, Y., & Foekens, J. A. (2007). Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC cancer*, 7(1), 182.
- Jackman, S. (2000). Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo. *American Journal of Political Science*, (pp. 375–404).
- Jackson, S. P., & Bartek, J. (2009). The dna-damage response in human biology and disease. *Nature*, 461(7267), 1071.
- Jedy-Agba, E., Curado, M. P., Ogunbiyi, O., Oga, E., Fabowale, T., Igbinoba, F., Osu- bor, G., Otu, T., Kumai, H., Koechlin, A., et al. (2012a). Cancer incidence in nigeria: a report from population-based cancer registries. *Cancer epidemiology*, 36(5), e271–e278.
- Jedy-Agba, E., McCormack, V., Olaomi, O., Badejo, W., Yilkudi, M., Yawe, T., Ezeome, E., Salu, I., Miner, E., Anosike, I., et al. (2017). Determinants of stage at diagnosis of breast cancer in nigerian women: sociodemographic, breast cancer awareness, health care access and clinical factors. *Cancer causes & control*, (pp. 1–13).
- Jedy-Agba, E. E., Curado, M.-P., Oga, E., Samaila, M. O., Ezeome, E. R., Obiorah, C., Erinomo, O. O., Ima-obong, A. E., Uka, C., Mayun, A., et al. (2012b). The role of hospital-based cancer registries in low and middle income countries: the nigerian case study. *Cancer epidemiology*, 36(5), 430–435.
- Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2), 69–90.
- Jemal, A., Center, M. M., DeSantis, C., & Ward, E. M. (2010). Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiology and Prevention Biomarkers*, 19(8), 1893–1907.
- Jensen, L., Vestergaard, P., Hermann, A., Gram, J., Eiken, P., Abrahamsen, B., Brot, C., Kolthoff, N., Sørensen, O., Beck-Nielsen, H., et al. (2003). Hormone replacement therapy dissociates fat mass and bone mass, and tends to reduce weight gain in early postmenopausal women: A randomized controlled 5-year clinical trial of the danish osteoporosis prevention study. *Journal of bone and mineral research*, 18(2), 333–342.
- Jolly, T. A. (2015). External validity of a trial comprised of elderly patients with hormone receptor-positive breast cancer. *Breast Diseases*, 25(4), 307–308.

- Karayurt, Ö., Özmen, D., & Çetinkaya, A. Ç. (2008). Awareness of breast cancer risk factors and practice of breast self examination among high school students in turkey. *BMC Public Health*, 8(1), 359.
- Key, T. J., Verkasalo, P. K., & Banks, E. (2001). Epidemiology of breast cancer. *The lancet oncology*, 2(3), 133–140.
- Kiderlen, M., Walsh, P. M., Bastiaannet, E., Kelly, M. B., Audisio, R. A., Boelens, P. G., Brown, C., Dekkers, O. M., de Craen, A. J., van de Velde, C. J., et al. (2015). Treatment strategies and survival of older breast cancer patients—an international comparison between the netherlands and ireland. *PloS one*, 10(2), e0118074.
- Kotsopoulos, J., Chen, W. Y., Gates, M. A., Tworoger, S. S., Hankinson, S. E., & Rosner, B. A. (2010). Risk factors for ductal and lobular breast cancer: results from the nurses' health study. *Breast Cancer Research*, 12(6), R106.
- Krieger, N., Chen, J. T., & Waterman, P. D. (2010). Decline in us breast cancer rates after the women's health initiative: socioeconomic and racial/ethnic differentials. *American journal of public health*, 100(S1), S132–S139.
- Kumar, V., Abbas, A. K., Fausto, N., & Aster, J. C. (2014). *Robbins and Cotran pathologic basis of disease*. Elsevier Health Sciences.
- Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Blackwell Oxford.
- Landman, G. W., Kleefstra, N., van Hateren, K. J., Groenier, K. H., Gans, R. O., & Bilo, H. J. (2010). Metformin associated with lower cancer mortality in type 2 diabetes. *Diabetes care*, 33(2), 322–326.
- Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 619–678).
- Lehmann-Che, J., Hamy, A.-S., Porcher, R., Barritault, M., Bouhidel, F., Habuellelah, H., Leman-Detours, S., De Roquancourt, A., Cahen-Doidy, L., Bourstyn, E., et al. (2013). Molecular apocrine breast cancers are aggressive estrogen receptor negative tumors overexpressing either her2 or gcdfp15. *Breast Cancer Research*, 15(3), R37.
- Lesaffre, E., & Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.
- Little, R. J., & Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, (pp. 98–111).

- Loi, S., Sotiriou, C., Haibe-Kains, B., Lallemand, F., Conus, N. M., Piccart, M. J., Speed, T. P., & McArthur, G. A. (2009). Gene expression profiling identifies activated growth factor signaling in poor prognosis (luminal-b) estrogen receptor positive breast cancer. *BMC medical genomics*, 2(1), 37.
- Lønning, P. E., Chrisanthar, R., Staalesen, V., Knappskog, S., & Lillehaug, J. (2007). Adjuvant treatment: the contribution of expression microarrays. *Breast Cancer Research*, 9(2), S14.
- LOZANO-FERNÁNDEZ, L. M. (2008). Bayesian inference for binomial populations. bayesian estimation for child depression prevalence. *Advances and Applications in Statistics*, 9(1), 13–35.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4), 325–337.
- Mamun, A. A., Kinarivala, M., O’Callaghan, M. J., Williams, G. M., Najman, J. M., & Callaway, L. K. (2010). Associations of excess weight gain during pregnancy with long-term maternal overweight and obesity: evidence from 21 y postpartum follow-up. *The American journal of clinical nutrition*, 91(5), 1336–1341.
- Marrelec, G., Benali, H., Ciuciu, P., Pélégrini-Issac, M., & Poline, J.-B. (2003). Robust bayesian estimation of the hemodynamic response function in event-related bold fmri using basic physiological information. *Human Brain Mapping*, 19(1), 1–17.
- McPherson, K., Steel, C., & Dixon, J. (2000). Abc of breast diseases: breast cancer epidemiology, risk factors, and genetics. *BMJ: British Medical Journal*, 321(7261), 624.
- Mieog, J. S. D., de Kruijf, E. M., Bastiaannet, E., Kuppen, P. J., Sajet, A., de Craen, A. J., Smit, V. T., van de Velde, C. J., & Liefers, G.-J. (2012). Age determines the prognostic role of the cancer stem cell marker aldehyde dehydrogenase-1 in breast cancer. *BMC cancer*, 12(1), 42.
- Mila, A. L., & Michailides, T. J. (2006). Use of bayesian methods to improve prediction of panicle and shoot blight severity of pistachio in california. *Phytopathology*, 96(10), 1142–1147.
- Møller, H., Henson, K., Lüchtenborg, M., Broggio, J., Charman, J., Coupland, V. H., Davies, E., Jack, R. H., Sullivan, R., Vedsted, P., et al. (2016). Short-term breast cancer survival in relation to ethnicity, stage, grade and receptor status: national cohort study in england. *British journal of cancer*, 115(11), 1408–1415.

- Naylor, J. C., & Smith, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Applied Statistics*, (pp. 214–225).
- Nelder, J. A., & Baker, R. J. (1972). *Generalized linear models*. Wiley Online Library.
- Nelson, H. D., Zakher, B., Cantor, A., Fu, R., Griffin, J., O'meara, E. S., Buist, D. S., Kerlikowske, K., van Ravesteyn, N. T., Trentham-Dietz, A., et al. (2012). Risk factors for breast cancer for women aged 40 to 49 years: a systematic review and meta-analysis. *Annals of internal medicine*, 156(9), 635–648.
- Ngesa, O., Mwambi, H., & Achia, T. (2014). Bayesian spatial semi-parametric modeling of hiv variation in kenya. *PloS one*, 9(7), e103299.
- Ntekim, A., Nufu, F., & Campbell, O. (2009). Breast cancer in young women in ibadan, nigeria. *African health sciences*, 9(4).
- Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, vol. 698. John Wiley & Sons.
- Ochicha, O., Edino, S., Mohammed, A., & Amin, S. (2002). Benign breast lesions in kano. *Nigerian Journal of Surgical Research*, 4(1), 1–5.
- Odetunmibi, O., Adejumo, A., & OOM, S. (2013). Loglinear modelling of cancer patients cases in nigeria: An exploratory study approach. *Loglinear modelling of cancer patients cases in Nigeria: An exploratory study approach*, (pp. 1–5).
- O'hagan, A., Woodward, E., & Moodaley, L. (1990). Practical bayesian analysis of a simple logistic regression: predicting corneal transplants. *Statistics in Medicine*, 9(9), 1091–1101.
- Ojewusi, A. A., & Arulogun, O. S. (2016). Breast cancer knowledge and screening practices among female secondary schools teachers in an urban local government area, ibadan, nigeria. *Journal of Public Health and Epidemiology*, 8(5), 72–81.
- Ojo, O. B., Lougue, S., & Woldegerima, W. A. (2017). Bayesian generalized linear mixed modeling of tuberculosis using informative priors. *PloS one*, 12(3), e0172580.
- Okobia, M. N., Bunker, C., Lee, L., Osime, U., & Uche, E. (2005). Case-control study of risk factors for breast cancer in nigerian women: a pilot study. *East African medical journal*, 82(1).
- Okobia, M. N., Bunker, C. H., Okonofua, F. E., & Osime, U. (2006). Knowledge, attitude and practice of nigerian women towards breast cancer: a cross-sectional study. *World journal of surgical oncology*, 4(1), 1.

- Oladimeji, K. E., Tsoka-Gwegweni, J. M., Igbodekwe, F. C., Twomey, M., Akolo, C., Balarabe, H. S., Atilola, O., Jegede, O., & Oladimeji, O. (2015). Knowledge and beliefs of breast self-examination and breast cancer among market women in ibadan, south west, nigeria. *PloS one*, *10*(11), e0140904.
- OLUGBENGA, A. M., OLANREWAJU, M. J., & KAYODE, O. M. (2012). Profile of cancer patients attending tertiary health institutions in southwestern nigeria. *Asian J Pharm Clin Res*, *5*(1), 34–37.
- Oluwatosin, O. A., & Oladepo, O. (2006). Knowledge of breast cancer and its early detection measures among rural women in akinyele local government area, ibadan, nigeria. *BMC cancer*, *6*(1), 271.
- O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M., & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *49*(4), 517–542.
- Onland-Moret, N., Peeters, P., Van Gils, C., Clavel-Chapelon, F., Key, T., Tjønneland, A., Trichopoulou, A., Kaaks, R., Manjer, J., Panico, S., et al. (2005). Age at menarche in relation to adult height: the epic study. *American journal of epidemiology*, *162*(7), 623–632.
- Onyeanusi, C. G. (2015). *Effect of Family Support on Medication Adherence and Glycemic Control of Type 2 Diabetes Outpatients in A Tertiary Hospital In South-Eastern Nigeria*. Ph.D. thesis.
- Organization, W. H. (2013). *WHO report on the global tobacco epidemic, 2013: enforcing bans on tobacco advertising, promotion and sponsorship*. World Health Organization.
- ONEILL, P. D. (2002). A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain monte carlo methods. *Mathematical biosciences*, *180*(1), 103–114.
- Parkin, D. (2011a). 1. the fraction of cancer attributable to lifestyle and environmental factors in the uk in 2010: Introduction. *British Journal of Cancer*, *105*(Suppl 2), S2.
- Parkin, D. (2011b). 10. cancers attributable to exposure to hormones in the uk in 2010. *British journal of cancer*, *105*, S42–S48.
- Parkin, D. M., Bray, F., Ferlay, J., & Pisani, P. (2005). Global cancer statistics, 2002. *CA: a cancer journal for clinicians*, *55*(2), 74–108.

- Pettersson, A., Graff, R. E., Ursin, G., dos Santos Silva, I., McCormack, V., Baglietto, L., Vachon, C., Bakker, M. F., Giles, G. G., Chia, K. S., et al. (2014). Mammographic density phenotypes and risk of breast cancer: a meta-analysis. *JNCI: Journal of the National Cancer Institute*, 106(5).
- Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, 4(1), 12–35.
- Porter, P. (2008). westernizing women's risks? breast cancer in lower-income countries. *New England Journal of Medicine*, 358(3), 213–216.
- Pruitt, L., Mumuni, T., Raikhel, E., Ademola, A., Ogundiran, T., Adenipekun, A., Morhason-Bello, I., Ojengbede, O. A., & Olopade, O. I. (2015). Social barriers to diagnosis and treatment of breast cancer in patients presenting at a teaching hospital in ibadan, nigeria. *Global public health*, 10(3), 331–344.
- Pudrovska, T., & Anikputa, B. (2012). The role of early-life socioeconomic status in breast cancer incidence and mortality: unraveling life course mechanisms. *Journal of aging and health*, 24(2), 323–344.
- Raftery, A., & Lewis, S. (1992). How many iterations in the gibbs sampler? in bayesian statistics 4.(eds jm bernardo, jo berger, ap dawid and afm smith.) pp. 763–773.
- Rakha, E. A., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V., Fox, S. B., Ichihara, S., Jacquemier, J., Lakhani, S. R., et al. (2010). Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, 12(4), 207.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, vol. 1. Sage.
- Reeves, G. K., Beral, V., Green, J., Gathani, T., Bull, D., Collaborators, M. W. S., et al. (2006). Hormonal therapy for menopause and breast-cancer risk by histological type: a cohort study and meta-analysis. *The lancet oncology*, 7(11), 910–918.
- Reeves, G. K., Pirie, K., Beral, V., Green, J., Spencer, E., & Bull, D. (2007). Cancer incidence and mortality in relation to body mass index in the million women study: cohort study. *Bmj*, 335(7630), 1134.
- Rekkas, M. (2009). Approximate inference for the multinomial logit model. *Statistics & Probability Letters*, 79(2), 237–242.

- Ries, L. A., Harkins, D., Krapcho, M., Mariotto, A., Miller, B. A., Feuer, E. J., Clegg, L. X., Eisner, M., Horner, M.-J., Howlader, N., et al. (2006). Seer cancer statistics review, 1975-2003.
- Ritte, R., Lukanova, A., Berrino, F., Dossus, L., Tjønneland, A., Olsen, A., Overvad, T. F., Overvad, K., Clavel-Chapelon, F., Fournier, A., et al. (2012). Adiposity, hormone replacement therapy use and breast cancer risk by age and hormone receptor status: a large prospective cohort study. *Breast cancer research*, 14(3), R76.
- Ronckers, C. M., Erdmann, C. A., & Land, C. E. (2004). Radiation and breast cancer: a review of current evidence. *Breast Cancer Research*, 7(1), 21.
- Rubinstein, R. Y., & Kroese, D. P. (2016). *Simulation and the Monte Carlo method*, vol. 10. John Wiley & Sons.
- Rust, K. F., & Rao, J. (1996). Variance estimation for complex surveys using replication techniques. *Statistical methods in medical research*, 5(3), 283–310.
- Salameh, P., Waked, M., Khayat, G., & Dramaix, M. (2014). Bayesian and frequentist comparison for epidemiologists: A non mathematical application on logistic regressions. *The Open Epidemiology Journal*, 7(1).
- Saldova, R., Asadi Shehni, A., Haakensen, V. D., Steinfeld, I., Hilliard, M., Kifer, I., Helland, Å., Yakhini, Z., Børresen-Dale, A.-L., & Rudd, P. M. (2014). Association of n-glycosylation with breast carcinoma and systemic features using high-resolution quantitative uplc. *Journal of proteome research*, 13(5), 2314–2327.
- Schonberg, M. A., Marcantonio, E. R., Li, D., Silliman, R. A., Ngo, L., & McCarthy, E. P. (2010). Breast cancer among the oldest old: tumor characteristics, treatment choices, and survival. *Journal of Clinical Oncology*, 28(12), 2038–2045.
- Shaw, T. E., Currie, G. P., Koudelka, C. W., & Simpson, E. L. (2011). Eczema prevalence in the united states: data from the 2003 national survey of children's health. *Journal of Investigative Dermatology*, 131(1), 67–73.
- Shimizu, H., Ross, R., Bernstein, L., Yatani, R., Henderson, B., & Mack, T. (1991). Cancers of the prostate and breast among japanese and white immigrants in los angeles county. *British journal of cancer*, 63(6), 963.
- Shulman, L. N., Willett, W., Sievers, A., & Knaul, F. M. (2010). Breast cancer in developing countries: opportunities for improved survival. *Journal of oncology*, 2010.

- Siegel, R. L., Miller, K. D., Fedewa, S. A., Ahnen, D. J., Meester, R. G., Barzi, A., & Jemal, A. (2017). Colorectal cancer statistics, 2017. *CA: a cancer journal for clinicians*, 67(3), 177–193.
- Siegel, R. L., Miller, K. D., & Jemal, A. (2015). Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1), 5–29.
- Simkin-Silverman, L. R., & Wing, R. R. (2000). Weight gain during menopause: is it inevitable or can it be prevented? *Postgraduate medicine*, 108(3), 47–56.
- Simpson, J. F., Gray, R., Dressler, L. G., Cobau, C. D., Falkson, C. I., Gilchrist, K. W., Pandya, K. J., Page, D. L., & Robert, N. J. (2000). Prognostic value of histologic grade and proliferative activity in axillary node-positive breast cancer: results from the eastern cooperative oncology group companion study, est 4189. *Journal of Clinical Oncology*, 18(10), 2059–2069.
- Sinharay, S. (2003). Assessing convergence of the markov chain monte carlo algorithms: A review. *ETS Research Report Series*, 2003(1).
- Smith, B. J., et al. (2007). boa: an r package for mcmc output convergence assessment and posterior inference. *Journal of Statistical Software*, 21(11), 1–37.
- Sotiriou, C., Neo, S.-Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., & Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*, 100(18), 10393–10398.
- Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). Winbugs user manual.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Stewart, B. W., Kleihues, P., et al. (2003). *World cancer report*, vol. 57. IARC press Lyon.
- Sundquist, M., Thorstenson, S., Brudin, L., & Nordenskjöld, B. (1999). Applying the nottingham prognostic index to a swedish breast cancer population. *Breast cancer research and treatment*, 53(1), 1–8.
- Sylla, B. S., & Wild, C. P. (2012). A million africans a year dying from cancer by 2030: what can cancer research and control offer to the continent? *International journal of cancer*, 130(2), 245–250.

- Taib, N. A., Yip, C. H., & Low, W. Y. (2014). A grounded explanation of why women present with advanced breast cancer. *World journal of surgery, 38*(7), 1676–1684.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association, 82*(398), 528–540.
- Tavani, A., Braga, C., La Vecchia, C., Negri, E., Russo, A., & Franceschi, S. (1997). Attributable risks for breast cancer in Italy: education, family history and reproductive and hormonal factors. *International journal of cancer, 70*(2), 159–163.
- Thompson, D. J., Leach, M. O., Kwan-Lim, G., Gayther, S. A., Ramus, S. J., Warsi, I., Lennard, F., Khazen, M., Bryant, E., Reed, S., et al. (2009). Assessing the usefulness of a novel MRI-based breast density estimation algorithm in a cohort of women at high genetic risk of breast cancer: the UK MARIAS study. *Breast Cancer Research, 11*(6), R80.
- Thrift, A. P. (2016). The epidemic of oesophageal carcinoma: Where are we now? *Cancer epidemiology, 41*, 88–95.
- Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., & Jemal, A. (2015). Global cancer statistics, 2012. *CA: a cancer journal for clinicians, 65*(2), 87–108.
- Townsley, C. A., Selby, R., & Siu, L. L. (2005). Systematic review of barriers to the recruitment of older patients with cancer onto clinical trials. *Journal of Clinical Oncology, 23*(13), 3112–3124.
- Ursin, G., Bernstein, L., Lord, S. J., Karim, R., Deapen, D., Press, M. F., Daling, J. R., Norman, S. A., Liff, J. M., Marchbanks, P. A., et al. (2005). Reproductive factors and subtypes of breast cancer defined by hormone receptor and histology. *British journal of cancer, 93*(3), 364.
- Uwaezuoke, S. C., & Udoe, E. P. (2014). Benign breast lesions in Bayelsa state, Niger Delta Nigeria: a 5 year multicentre histopathological audit. *The Pan African medical journal, 19*.
- Vainshtein, J. (2008). Disparities in breast cancer incidence across racial/ethnic strata and socioeconomic status: a systematic review. *Journal of the National Medical Association, 100*(7), 833–839.
- Van Den Brandt, P. A., Spiegelman, D., Yaun, S.-S., Adami, H.-O., Beeson, L., Folsom, A. R., Fraser, G., Goldbohm, R. A., Graham, S., Kushi, L., et al. (2000). Pooled analysis of prospective cohort studies on height, weight, and breast cancer risk. *American journal of epidemiology, 152*(6), 514–527.

- Vineis, P., & Wild, C. P. (2014). Global cancer patterns: causes and prevention. *The Lancet*, 383(9916), 549–557.
- Weigelt, B., Baehner, F. L., & Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *The Journal of pathology*, 220(2), 263–280.
- Wilson, C., Tobin, S., & Young, R. (2004). The exploding worldwide cancer burden: the impact of cancer on women. *International Journal of Gynecological Cancer*, 14(1), 1–11.
- Wong, R. S. Y., & Ismail, N. A. (2016). An application of bayesian approach in modeling risk of death in an intensive care unit. *PloS one*, 11(3), e0151949.
- Yancik, R., Wesley, M. N., Ries, L. A., Havlik, R. J., Edwards, B. K., & Yates, J. W. (2001). Effect of age and comorbidity in postmenopausal breast cancer patients aged 55 years and older. *Jama*, 285(7), 885–892.
- Yip, C., & Taib, N. (2014). Breast health in developing countries. *Climacteric*, 17(sup2), 54–59.
- Youlden, D. R., Cramb, S. M., Yip, C. H., & Baade, P. D. (2014). Incidence and mortality of female breast cancer in the asia-pacific region. *Cancer biology & medicine*, 11(2), 101.
- Yu, L., & Wang, L. (2011). Bayesian and classical estimation of mixed logit model for simulated experimental data. In *Service Operations, Logistics, and Informatics (SOLI), 2011 IEEE International Conference on*, (pp. 255–258). IEEE.
- Yüksel, S., Ugras, G. A., Çavdar, I., Bozdogan, A., Gürdal, S. Ö., Akyolcu, N., Esencan, E., Saraçoğlu, G. V., & Özmen, V. (2017). A risk assessment comparison of breast cancer and factors affected to risk perception of women in turkey: A cross-sectional study. *Iranian Journal of Public Health*, 46(3), 308.
- Yusufu, L., Odigie, V., & Mohammed, A. (2003). Breast masses in zaria, nigeria. *Annals of African Medicine*, 2(1), 13–16.
- Zare, N., Haem, E., Lankarani, K. B., Heydari, S. T., & Barooti, E. (2013). Breast cancer risk factors in a defined population: weighted logistic regression approach for rare events. *Journal of breast cancer*, 16(2), 214–219.
- Zellner, A., & Rossi, P. E. (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, 25(3), 365–393.

- Zhao, Y., Staudenmayer, J., Coull, B. A., & Wand, M. P. (2006). General design bayesian generalized linear mixed models. *Statistical Science*, (pp. 35–51).
- Zheng, W., McLerran, D. F., Rolland, B., Zhang, X., Inoue, M., Matsuo, K., He, J., Gupta, P. C., Ramadas, K., Tsugane, S., et al. (2011). Association between body-mass index and risk of death in more than 1 million asians. *New England Journal of Medicine*, 364(8), 719–729.
- Ziegler, R. G., Hoover, R. N., Pike, M. C., Hildesheim, A., Nomura, A. M., West, D. W., Wu-Williams, A. H., Kolonel, L. N., Horn-Ross, P. L., Rosenthal, J. F., et al. (1993). Migration patterns and breast cancer risk in asian-american women. *Journal of the National Cancer Institute*, 85(22), 1819–1827.
- Zuur, G., Garthwaite, P. H., & Fryer, R. J. (2002). Practical use of mcmc methods: lessons from a case study. *Biometrical Journal*, 44(4), 433–455.

Appendix A

Appendix A

A.1 Diagnostic plots for the fixed and random effects for Bayesian multilevel model with blocking

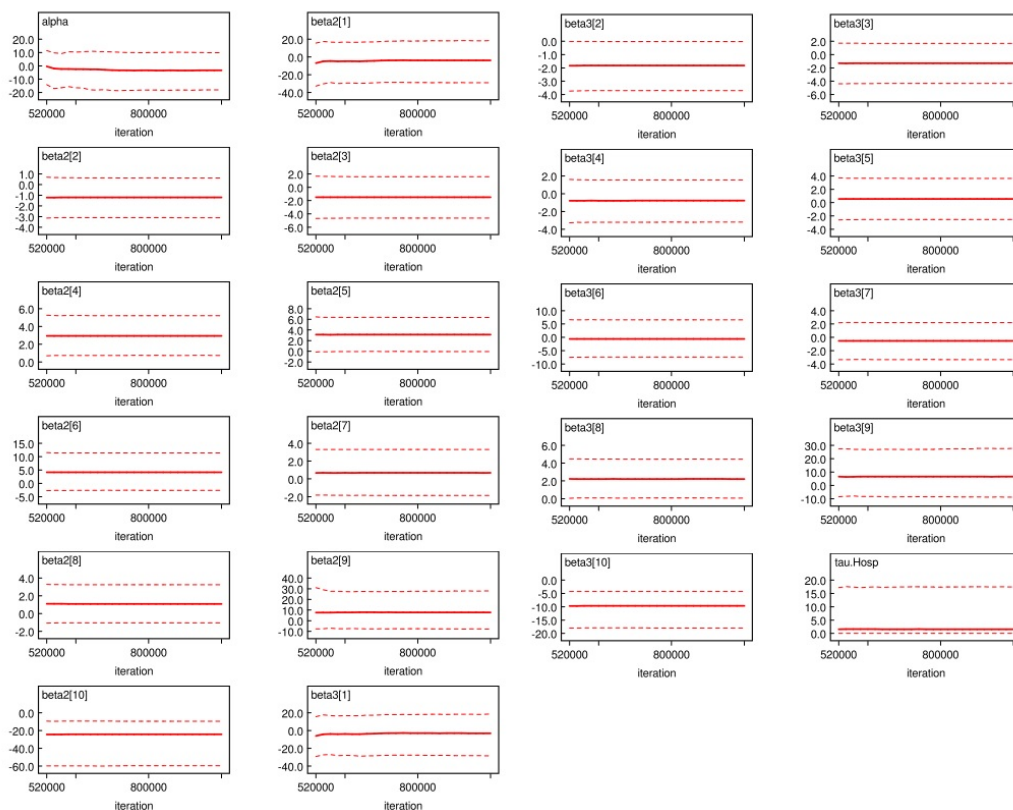


Figure 5.1. WinBUGS' output of Gelman Rubin Statistic for some independent variable

Fig 5.2 showed the posterior distribution of the random effect part of the model with blocking, based on the analysis of 1500000 iterations, 500000 burn in period, and with thin=40 in MCMCglmm.

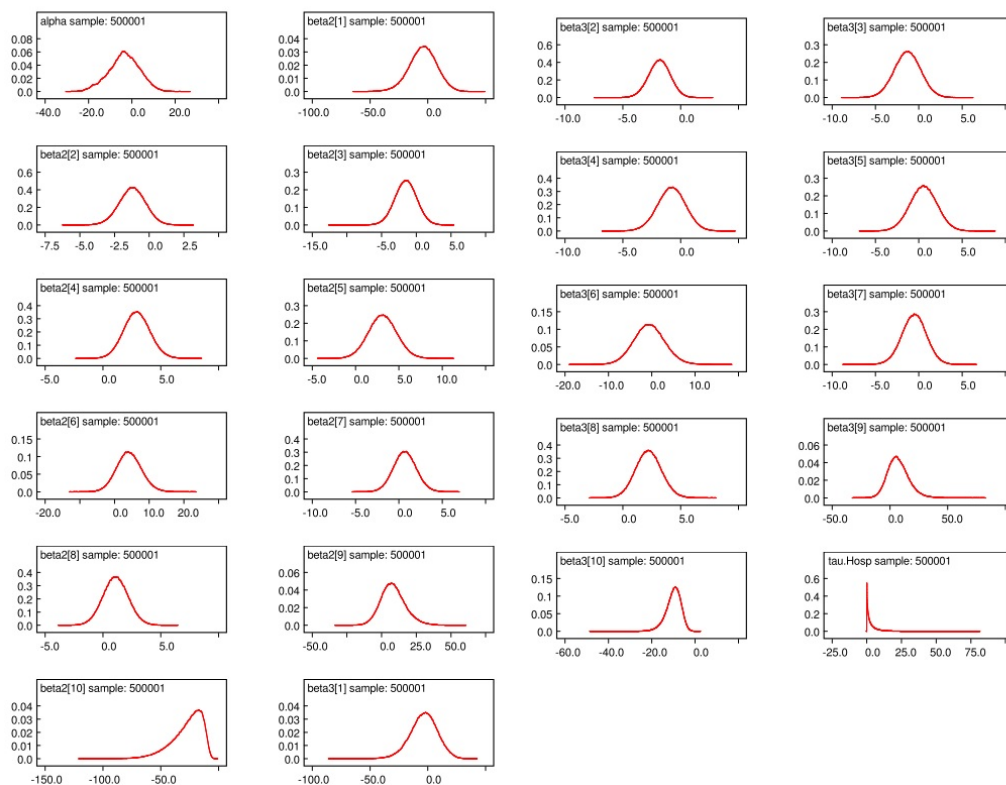


Figure 5.2. WinBUGS' output of posterior density for some independent variable

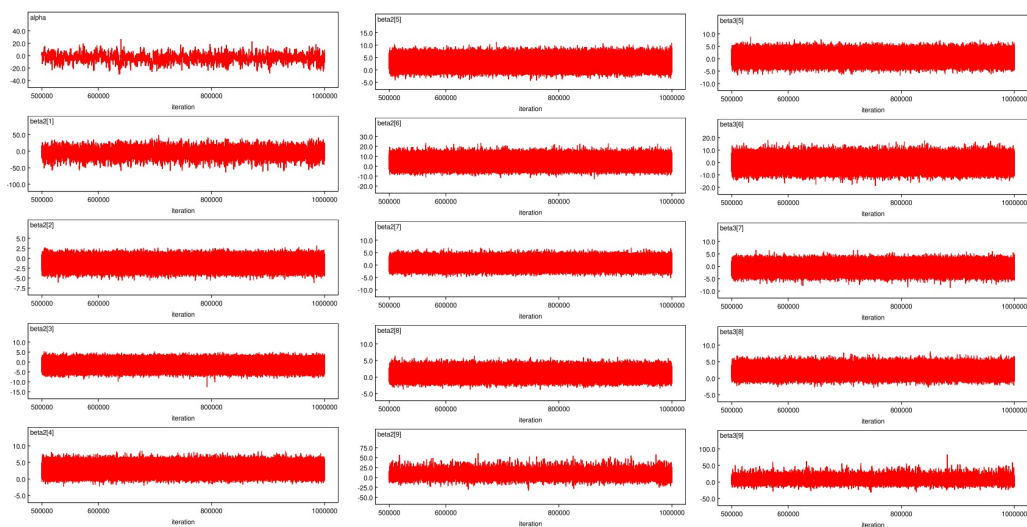


Figure 5.3. WinBUGS' output of time series for some independent variable

Appendix B

B.1 Diagnostic plots for the fixed and random effects for Bayesian multilevel multinomial model without blocking

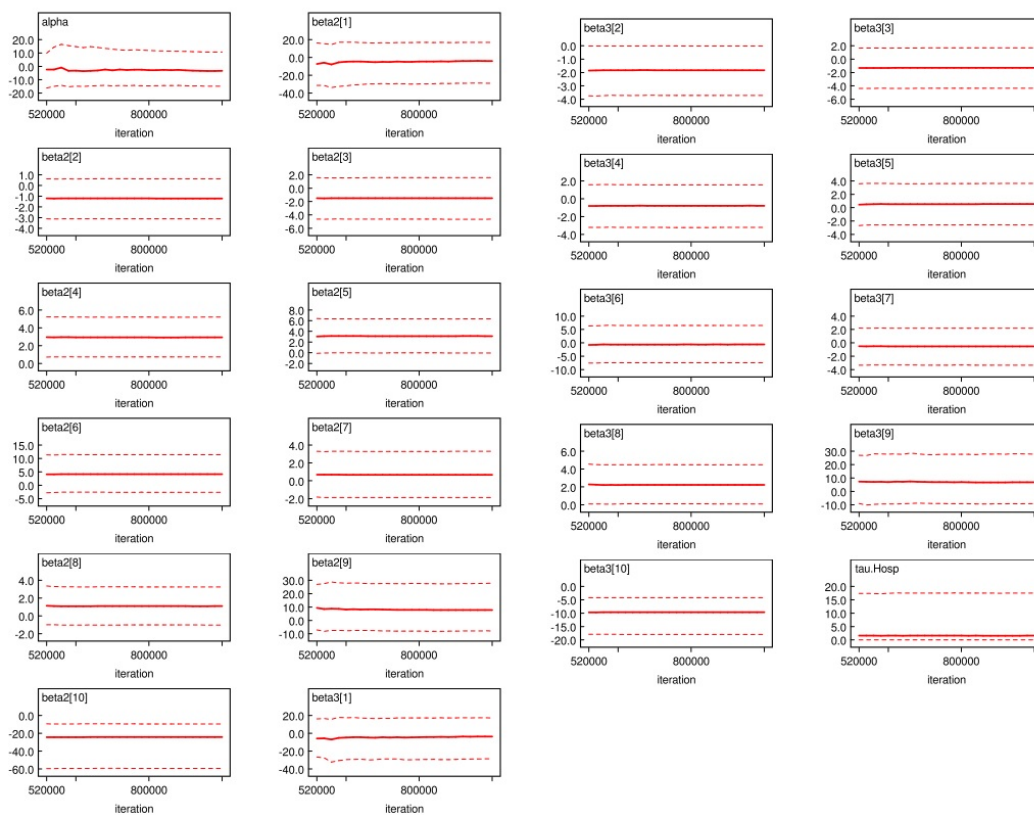


Figure 5.4. WinBUGS' output of Gelman Rubin Statistic for some independent variable

Fig 5.5 provides a graphical representation of the posterior density estimate for each parameter. This plot indicate normality for the posterior distribution of each parameter.

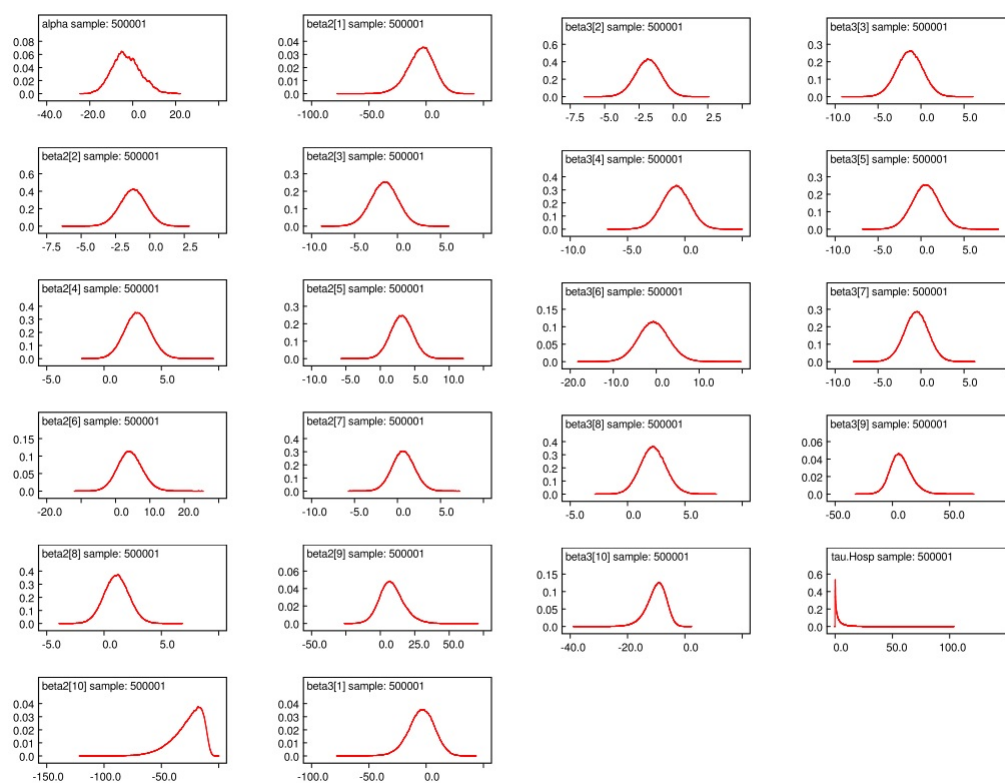


Figure 5.5. WinBUGS' output of posterior density for some independent variable

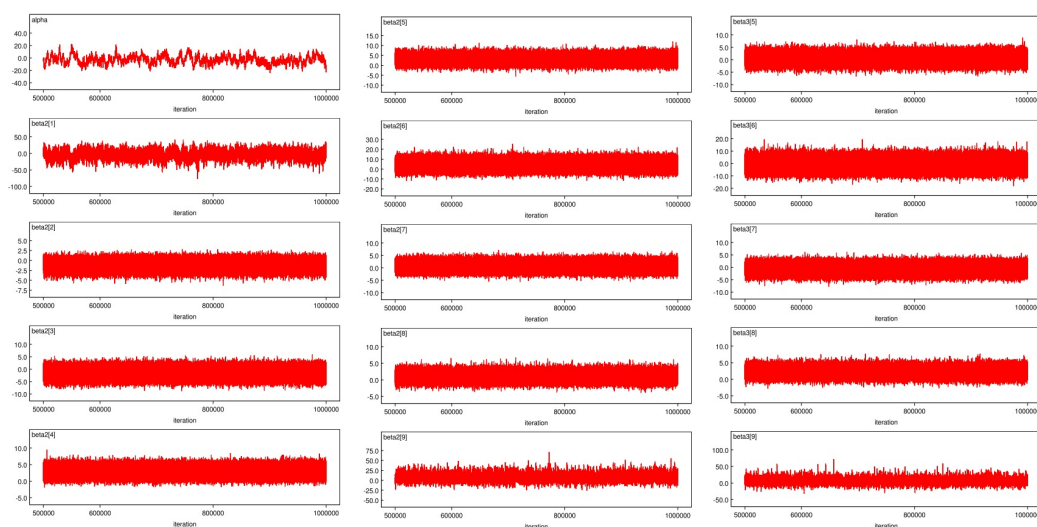


Figure 5.6. WinBUGS' output of time series for some independent variable

Fig 5.6 shows the time series plots and the plots suggest stability which is an indication that the Markov chain have converged.

Appendix C

C Diagnostic plots for Bayesian multilevel logistic regression model

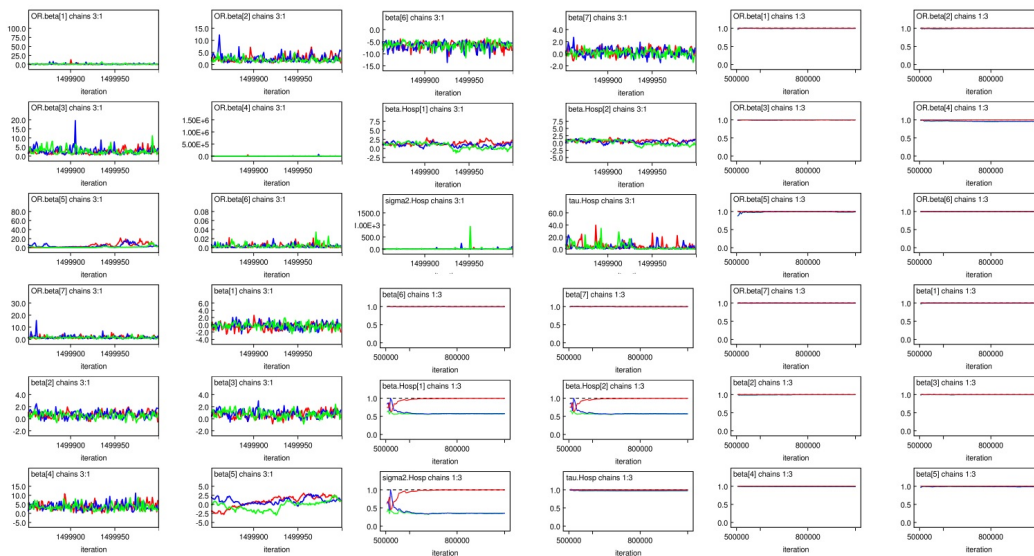


Figure 5.7. WinBUGS' output of Gelman Rubin Statistic

Fig 5.7 illustrates the convergence of the Bayesian approach with non-informative prior using Gelman-Rubin diagnostic test. The algorithm converged after 1,500, 000 iterations. The output of Gelman-Rubin convergence diagnostic test displays the red lines representing the \hat{R} . The graph shows that all the $\hat{R} \rightarrow 1$. Also, the blue and green lines which represent the within sample variance and the pooled posterior variance, are stationary. Thus, the Gelman-Rubin Convergence Diagnostic test suggests that the algorithm converges.

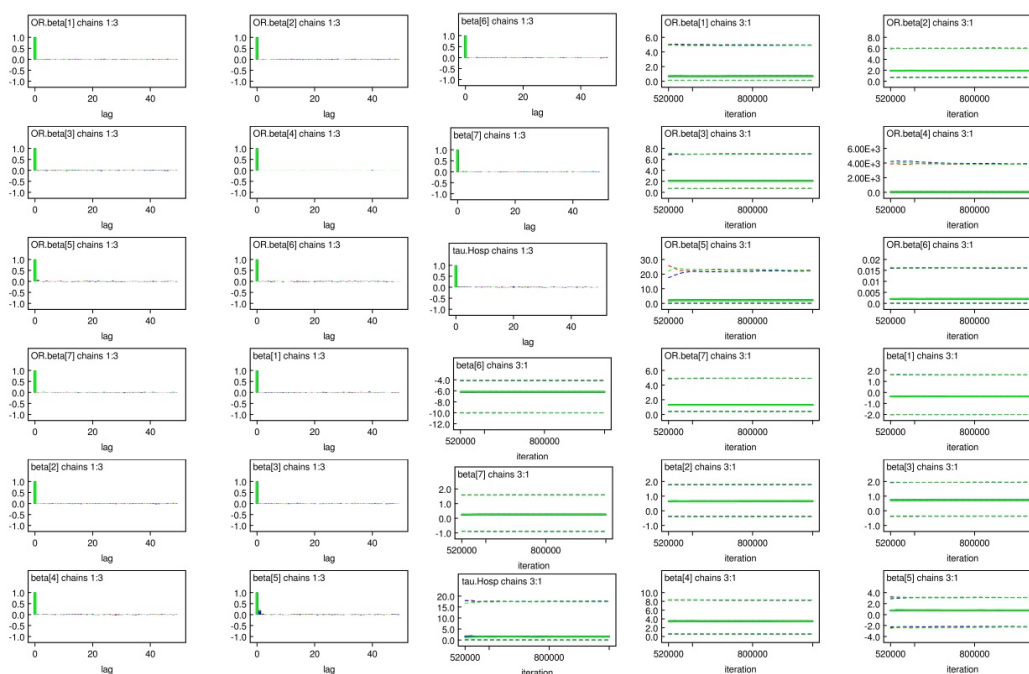


Figure 5.8. WinBUGS' output autocorrelation

Fig 5.8 illustrates the convergence of the Bayesian approach with non-informative prior. The algorithm converged after 1,500,000 iterations. We took a lag of 40 in order to remove the autocorrelation which requires an iterations up to 1,500,000 iterations and the first 500,000 iterations removed to cater for the burn-in period.

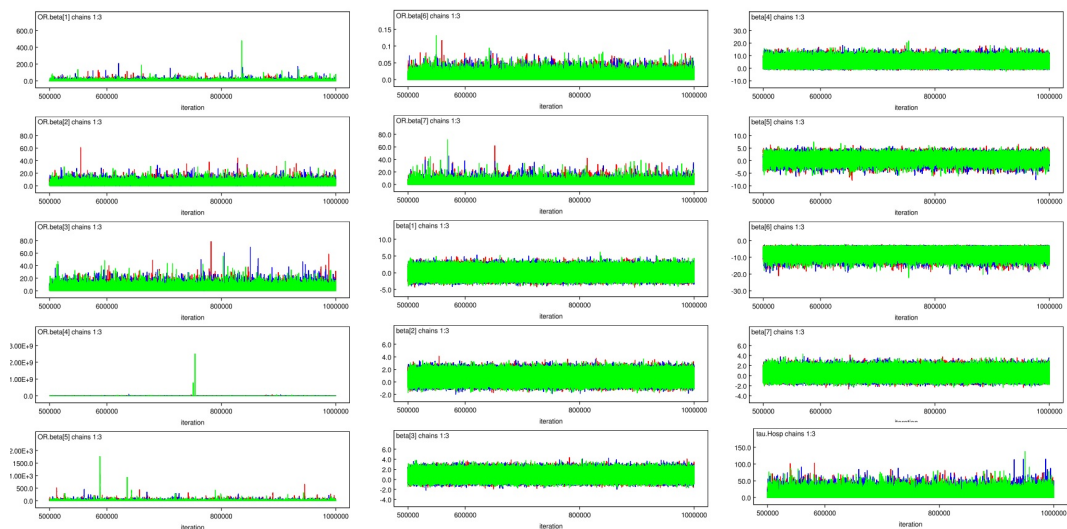


Figure 5.9. WinBUGS' output autocorrelation

The time series plot in Fig 5.9 suggests that the chain is wandering through the same region of the parameter space and has found the stationarity.