STATISTICAL TECHNIQUES FOR THE ANALYSIS OF REPEATED MEASURES: AN APPLICATION TO DIABETES MELLITUS DATA

BY

AUGUSTINE NARKORLI KWESI SIAKWA

(10508664)

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF MPHIL STATISTICS DEGREE

JUNE, 2018

DECLARATION

I hereby declare that this thesis is the result of my own research work and that no part of it has been presented for another degree in this university or elsewhere.

.....

Date..28/07/2017.....

SIAKWA AUGUSTINE NARKORLI KWESI

(Candidate)

I hereby declare that the preparation and presentation of this thesis was supervised in accordance with the guidelines on the supervision of thesis laid down by the University of Ghana.

.....	.....
.....	.....
Dr. Samuel Iddi	**Dr. Louis Asiedu**
(PRINCIPAL SUPERVISOR)	(CO – SUPERVISOR)
Date.....	**Date**.....

DEDICATION

This study is dedicated to my late Dad. To my siblings and mom, you guys have made me what I am today and I am really grateful for all that you have been doing for me. God richly bless all these wonderful people. Amen!

ACKNOWLEDGEMENT

This thesis is the culmination of the effort of many individuals to whom I owe my most sincere gratitude. My first thanks goes to my supervisors: Dr. Samuel Iddi who in spite of his busy schedules manage to see this work comes to a successful end.

My sincere gratitude is extended to Dr. Louis Asiedu who is a co- supervisor of this work.

I would like to extend my thanks to staff of the Diabetes Centre of the Maamobi Polyclinic in the greater Accra region for their help in the collection of the data for the study.

My entire course mates deserve the best recognition. Clarice, Umar, Mawuli, Sarah, Dela, Atsu, Key, Israel, Daniel, Aminu, Julius, Flex, Alfred and DD; you guys have been a great companion throughout the study period.

Finally, I would like to thank my family for their love, trust and prayers. Without them this accomplishment would have been unachievable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The focus of the research is to model some risk factors called covariates that influence the blood sugar level of diabetes patients using Generalized Linear Mixed Models (GLMM) and Generalized Estimating Equations (GEE). Both primary and secondary data was obtained from the Diabetes Centre of Maamobi Polyclinic in Accra. Data was obtained on 155 patients for four consecutive times they visit the hospital for review. Analysis of the data was done using SPSS version 20 and R. The findings revealed that the covariates (Age, Weight, Systolic blood pressure, Family history) were significant in both the GLMM and GEE models. A conclusion was drawn to the effect that since factors influencing the rise and fall of the sugar content in the blood of diabetes have been established, health providers should note these and lay emphases on them in controlling the epidemic (diabetes). Health providers must also be more vigilant when controlling high or low systolic and diastolic blood pressure of patients. Patients must be educated regularly on these factors so they can control their sugar levels.

CHAPTER ONE

INTRODUCTION

This chapter gives a general overview of the study by providing a general background of the research, problem statement, research objectives, scope of the study, significant of the research, and limitations. The chapter ends by presenting the layout of the thesis.

1.1. Background of the study

According to World Health Organization (WHO) diabetes is defined as a genuine, ceaseless malady that happens either when the pancreas does not deliver enough insulin (a hormone that directs glucose, or glucose), or when the body cannot successfully utilize the insulin it produces. It is among the leading cause of death worldwide (WHO, 1999). Diabetes is a worldwide burden with an expected 422 million grown-ups living with diabetes universally (WHO, 2016). WHO in 2009 assessed that 8% of African populace over 25 years have diabetes and about 820,000 Ghanaians are at risk of developing diabetes by 2035 (Guariguata et al. 2014).

There are three major types of diabetes; type '1', type '2', and gestational diabetes. The body of diabetic patients cannot make or use insulin in all the three types. This study focuses on type '2', which is common among all the diabetes cases. The type '2' formally known as grown-up onset or non-insulin-dependent diabetes can start at any age yet obvious amid adulthood. Among all the types, about 90-95 out of 100 diabetes patients are type '2' (Centers for Disease Control and Prevention, 2011). Type '2' can be classified as insulin resistance where the body is not able to

use the insulin produced or insulin deficiency where the pancreas makes less insulin needed by the body.

There are several risk factors associated with type ‘2’ diabetes, which include:

- Family history of diabetes
- Overweight
- Unhealthy diet
- Physical inactivity
- Increasing age
- High blood pressure
- Ethnicity
- Impaired glucose tolerance (IGT)
- Gestational diabetes history
- Poor nutrition during pregnancy

A common study of diabetes is a longitudinal type where people are studied for a time period. It varies from cross-sectional investigation in which diverse people with similar qualities are looked at. There is a dependency in the response variable made on the same person in a longitudinal study. The dependency accordingly should be considered in investigating longitudinal information to abstain from misleading inferences (Rahman and Islam, 2007). Utilization of Generalized Linear Mixed Models (GLMM) and Generalized Estimating Equation (GEE) are used to scrutinize information obtained from enlisted diabetes patients at the Maamobi Polyclinic in greater Accra region of Ghana. The information contains both time-differing and time-stationary covariates.

1.2. Problem statement

According to Peer, Kengne, Motala, and Mbanya (2014), Africa Region (AFR) where diabetes was not common has now witnessed an increase in the state of the condition. Type '2' diabetes incidence is higher within the elderly, 20-79 years. The percentage prevalence among this age group was about 4.9% with most of patients less than 60 years of age. The highest percentage is those in the age group between 40-59years with 43.2%. Undiagnosed diabetes cases, pose an increased risk of complications for the people living with diabetes without knowing it. Delayed diabetes diagnosis can increase risk of complications such as Kidney damage, heart diseases, blindness, stroke, neural damage leading to amputations and an overall reduced life expectancy.

Two diabetes centres have been lunched in Accra at the Maamobi Polyclinic and the Ga South Hospital with the objectives to increase diabetes awareness, earlier diagnosis, and access to treatment.

This research, therefore, seeks to model some risk factors herewith referred as covariates that influence the risk of developing diabetes using GLMM and GEE. The two models are chosen because the responses are binary (i.e. either one has diabetes or not at a particular time of observation) and there is a correlation in the data due to repeated measure overtime taken for each patient.

1.3. Objectives of the study

The main objectives of this research are to

1. Identify the most appropriate GLMM and GEE models that best fit longitudinal diabetes mellitus data

2. Study the effect of risk factors on the conditional distribution of the probability of having diabetes.

3. Obtain a model to predict disease probabilities based on observed risk factors in the longitudinal models.

1.4. Scope of the study

The study covers selected diabetes cases recorded at the Maamobi Polyclinic in Accra. It used monthly secondary data of diabetes patients covering four months of a visit to the hospital. The glucose level of a patient is recorded any time he or she visits the hospital to determine whether the glucose level is going up or down. The study also uses the risk factor information on patients for four consecutive times they visit the hospital for review.

1.5. Significance of the study

This study will determine which factors influence the sugar levels in the blood of patients. These may help health policy makers to take those factors seriously when treating diabetes patients

1.6. Research questions

1. By how much is a person Fasted Blood Sugar (FBS) change over time?
2. Does Systolic and Diastolic measurement play important role in diagnosing a person with Diabetes?
3. Does increase in weight increase one's probability of an increase in sugar content in the blood?

1.7. Limitation of the study

- There was limited information on patients in our hospitals. In a study like these, one is expected to have more information on another risk factor for diagnosing diabetes.
- Missing Data
- Some subjects are lost during the entire period of the study. The reason discontinuing of hospital visit may be differentially related to the treatment. For example, some subjects may developed side effects to an otherwise effective treatment and must discontinue the study.
- Irregular Spaced Measurement Occasions

  The time interval for patients to visit the Hospital for review varies from patient to patient. Some patients also violate the scheduled time giving them for review. These may not allow for using other statistical application like Markov structure based logistic model.

1.8. Methodology

The study seeks to use Generalized Linear Mixed Model (GLMM) and Generalized Estimating Equations. Statistical software such as SPSS and R-Studio were used for the study. The GLMM include subject effect to the logistic regression equation whilst the GEE does not.

1.9. Organization of the study

Chapter one introduces the research topic, where the study has been introduced and the research objectives presented. This also briefly explains which methods would be used for the conduction of the research.

The second chapter of the thesis is the Literature Review. In this review, a theoretical background about the main study is provided. The chapter starts by explaining what diabetes mellitus is and some risk factors which may affect the sugar level of diabetes patients. It further looks at the research analytical tools used in longitudinal studies. Some related reviews specifically on the longitudinal study were also discussed.

The third chapter presents the research methodologies, which include research populations, details description of GLMM and GEE techniques in the analysis.

In chapter four, the result and discussions were presented.

The final chapter, chapter five, presents findings from the study, conclusions of the key outcomes of the research, challenges faced and recommendations for future directions of the research.

The references cited and appendices, which include the data and, R codes used for the analysis, are also attached.

CHAPTER TWO

LITERATURE REVIEW

2.0. Introduction

The chapter two reviews the works others have done on the subject area of longitudinal study and analysis and the proposed statistical methodologies to achieve its intended purpose.

2.1. Diabetes Risk factors

Zimmet (1992) first referred to the "epidemic of diabetes," taking note of that its costs both in terms of financial burden and human enduring are increasing at a disturbing rate. The worldwide predominance of diabetes mellitus has been anticipated to about double from a benchmark of 2.8% out of 2000 to 4.4% by 2030, influencing more than 350 million people. In the decade starting in 1997, the pervasiveness of diabetes in the USA has expanded by 48% (Whiting, Guariguata, Weil, and Shaw, 2011).

Diabetes is pandemic in both developed and developing nations. In 2000, there were an expected 175 million individuals with diabetes worldwide and by 2030, the anticipated gauge of diabetes is 354 million (Mehta, Kashyap, and Das, 2009).

Diabetes in all forms has serious effect on health. In addition to the consequence of abnormal digestion of glucose (e.g., hyperlipidemia, glycosylation of proteins, and so on.), there are various long complexities related with the malady. These include cardiovascular, fringe vascular, visual, neurologic and renal variations, which are responsible for morbidity, disability and sudden death in youthful grown-ups. Diabetes remains an extremely huge reason for social, mental and money related weights in populaces around the world (Assal and Groop, 1999).

2.1.1 Risk factors for Type 1 diabetes

Type '1' diabetes (T1D) is caused by the damage of the beta cells of the pancreas by the auto immune system and there are close to 10% of all cases with diabetes (Pak et al., 1998). Presently, lifetime insulin treatment is the chief treatment for the illness. People with T1D will die without external insulin infusions. In spite of the fact that the predominance of T1D is less than 1% in many populaces, the geographical variety in occurrence is huge, spanning from <1/100,000 yearly in China to around 40/100,000 yearly in Finland (Karvonen, Tuomilehto, Libman, and LaPorte, 1993). It has been evaluated that roughly 20 million individuals around the world, mostly kids and youthful grown-ups, have T1D (Holt, 2004).The occurrence of T1D is rising globally at about 3% yearly (Onkamo, Väänänen, Karvonen, and Tuomilehto, 1999). This pattern seems to be most striking in the most youthful age groupings and is totally uncorrelated to the current increment in Type '2' diabetes (T2D) in kids. Many kids having beta cell autoantibodies, a sign of T1D, are being identified to have the T1D globally every year. Even though the peak age of inception is at puberty, T1D can likewise form in grown-ups. The study of disease transmission examines have uncovered no significant gender contrasts in frequency among people analyzed before age 15 (Kyvik et al., 2004). Nonetheless, after age 25, the proportion of male to female is around 1:5. Also, there is a striking cyclic variety in the incidence of T1D in many nations, with reduced rates in the warm summer months, and greater rates amid the icy winter (Dorman, Steenkiste, Burke, and Sonjini, 2003).

2.1.2. Risk factors for Type 2 diabetes

Type ‘2’ two diabetes mellitus (T2DM) is the most widely recognized type of diabetes mellitus (DM), which represents 90% to 95% of every single diabetic patient and has been projected to rise to 439 million by 2030 (Wu, Ding, Tanaka, and Zhang, 2014). In China, the most recent measurable information demonstrates that pre-diabetes and diabetes are predominant amongst individuals more than 20-year-old of age, with the rates 9.7% and 15.5% for T1DM and T2DM, separately. T2DM generally comes about because of the cooperation of hereditary, natural and other hazard variables.

In spite of the fact that T2DM patients normally free of external insulin, they may require it when levels of blood glucose are not adequately regulated with only diet or with oral hypoglycemic medications. Likewise, individuals with T2DM are regularly joined by complexities, for example, cardiovascular illnesses, diabetic neuropathy, nephropathy, and retinopathy. Diabetes and its related complexities affect patients’ lives and create huge financial and social burdens.

T2DM has turned into a noticeable worldwide public medical issue. Investigation of late measurable information uncovers that T2DM has a few new epidemiological qualities. Firstly, diabetes keeps an unfaltering increment in developed nations, for example, Japan and the United States. In addition, it is noteworthy that T2DM has turned into a significant issue at a disturbing rate in developing nations. The expectation is that T2DM would keep on increasing in the following twenty years, and over 70% of its patients will show up in developing nations, with the most of them in the age groups of 45-64 years. T2D is the most widely recognized type of the illness, representing around 90% of all affected persons. A detection of T2D is made if a fasting plasma glucose fixation is > 7 mmol/L (> 126 mg/dl) or plasma glucose 2 hours after a standard

glucose challenge is > 11.1 mmol/L (> 200 mg/dl). T2D is due to the comparative incapacitated insulin discharge and fringe insulin resistance. Normally, T2D is controlled workouts, diet, oral hypoglycemic specialists and occasionally exogenous insulin. Nevertheless, it is related with the same long-term difficulties as T1D (WHO, 1999).

T2DM has turned into a detectable worldwide general medical issue. Investigation of late measurable information uncovers that T2DM has a few new epidemiological qualities. Initially, diabetes keeps an unfaltering increment in well-to-do nations; for example, in the United States Native Americans have the most elevated rates of T2D, especially the Pima Indians who lives in Arizona, and in Nauru of the South Pacific islands (Wild, Roglic, Green, Sicree, and King, 2004). T2D is additionally identified to be more overwhelming in African American and Hispanic populaces than in Caucasians. In the year 2000, it was assessed that 171 million individuals (2.8% of the total populace) had diabetes and that by 2030 the figures would be 366 million (4.4% of the total populace). By far most of this expansion will happen in men and women matured 45 to 64 years living in third world nations. As per Wild et al. (2004), the "top" three nations regarding the quantity of T2D people with diabetes are India (31.7 million out of 2000; 79.4 million of every 2030), China (20.8 million in 2000; 42.3 million by 2030) and the US (17.7 million in 2000; 30.3 million by 2030).

Notwithstanding the troubles of T2D a much higher number of individuals exist with high levels of blood glucose yet beneath the level for diabetes. The World Health Organization characterizes impaired fasting glucose as a fasting plasma glucose level of > 6.1mmol and under 7mmol, and disabled glucose resilience as 2-hour plasma glucose, post glucose test, of 7.8 to under 11.1mmol (WHO, 1999).

2.1.3. Environmental Risk Factors

The epidemiological examples depicted above recommend that natural components add to the etiology of the T1D (Sperling, 2003). The current transient increment in T1D rate focuses to a varying worldwide condition as opposed to variety in the quality pool that needs the section of various eras. Research studies of twins additionally give confirmation to the significance of natural hazard elements for T1D. T1D concordance rates are higher for monozygotic twins than for dizygous twins (around 30% versus 10%, separately) (Hirschhorn, 2003). In any case, most homozygous twin sets continue conflicting. Hence, T1D cannot be determined completely by heredity.

Environmental hazard causes are perceived as being "initiators" or "quickening agents" of beta cell autoimmunity, or "precipitators" of clear indications in people who as of now have proof of beta cell demolition. They additionally might work by instruments that are explicitly injurious to the pancreas, or by implicit procedures that offer an irregular strong reaction to proteins typically existing in the cells. The environmental hazard figures of T1D that have gotten the most consideration are infections and baby nourishment.

Most ecological and case-control research studies have been centered on Enteroviruses, particularly Coxsackievirus B (CVB) (Dahlquist, 1998). CVB diseases are common amid adolescence and are known to have systemic influence on the pancreas. Potential examinations are explaining the part of infections to the etiology of T1D. For instance, enteroviral contaminations happening as right on time as in utero seem to build a youngster's resulting danger of building up the malady (Dahlquist et al., 1995; Hyoty et al., 1995). Different infections, including mumps (Hyöty et al, 1993), cytomegalovirus (Pak, Mcarthur, Eun, and

Yoon, 1988), rotavirus (Honeyman et al., 2000) and rubella, (McIntosh and Menser, 1992) have additionally been related with the disease.

Another theory that has been the centre of appreciable concern identifies with being exposed early to cow milk protein and the resulting formation of T1D (Kolb and Pozzilli, 1999). The main epidemiologic perception of such a connection was by Borch-Johnsen et al. (1984) who discovered that T1D children were breastfed for shorter timeframes than their non-diabetic peers from whole population. The researchers proposed that the absence of immunologic security from inadequate breastfeeding may build the hazard for T1D afterward amid adolescence. It was likewise proposed that shorter term of breastfeeding may implicitly echo early vulnerability to dietary proteins that invigorate an irregular insusceptible reaction in babies.

Obesity

Having a high measure of abundance muscle to fat ratio additionally ups type '2' diabetes hazard. Body Mass Index (BMI), which is a measure of muscle to fat ratio in light of stature and weight. The American Diabetes Association (ADA) says that all individuals with a BMI higher than 25 in addition to at least 1 other hazard variables ought to be tried for type '2' diabetes, regardless of what their age is (ADA, 2015).

Ethnicity

Inside the U.S., African-Americans have a twofold increment in the danger of the advancement of type'2' diabetes compared with Caucasians, and this hazard is marginally higher in ladies, most likely because of high corpulence. The risk is alarming, especially people of either Puerto Rican or Mexican origin, and was approximately 2.5 times more noteworthy than in Caucasians, while Native Americans demonstrate a fivefold increment in chance (ADA, 2015).

Age

The risk for type ‘2’ diabetes varies with age. The American Diabetes Association (ADA) prescribes that individuals be tried for type‘2’ diabetes beginning at age 45, particularly on the off chance that they are overweight or stout.

Lifestyle factors

A wide assortment of way of life components is likewise of extraordinary significance to the improvement of T2DM, for example, stationary way of life, physical latency, smoking and liquor utilization. Generous epidemiological investigations have demonstrated that weight is the most imperative risk consider for T2DM, which may influence the improvement of insulin resistance and illness movement (WHO, 2011).

2.2. Longitudinal data Analysis

Longitudinal investigation plans in biomedical research are propelled by the demands or yearning of a scientist to evaluate the variation of an outcome overtime. In a longitudinal study outline, the result is measured recurrently after some time for each person in the investigation. Contrasted with cross-sectional investigation plans, longitudinal examination outlines can be more effective, not so costly, but rather more vigorous to model selection and they can have increased statistical power (Edwards, 2000). There are a few classes of longitudinal investigation outlines, including, retrospective (case-control) designs, prospective (cohort or follow-up) designs, experimental designs and observational designs. The retrospective longitudinal investigation configuration is utilized to gather information on subjects going in reverse in time where the result variable for both cases (those definitely known to have sickness in view of their result) and controls (those definitely known to not have the disease) is more than once gathered

in reverse in time. The prospective longitudinal research design does the direct opposite i.e. it is utilize to gather information on subjects moving ahead in time. For the most part, subjects are chosen with and without risk components, and repeatedly followed over time to measure a characterized outcome variable.

Amid the current past, an extensive variety of factual methods have been utilized as a part of the examinations of longitudinal information. One of the most punctual procedures suggested for examining longitudinal information was a mixed effect ANOVA, with a solitary arbitrary subject effect (Fisher, 1918). The incorporation of an irregular subject effect incited positive connection among the repeated estimations on a same subject. Airy (1861) established the frameworks for linear mixed model formulation, prior to it being put on a more official hypothetical balance in the fundamental work of R. A. Fisher.

While ANOVA strategies can give a sensible premise to a longitudinal examination in situations where the investigation configuration is extremely basic, they have numerous deficiencies that have constrained their handiness in practice (Tabachnick and Fidell, 2007). In numerous longitudinal investigations, there is extensive variety among people in both the frequency and timeline of data collections. The subsequent information is profoundly uneven and not promptly manageable to ANOVA techniques created for balance design. It was these components of longitudinal information that gave the force to analysts to create significantly more flexible methods that can deal with the ordinarily experienced issues of information that are unbalanced and incomplete, mistimed measurements, time-fluctuating and time-invariant covariates, and reactions that are discrete instead of continuous (Fitzmaurice and Molenberghs, 2008).

With a given binary response gathered at successive time, most statisticians focuses on the connections between the probability of a success response and covariates. A large portion of the

takes a shot at Markov models managed estimation of move probabilities for first or higher order. Muenz and Rubinstein (1985) utilized calculated logistic models to break down the move probabilities starting with one state then onto the next for the first order. Islam and Chowdhury (2006) stretched out Muenz and Rubinstein model to higher order Markov demonstrate with covariate reliance for binary results. Instead of conditional probabilities, they utilized transitional probabilities generated from Markovian presumptions. Azzalini (1994) concentrates on models in which the marginal expectation of the response is related with the arrangement of covariates by some known link function. At the point when the response is binary, a characteristic decision is to utilize a logit link function to connect the marginal expectation of the responses to the covariates. The covariates can be time-differing and time- stationary.

A distinction among transitional and marginal models have been made by Ware, Lipsitz and Speizer (1988) relying upon the covariates determining the marginal distribution or conditional probabilities of any nominal observation. According to Stiratelli, Laid, and Ware (1984), Zeger and Qaquish (1988), Cox and Snell (1989), inside transitional models, the covariates' impact is likely to be on the of the Markov chain probabilities or on its mean values. Regardless, it gives off an impression of being unreasonably possible that, after fitting the model, one might want to use it for find the mean estimation of the response based on the covariates alone. Therefore, Azzalini favoured marginal model. Liang and Zeger (1986) and Prentice (1988) present the Generalized Estimating Equation. The key component of this approach is that one doesn't endeavor to show the joint distribution of the subject profile; just the marginal distribution at each time point is displayed as an element of the covariates, and the standard deviations of the regression coefficients are changed in accordance with the autocorrelation.

Lipsitz, Laid, and Harrington (1992) altered the estimation condition of Prentice to estimate the odd proportions. Lipsitz, Kim, and Zhao (1994) stretched out Liang and Zeger's strategies to show for the connection between ordinal categorical or repeated nominal responses. Although this method has a lot of advantages, it doesn't develop a model for the stochastic instrument which can produce the data; however Azzalini's proposed show has stochastic properties. Kosorok (1993) proposes a model called the equilibrium model. The basic element of this model is the way it breaks down a Markov procedure into two institutive and parametrically different segments: one corresponding to the equilibrium dispersion and the other to the rate of exchange between neighboring classes. The parameters in the equilibrium distribution part can be interpreted similarly that parameters in a marginal regression model; nonetheless, the balance appropriation display how covariates influence the future response. Laird and Ware (1982) demonstrated that generalized mixed-effects regression models could be utilized to carry out a more finished analysis of all of the existing longitudinal data under significantly more broad assumptions with respect to the missing data (i.e., missing at random; MAR)

2.3. Problems with longitudinal data

Repeated-measures designs have the apparent benefit of eliminating the distinction in responses between subjects. In any case, there are likewise some issues, which ought to be dealt with (Ware et al., 1988). These are a latent effect, carry-over effect and learning effect.

The latent effect happens when one treatment can initiate the lethargic impacts of a past treatment. At a point when latency effect is suspected, it is best to avoid repeated-measures design.

Learning effect happens in a situation where response improves each time the test is repeated on subjects (Fitzmaurice and Molenberghs, 2008).

The carry-over effect happens when a treatment is administered before the impact of the earlier treatment has subsided. The carry-over effect is dealt with by permitting enough time between treatments.

Three wellsprings of irregular variety in a regular arrangement of longitudinal information:

• Random impacts

• Serial connection

• Measurement mistake

At the point when estimations include sensitive judgments, copy estimations at same time on a similar subject may indicate generous variety (Diggle, Hergerty, Liag, and Zeger, 2002).

2.4. Data Collection Schedule

It is standard in a longitudinal report configuration to address the timelines for gathering repeated measures. Goldstein, Baxter-Jones, and Helms (1993) give two great definitions with respect to data collection planning for a longitudinal report:

a) A longitudinal research has a frequently planned schedule if estimations are stipulated in an even interval of the longitudinal study and furthermore has routinely timed data information if measures are really gotten at consistent intervals of the longitudinal investigation.

b) A longitudinal report has a reliably planned timetable if each subject has a similar time schedule, i.e., it is schedule to be assessed at a similar arrangement of longitudinal examination esteems, regardless of whether the calendar is frequently coordinated. A longitudinal report has

reliably planned information if each subject is assessed at a similar arrangement of longitudinal examination values.

In the investigation of numerous ceaseless ailments, one may anticipate having consistently planned schedules; however, the real information accumulation is unpredictably timed. For instance, in a longitudinal investigation of pulmonary function in cystic fibrosis with a frequently planned time, say toward the finish of every month for 6 months, cystic fibrosis patients may have sudden pneumonic intensification amid the month, which may require measurements at some point amid the month notwithstanding the finish of the month. A further illustration is that patients may miss the window of chance of pulmonary measurement and must be re-scheduling for another time. Watch that a longitudinal report can have a reliably time schedule, yet the timetable can be irregular planned. For instance, if information were schedule to be gathered at months 1, 3, and 6, at that point the investigation would have a reliably planned time yet the timetable would likewise be irregularly timed, (Edward, 2000).

2.5. Missing Data

Since longitudinal investigations seldom entire because of patient steady loss, ill-timed visits, untimely research end, demise and other causes, missing information in longitudinal investigations could be a troublesome issue to overcome. Missing information is reasonable just about a routine or fluctuating information gathering plan. Missing information can be ordered into two general classifications: randomly missing information, and non-randomly missing information. Randomly missing information can be additionally separated into "missing totally at random" (MCAR) or "missing at random" (MAR). Non-random missing information is frequently alluded to as usefully missing information.

Illustration: In a longitudinal clinical trial, completely at random means, patients leave the examination since they change habitation. At random: patients abandon the study on the recommendation of a medical doctor, in light of watched past measurement records. Not at random: patients miss their course of action since they are feeling unwell.

2.6. Statistical Models for Longitudinal Data Analysis

There are techniques to assess the impact of a set of predictor variables on dependent variables. Some of these techniques include Simple and multiple linear regression, discriminant analysis and logistic analysis among others. A linear regression is suitable for a continuous dependent variable and the logistic regression and discriminant analysis is use if the response variable is categorical.

Also in study where groups are notice, the focus is usually on predicting group attachment from a set of variables. Logit analysis, Discriminant analysis, and logistic regression are built to achieve this forecast. Logit analysis is use when independent variables are all discrete, discriminant analysis when all independent variables are continuous and nicely distributed, and logistic regression when independent variables are a mix of discrete and continuous and / or poorly distributed. The discriminant analysis unlike the logistic regression requires stringent assumptions to be met. The data should be normally distributed. Assumptions of linear relations and homoscedasticity should not be violated.

The data collected for this research have the predictor variables as both continuous and categorical and the dependent variable, a high or low sugar level is also categorical, and there is dependence in the data thus making the data suitable to be analyzed using generalized mixed-effects model and generalized estimating equation.

Rigorous and wide approaches in analyzing longitudinal data have been developed by statisticians. Among them incorporate mixed–effects regression models, Laird and Ware (1982) and generalized estimating equations (GEE) models, Liang and Zeger (1986). The essential refinement between the two general methodologies is that GEE models are partial likelihood methods and mixed-effects models are full-likelihood methods. The importance of the statistical models in light of partial likelihood is that (1) They generalized effortlessly to a diverse range of result measures with very unique distributional structures (2) it is less demanding to compute them than full-likelihood techniques. However, partial likelihood methods are more prohibitive in their assumption with respect to missing information than the full-likelihood techniques. What's more, full-likelihood methods give assessments of individual particular impacts that are valuable in understanding between subject variations in the longitudinal response method and in forecasting future responses for a given subject or an arrangement of subjects from a specific subgroup.

2.6.1. Generalized Linear Mixed Models (GLMMs)

Generalized linear models (GLMs) are a class of fixed effects regression models for multiple response variables (i.e., binary, continuous counts) without a subject effect. The GLM generalizes regression by

- relating the linear model with the response variable through connecting link function and
- making the absolute difference of every measurement to be a component of its predicted value.

In a GLM, every result of the response variable (y) is thought of being produced from a specific distribution of the exponential family distributions. The most widely recognized distributions from this family are Binomial, Poisson, and Normal. GLMMs expand GLMs by the

incorporation of random effects in the indicator variables. The general linear mixed model also known as the mixed model is a multivariate regression technique that generalized the analysis of variance (ANOVA) and general linear regression techniques. It is a general statistical method for investigating longitudinal information. It is a statistical approach for modeling continuous response measures as a member of fixed (populace) effects, while at the same time demonstrating singular subject parameters as random effects. Finally, the mixed model can contain both time-dependent and independent covariates.

As indicated by Manning (2007), a GLMM offers all the benefits of a logistic regression model,

a) Deals with a multinomial response variable.

b) Manages unbalanced data

c) Offers a lot of information on the size and direction of effects

d) Has a straightforward model composition, versatile post hoc for various investigations (instead of requiring diverse exploratory outlines)

e) Can do one combined study, which inculcates in it all random effects simultaneously.

CHAPTER THREE

METHODOLOGY

3.0. Introduction

This chapter outline the research technique utilized to accomplish the objectives of the study which were set out in chapter one. The chapter discusses the study research design, the study population, the sample size determination, source of data, data collection, data exploration (preliminary analysis) and the tools used for the data analysis.

3.1. Study Population

The study includes all the registered diabetes patients in the Maamobi polyclinic. As the Centre is a newly established one to help control diabetes, its population of patients is about 155. The researcher, therefore, collected data on the 155 available records at the time of the study.

3.2. Data Collection

Both primary and secondary data on diabetes were obtained from the Maamobi Polyclinic in the Greater Accra Region. Data on individual patients was record each time they visit the hospital for review for four consecutive times. However, variables such as educational level, occupation and family history were obtained from the patients on their last visit. The diabetes Centre was visited each week to meet patient schedule for review. Patients were expected to visit the Centre

once every month. The researcher, therefore, used four months to obtain the required data for the study.

The measurements collected on each individual include: Age, Sex, Weight, Systolic, Diastolic, and Fasted Blood Sugar (FBS) levels were recorded. The FBS measurement was used to diagnose whether an individual is at risk for diabetes to progress at a particular visit or not. A patient with FBS greater than or equal to 7.1mmol is considered as likely for diabetes to progress otherwise not showing progression but at risk of developing the disease. The data was analyzed to verify whether relationships exist between the sugar level and the other variables.

3.3. Data description

The researcher took all information recorded in the folders of the patients that are relevant to the study and how they relate to the sugar level. The variables that are recorded on each visit include:

- **Blood Pressure**. This measures the force pushing blood outwards on the blood vessel. It is normally recorded as two numbers, composed as a proportion like $\frac{120}{80} mmHg$. The top number which is called Systolic and furthermore larger between the two numbers, quantifies the pressure in the veins when the heart thumps (when the heart muscles contracts). The base number, which is likewise the smaller between the two numbers, quantifies the pressure in the arteries between heartbeats (when the heart muscles are resting between pulsates and refilling with blood). It is should be noted that

blood pressure above $\frac{140}{90} mmHg$ is termed as High Blood Pressure, or hypertension affects people with type 2 diabetes.

- **Weight.** Research demonstrates that one best indicator of type ‘2’diabetes is overweight or obesity. Individuals who are overweight for the most part include pressure their body's capacity to utilize insulin to appropriately control sugar levels and are in this way prone to create diabetes (Centre for Disease Control and Prevention, 2011).

- **Family History.** Family health history is a vital risk factor for developing type ‘2’ diabetes (Centre for Disease Control and Prevention, 2011).

- **Age.** There is a positive connection amongst age and one's risk for T2D, heart infections, and stroke (Centers for Disease Control and Prevention, 2011).

- **Gender.** As per American Diabetes Association, men will probably create heart illnesses than ladies. However, men with diabetes live 7.5 years less all things considered than the individuals who do not have, among ladies the distinction is more prominent than 8.2 years by and large (Centers for Disease Control and Prevention, 2011).

3.4. Data Analysis

This section describes the statistical techniques employed to analyze the data. The techniques discussed include logistic regression, generalized linear mixed models and generalized estimating equations.

3.4.1. Introduction to Logistic Regression

Regression analysis is utilized for displaying the connection between a single variable Y, called the response or dependent variable: and at least one indicator $\left(X_1,...,X_p\right)$. When p =1, it is called the simple regression but p >1 it is termed as multiple regressions. Where the response variable Y can take more than one value, the model is called multivariate regression. Now, when Y can take only two possible outcomes either 0 or 1, the model is referred to as logistic regression.

The study makes use of the binary logistic regression to establish a relationship between patients Fasted Blood Sugar (FBS) level (coded as 0 if FBS $\leq$ 7mmol and 1 otherwise) and a number of predictors (i.e. Age, Sex, Weight, Systolic, Diastolic and Family history of the patient).

To specify the logistic regression, we define the following: the logit is the natural logarithm (ln) of odds Y and odds are ratios of probabilities (p) of Y happening (i.e., FBS greater than 7mmol) to probability (1- p) of Y not happening (i.e., FBS less or equal to 7mmol)

Then the simplest logistic regression model is represented in equation (3.1)

$$\text{logit}\left(Y\right)=\ln\left(\text{odds}\right)=\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1X \tag{3.1}$$

where,

β_1 is the regression coefficient and

β_0 is the intercept at the logit scale

Re-arranging (3.1), we obtained the equation for the predicted probability of the outcome of interest as

$$\Pr(Y = \text{response of interest} | X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \qquad (3.2)$$

where,

e = 2.71828 is base of the system of natural logarithms.

The relationship between $\text{logit}(Y)$ and X in equation (3.1) is linear whereas from equation (3.2), the relation between the probability of Y and X is nonlinear.

Extending the logic of the simple linear regression to multiple predictors $X = (x_1, x_2 ..., x_n)$ generate equation 3.3 and 3.4

$$\text{logit}(Y) = \ln(\text{odds}) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \qquad (3.3)$$

and,

$$\Pr(Y = \text{outcome of interest} | X = x) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n}} \qquad (3.4)$$

Typically,

$\beta_0, \beta_1, ..., \beta_n$ are assessed by maximum likelihood (ML) strategy which is picked over weighted least squares technique by a few researchers like as Haberman (1978) and Schlesselman (1982). The ML technique is intended to maximize the likelihood of generating the data given the parameter estimates.

The general likelihood of the data set can be represented as the product across all cases of the probabilities as stated in equation 3.5.

$$L(y;\beta,\alpha) = \prod_{i=1}^{n} P\left(Y_i \mid x_{i1}, ..., x_{ip}\right) =$$

$$\prod_{i=1}^{n}\left\{\left(\frac{e^{\beta_0+\sum_{j=1}^{m}\beta_j x_j}}{1+e^{\beta_0+\sum_{j=1}^{m}\beta_j x_j}}\right)^{Y_i} \times \left(\frac{1}{1+e^{\beta_0+\sum_{j=1}^{m}\beta_j x_j}}\right)^{1-Y_i}\right\} = \prod_{i=1}^{n} p^{Y_i}\left(1-p\right)^{1-Y_i} \quad (3.5)$$

where, Y is the 0 or 1outcome for the i^{th} case based on a sample of n. β and α are fixed effects and variance components respectively. The use of Y_i and $1-Y_i$ as exponents in equation 3.5 indicates that the suitable probability term is dependent upon whether Y_i =1 or Y_i = 0 (Josephat and Ismail, 2012).

3.4.2. Generalized Linear Mixed Model

The Generalized Linear models (GLMs) indicates a class of fixed effects regression models for a few types of response variables (i.e. counts, binary, continuous). The study focuses on the binary outcome where the logistic regression will be applied. The Generalized linear mixed-effect models will include adding the subject specific random effect to the logistic model.

The GLMM utilizes the logit link, namely

$$g\left(\mu_{ij}\right)=\text{logit}\left(\mu_{ij}\right)=\log\left[\frac{p_{ij}}{1-p_{ij}}\right]=\eta_{ij}=\eta_{ij}=x_{ij}'\beta+Z'\gamma_i \tag{3.6}$$

The conditional expectation $\mu_{ij}=E\left(Y_{ij}\mid\gamma_i,x_{ij}\right)$ is equal to $P\left(Y_{ij}=1\mid\gamma_i,x_{ij}\right)$ which is the conditional probability of a positive response given the random effects and covariate values, and

Subscripts i and j represents patients and visit time respectively, and

γ_i is the random effect for each subject which represent the influence of the subject i on his/her repeated observation that is not captured by the observed covariates and is independent and identically normally distributed with mean 0 and variance τ ,i.e. $\gamma_i\sim(0,\tau)$

The design matrix for the random effect is Z

Alternatively, the GLMM can be specified as,

$$\log\left(\frac{p(Y_{ij}=1\mid x_i,\gamma_i)}{1-p(Y_{ij}=1\mid x_i,\gamma_i)}\right)=\beta_0+\sum_{j=1}^{m}\beta_j X_i+Z\gamma_i \tag{3.7}$$

From (3.7), the model can be re-arranged in terms of probability in the form

$$p(Y_{ij}=1\mid x_i,\gamma_i)=\frac{e^{\beta_0+\sum_{j=1}^{m}\beta_j X_i+Z\gamma_i}}{1+e^{\beta_0+\sum_{j=1}^{m}\beta_j X_i+Z\gamma_i}} \tag{3.8}$$

Maximum Likelihood Estimation of GLMM is obtained as follows:

The distribution of the outcome variable Y is assumed to be a member of the exponential families of the form

$$f_{y_{i,j|\gamma_i}}\left(Y_{ij} \mid \gamma_i, \tau\right) = \exp\left\{\left[\frac{Y_{ij}\gamma_i - b\left(\gamma_i\right)}{a\left(\tau\right)}\right] + c\left(Y_i, \tau\right)\right\} \tag{3.9}$$

where,

The first and second derivatives of $b(\gamma)$ gives the mean and variance of the distribution respectively γ_i termed as a canonical parameter is a location parameter of the distribution. The parameter τ is the dispersion parameter, and $c(Y_{ij}, \tau)$ is some function of Y_{ij} and τ .

The likelihood function is giving by

$$L(y; \beta, \gamma, \tau) = \int \prod_i^n f_{y_{ij|\gamma_i}}\left(Y_{ij} \mid \gamma_i\right) f_{\gamma_i}(\gamma_i) d\gamma_i$$

$$= \prod_i \int_{-\infty}^{+\infty} e^{\sum_1^N \left[\frac{Y_{ij}\gamma_i - b(\gamma_i)}{a(\tau)}\right] + c\left(Y_{ij}, \tau\right)} \times \frac{e^{-\frac{\gamma_i^2}{2\sigma_\gamma^2}}}{\sqrt{2\pi\sigma_\gamma^2}} d\gamma_i \tag{3.10}$$

Integrating equation (3.10) is complex hence the Laplace Approximation (LA) method is used to estimates the model parameters.

The LA method is used if the integral of a function is of the form $\int_a^b e^{Mf(x)} dx$. The computation of this method to estimates the parameters is done using statistical software.

3.4.3. Generalized Estimating Equation (GEE)

One of the major models used in longitudinal studies is GEE. The GEE extends GLM to accommodate the correlation between the responses $(Y's)$ as in GLMM. The decision of which model to utilize relies upon the objective of the investigation. When the interest of the study is on population averaged, the GEE is used and when the concern is the heterogeneity among subjects, the GLMM is used. The GEE is a marginal model while GLMM is a subject-specific model.

Examples of the goals of such studies include;

a). Subject-specific effects estimate the odds of a patient having FBS > 7mmol if the patient has a family history of diabetes compared to the odds of the same patient having FBS > 7mmol if otherwise.

b). Population average effects estimate the odds of an average patient with FBS > 7mmol and a positive family history to the odds of an average patient with FBS > 7mmol and a non-positive family history.

To fit the GEE model, it is required to

i. Specify the distribution
ii. Specify the *link* function
iii. Estimate the model parameters using quasi-likelihood
iv. Estimate the variance – covariance matrix of the model parameter.

The *link* function is usually *logit* linked since the logistic regression is used in the modeling.

The link function of the mean is $g(\mu_i) = X_i^T \beta$.

The method of parameter estimation uses the quasi-likelihood methods since the equation depends only on the mean and the variance of the outcome

From the quasi-likelihood approach, the working covariance matrix for y_i is given by

$V_i = A_i^{\frac{1}{2}} R_i(\alpha) A_i^{\frac{1}{2}}$.

And the estimator of the score-like equation system is given as

$$\sum_i^K \frac{\partial \mu_i}{\partial \beta} V_i^{-1}(y_i - \mu_i) = 0 \qquad (3.11)$$

where,

μ_i is the mean vector and a function of β (i.e., $\mu_i = (\mu_{i1}, ..., \mu_{it})'$)

$R_i(\alpha)$ is the working correlation matrix of Y_i which is the same to all patients. $R_i(\alpha)$ must be specified for the estimation. There are four main choices,

i. Correlation between Y_i are zero (independent)
ii. All correlation are assumed to be equal (Exchangeable)
iii. Every observation is only correlated with the adjacent observation (Independent or AR(1))
iv. All correlation are estimated from data (Unspecified)

A_i is t_i x t_i the diagonal matrix with the variance of $g(\mu_{ij})$ as the j^{th} diagonal elements

V_i is the corresponding covariance of $R_i(\alpha)$.

For the case where the outcomes are normally distributed with constant variance across time, we have

$$V(\alpha) = \varphi R_i(\alpha) \tag{3.12}$$

Park (1993) extents equation 3.12 to heterogeneity of variance across time by allowing the scale parameter φ_j to vary throughout time $(j = 1,...,n)$

The solution for β is

$$\beta = \left[\sum_{i=1}^{N} X_i^{'} \left[R_i(\hat{\alpha}) \right]^{-1} X_i \right]^{-1} \left[\sum_{i=1}^{N} X_i \left[R_i(\hat{\alpha}) \right]^{-1} y_i \right] \tag{3.13}$$

Solving the GEE involves iteration for the regression coefficient, the correlation, and scale parameters, α and φ. Given $R_i(\alpha)$ and φ, β can be computed by iteratively reweighted least square proposed by McCullagh and Nelder (1989).

Given the estimate of β, the standardized residual computed using the Pearson residuals as

$r_{ij} = \dfrac{\left(y_{ij} - \hat{\mu}_{ij} \right)}{\sqrt{\left[\hat{V}(\alpha) \right]_{ij}}}$, which are consistently used to estimate α and φ

In other to perform hypothesis tests and build confidence intervals, the standard errors related with the evaluated regression coefficients will be of interest. Liang and Zeger (1986) exhibit the estimators for a few distinctive working relationship structures. The GEE version includes:

1. Naïve or model-based estimator:

The GEE estimator is equivalent to the inverse of the Fisher information matrix used in GLM as an estimator of the covariance estimate of the ML estimator of $\hat{\beta}$

$$V(\hat{\beta}) = \left[\sum_{i}^{N} D_i' \left(V_i^{-1} \right) D_i \right]^{-1}$$

where, $D_i = \dfrac{\partial \mu_i}{\partial \beta}$ and for $D_i = X_i$

$$V(\hat{\beta}) = \left[\sum_{i}^{N} X_i' \left(V_i^{-1} \right) X_i \right]^{-1} \tag{3.14}$$

If the mean model and $R_i(\alpha)$ are correctly specified, then $V(\hat{\beta})$ is a consistent estimator of $\hat{\beta}$.

2. Robust or 'empirical' or 'sandwich' estimator:

Here the estimator is given by

$$V(\hat{\beta}) = \sum_{i}^{N} M_0^{-1} M_1 M_0^{-1} \tag{3.15}$$

where, $M_0 = \left[\sum_{i}^{N} D_i' \left(\hat{V}_i^{-1} \right) D_i \right]^{-1}$ and

$$M_1 = \sum_{i}^{N} D_i' \hat{V}_i^{-1} \left(y_i - \hat{\mu} \right) \left(y_i - \hat{\mu} \right)' \hat{V}_i^{-1} D_i$$

This has a consistent estimator of the covariance matrix of $\hat{\beta}$ even if $R_i(\alpha)$ is not specified correctly.

According to Royall (1986), the robust estimator gives a steady estimator of $V(\hat{\beta})$ even if the working correlation structure is not the true correlation of y_i.

The working correlation matrix has to be specified when using GEE in modeling to estimate the covariance of the parameter estimates. This determination represents the within-subject relationship of the responses on the dependent variables. Four types of working correlation will be examine in this study to measure the relationship between patients fasted blood sugar levels over time for four visits. They are briefly summarized in table 1 with examples presented in a matrix form.

Table 1: Working correlation assumptions

Working correlation assumption	Correlation Matrix
Independence GEE assumes that there is no correlation within groups of patients FBS measures.	$Corr(y_{ij}, y_{i,j+k}) = \begin{cases} 1, j=k \\ 0, j \neq k \end{cases}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = I$
Exchangeable GEE permits for a constant correlation between two given measurement of the FBS within subject for all the visit time points	$Corr(y_{ij}, y_{ii}) = \begin{cases} 1, j=k \\ \rho, j \neq k \end{cases}, \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$
Autoregressive GEE weighs the correlation within FBS levels by their different time for the four time visits.	$Corr(y_{ij}, y_{i,j+k}) = \rho^k, \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$
Unstructured GEE assumes different correlations of any two FBS measures for every patient.	$Corr(y_{ij}, y_{i,i+k}) = \begin{cases} 1, j=k \\ \rho_{jk}, j \neq k \end{cases}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}$

CHAPTER FOUR

RESULT AND DISCUSSION

4.0. Introduction

This chapter is divided into two main parts; the result of preliminary analysis section and that of further analysis section. The preliminary section would focus on the few basic statistical tests that would give broad specifications of the data. The further analysis would mainly employ the use of statistical tools such as generalized linear mixed model (GLMM) and generalized estimating equations (GEE). Summary of the results are presented in a form of tables and figures. Other results are presented in the appendices. In the discussion of the findings, the results were also compared to the relevant theories and concepts discussed in the literature review.

This study used both primary and secondary data on diabetes patients of the Maamobi Polyclinic in the greater Accra region in assessing factors affecting the rise and fall of the sugar level of the patients and fitting a statistical model to the data.

4.1. Preliminary Analysis

The preliminary analysis section deals with the distribution of patients gender, their ages, fasted blood sugar (FBS) levels, weight, systolic blood pressure and diastolic blood pressure.

4.1.1 Gender Distribution

The frequency distribution in Table 2 displays the frequency cross tabulation of the gender and age of patients in the diabetes Centre of the hospital.

Table 2: Sex * Age Cross tabulation

		Age						Total
		<30	30-39	40-49	50-59	60-69	70+	
Gender	Female	2	6	26	35	32	27	128
	Male	0	2	3	5	12	5	27
Total		2	8	29	40	44	32	155

Source: Field data (2016)

Table 2 show that the total number of female patient is 128 representing 82.6% of the total patients sampled. This is more than that of the male respondents, which is 27 representing 17.4%. The table shows that majority of the patients are above 40 years, that is, out of 155 total numbers of patients. Only 10 patients out of the 155 are below 40 years with the number of male and female respondents being 2 and 8 respectively. Majority of the patients are within the age of 60-69. The age group with the highest number of female patients is 50-59 while that of males is 60-69.

4.1.2 Distribution of Fasted Blood Sugar across gender

To seek the distribution of fasted blood sugar measurement across gender, the boxplot displayed the result such distribution.

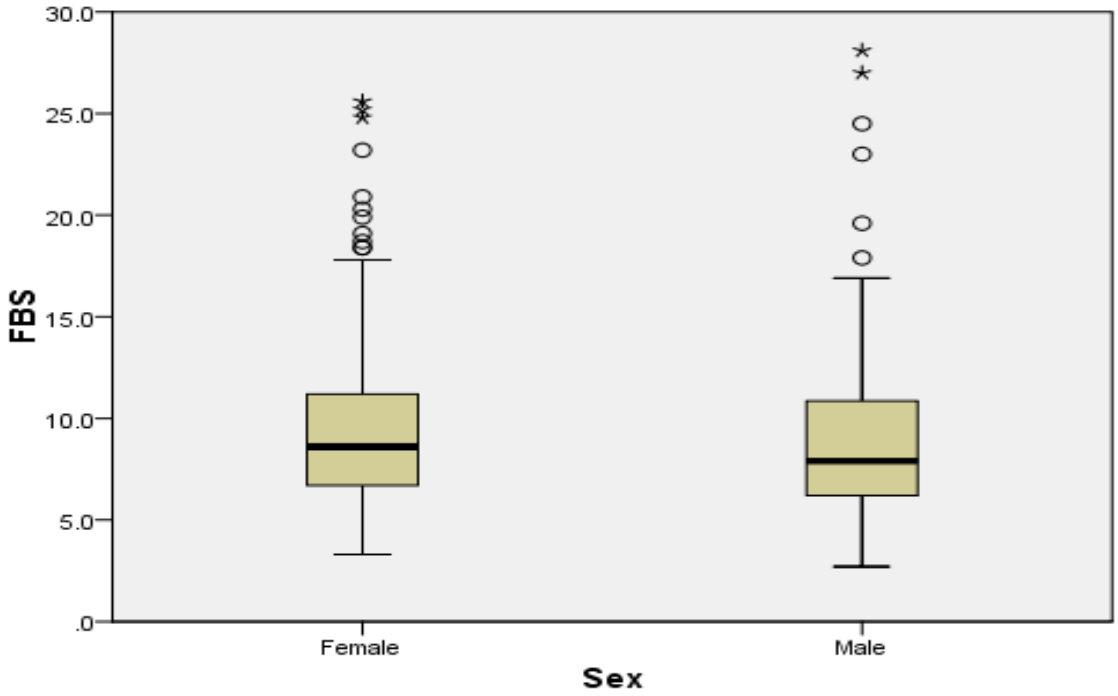


Figure 1: FBS distribution across gender

It can be observed from the Figure 1 that the FBS of all the patients has a positive skew data. Most of the patients FBS measurements fall between 5.0mmol and 15.0mmol with few patients having as high as 28.0mmol FBS. This means that most of the patients try to keep the sugar levels at a low level. The mean FBS is 9.14mmol, which is considered above a normal fasted blood sugar level. Males have fewer but high FBS values that can be considered as "outliers" than females. The standard deviation of 3.959 can prove that the measurements are not far apart from the mean. The descriptive statistics of the fasted blood sugar measurement is shown in appendix A.

4.1.3 Distribution of Educational level of patients against their fasted blood sugar level

While a few variables that add to the advancement of diabetes are past a person control, few elements can be control. Individuals can diminish their risk or postpone the advancement of diabetes in the event that they know about how to deal with these components. Education thus plays an important role if patients can read or learn about the risk factors that can be controlled (Rahman and Islam, 2007).

Table 3 gives a cross tabulation of the educational levels of patients and their FBS measurements.

Table 3: Educational Level * FBS Cross tabulation

		FBS		Total
		FBS≤7.0mmol	FBS>7.0mmol	
Educational Level	below secondary	44	84	128
	secondary and higher	6	21	27
Total		50	105	155

Source: Field data (2016)

Table 3 shows that out of 155 patients, the educational level of 128 are below secondary while only 27 have secondary education or higher. The number of patients with FBS higher than 7.0mmol is more than those with FBS less or equal to 7mmol across the educational levels. About 97 patients representing 62.6% of the FBS measurements are above 7mmol and 37.4% are less or equal to 7mmol.

4.1.4 Distribution of Occupation of patients

Table 4 gives the frequency distribution of the occupation of patients and their fasted blood sugar measurements.

Table 4: Occupation * FBS Cross tabulation

		FBS		Total
		FBS≤7.0	FBS>7.0	
Occupation	Unemployed	26	54	80
	Informal sector	22	47	69
	Formal sector	2	4	6
Total		50	105	155

Source: Field data (2016)

Table 4 displays the cross tabulation of the occupation of patients and their FBS measurements. Out of the 155 patients, 80 representing 51.6% are unemployed, 69 representing 44.5% are employed in the informal sector, and 6 representing 3.9% are employed in the formal sector. Here again, the number of patients with FBS greater than 7mmol across the occupational levels are more than those with less FBS level.

4.1.5 Distribution of systolic blood pressure and Fasted Blood Sugar levels of patients

Cross tabulation of the systolic blood pressure measurements of patients and their fasted blood sugar level is displayed in table 5.

Table 5: Systolic * FBS Cross tabulation

		FBS		Total
		FBS≤7.0	FBS>7.0	
Systolic	≤120	38	97	135
	121-130	29	63	92
	131-140	33	58	91
	141-160	29	72	101
	161-180	17	37	54
	>180	10	12	22
Total		156	339	495

Source: Field data (2016)

Out of the 495 systolic blood pressure recorded, 339 (68.5%) have fasted blood sugar greater than 7.0mmol and 156 representing 31.5% have FBS less or equal to 7.0mmol. The number of records with higher systolic blood pressure above the normal systolic value of 140mmHg is 177 out of 495 measurements.

Table 6 present the cross tabulation of systolic and diastolic blood pressure to measure the number of hypertension cases among the diabetes patients.

Table 6: Systolic * Diastolic Cross tabulation

		Diastolic Blood Pressure						
		≤80	81-85	86-90	91-100	101-110	>110	Total
Systolic Blood Pressure	≤120	122	0	12	1	0	0	135
	121-130	50	7	31	4	0	0	92
	131-140	46	5	36	4	0	0	91
	141-160	32	3	35	28	3	0	101
	161-180	9	1	14	14	14	2	54
	>180	0	0	2	9	5	6	22
Total		259	16	130	60	22	8	495

Source: Field data (2016)

It can be observed from Table 6 that 122 records give an optimal blood pressure, 7 has a normal blood pressure, 36 has high normal systolic value, 28 has mild hypertension, 14 has moderate hypertension and 6 has severe hypertension.

4.1.6 Distribution of diastolic blood pressure against fasted blood sugar level of patients

As indicated by Centers for Disease Control and Prevention (2011), hypertension, which is the measure of the systolic circulatory strain against diastolic pulse, assumes an essential part in deciding the sugar level of diabetic patients. This study, therefore, displays in Table 7 the cross tabulation of diastolic blood pressure of the patients under study and their fasted blood sugar level.

Table 7: Diastolic * FBS Cross tabulation

		FBS		Total
		FBS ≤ 7mmol	FBS>7mmol	
Diastolic BP	≤80	76	183	259
	81-85	6	10	16
	86-90	43	87	130
	91-100	19	41	60
	101-110	7	15	22
	>110	5	3	8
Total		156	339	495

Source: Field data (2016)

From Table 7, it can be seen that majority of the DBP measurements fall in the normal range of below 100mmHg with few above 100mmHg. About 405 out of 495 representing 81.8% of the diastolic measures are within the normal range (i.e., between 60mmHg and 90mmHg) and 18.2% have an abnormal diastolic condition.

In all the categories of the diastolic blood pressure, the numbers of records with FBS greater than 7.0mmol are more than those with less or equal to 7.0mmol except when the DBP is above 110mmHG.

4.1.7 Distribution of FBS against family history of patients

The family history of diabetes patients plays an important role in the diagnosing of an individual of diabetes. According to Centre for Disease Control and Prevention (2011), anybody with either

a parent or sibling with diabetes, stands a high risk of developing diabetes in future. Table 8, therefore, displays the cross tabulation of the family history of patients and their fasted blood sugar levels.

Table 8: Family History * FBS Cross tabulation

		FBS		Total
		FBS≤7.0	FBS>7.0	
Family History	No	33	17	50
	Yes	41	64	105
Total		74	81	155

Source: Field data (2016)

Out of 155 patients in the study (Table 8), 105 representing 67.7% had a family history of diabetes and 50 patient representing 32.3% have no diabetes in their families. For the patients without a diabetes history, 17 (34%) out of 50 have their sugar level been greater than 7.0mmol and 33 (66%) have their sugar level, not more than 7.0mmol at the first visit of the study. On the other hand, 64 representing 61% of those with family history recorded sugar level greater than 7mmol as against 41 representing 39% of less sugar level. It can therefore, be deduced that, patients with diabetes history in the family tend to have higher sugar level than those without any family history.

4.1.8 FBS distribution over time

The sugar level of patients is expected to reduce or be stable at a minimum level as the patients visit the hospital regularly for review. (i.e. $y_{t+1} \leq y_t$) Table 9, therefore, gives the descriptive statistics of FBS over the four-visit period.

Table 9: Descriptive statistics of FBS over visit

	Visit			
	1	2	3	4
Mean	9.972	8.937	9.296	9.056
Median	8.8	8.4	8.6	8.5
Variance	21.686	9.985	15.613	9.769
Std. Dev	4.6568	3.16	3.9513	3.1256
Minimum	3.6	3.3	2.7	3.6
Maximum	28.1	18.4	24.8	23.2
Skewness	1.501	0.766	1.59	1.637
Valid cases	155	142	109	89
FBS ≤ 7.0mmol	50	45	37	24
FBS > 7.0mmol	105	97	72	65

Source: Field data (2016)

Results displayed in Table 9 show that even though at baseline with 155 patients, the number reduces for the subsequent visits. Only 89 patients were had complete records. In all the visits, the number of patients with FBS>7mmol are more than those with less FBS.

The mean value of the sugar level is more on the first visit than the subsequent visits with a maximum record of 28.1mmol as against 18.4mmol in the second visit. The variance in the measurements is lower in the second and fourth visits than in the first and third visit. The FBS is positively skewed in all the four visits with the second visit having the lowest skewness value.

4.1.9 Weight distribution of patients

Obesity or being overweight can leads to diabetes. Figure 2 display the scatter diagram of the weight of patients and the fasted blood sugar levels.

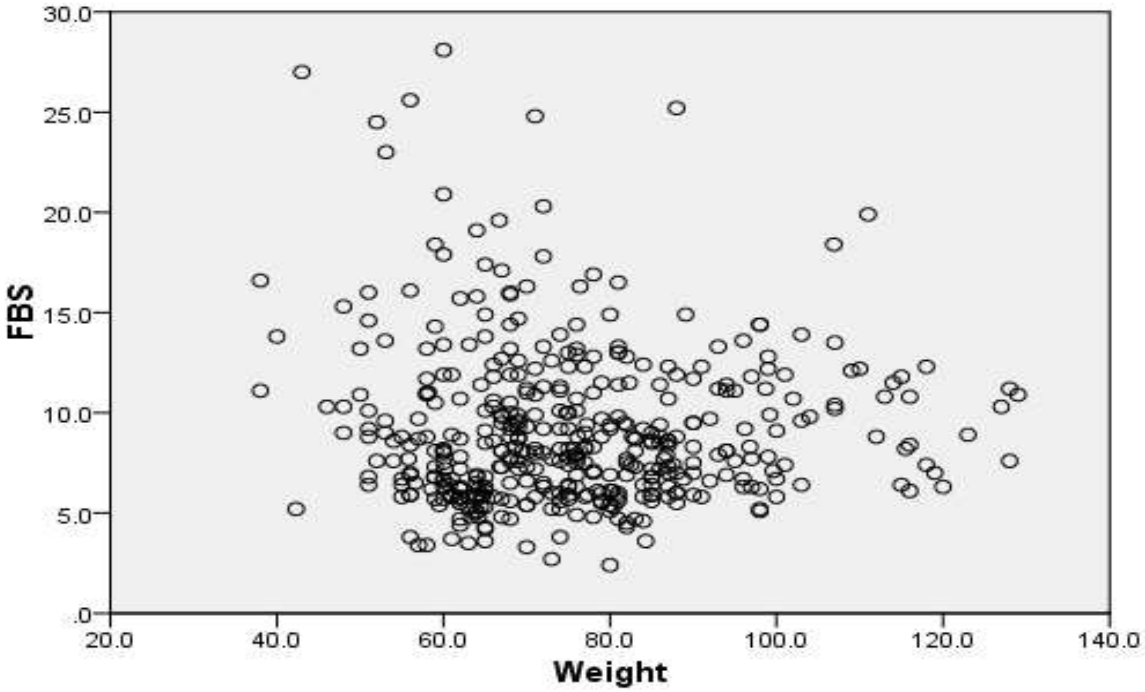


Figure 2: Distribution of FBS against weight of patients

It can be observed from figure 2 that the weight of the patients is concentrated between 50kg and 100kg while that of the FBS is between 5mmol and 15mmol. It appears from the diagram that the FBS does not change much as the weight increases. However, the plot displayed a nonlinear relationship between weight and fasted blood sugar level of patients

4.1.10 Combination of Systolic and Diastolic measures

Is there a situation where the systolic and diastolic measures move in the opposite direction over time?

According to WHO (2011), the heart beats-optimal pressure is being 120/80 mmHg. It is believed that systolic and diastolic have a linear relationship. If there is an increase in systolic measurement, there will be an increase in the diastolic measurement. However, this is not the case always. Researchers gave two conditions in which this exist.

i. **Isolated diastolic hypertension**: - This describes a situation at which the diastolic blood pressure is below 60mmHg and the systolic blood pressure remains higher than 100mmHg.

ii. **Isolated systolic hypertension**: - In this type hypertension, the systolic blood pressure is higher than 140mmHg, whilst the diastolic blood pressure remains normal. This condition as per WHO is basic in the elderly and it builds the danger of heart failure.

Table 10 and 11 shows the respective cross tabulation of isolated diastolic and systolic blood pressure.

Table 10: Isolated systolic hypertension

		Systolic	
		Normal	Not normal
Diastolic	Normal	195	110
	Not normal	46	144

Table 11: Isolated diastolic hypertension

		Systolic	
		Normal	Not normal
Diastolic	Normal	446	23
	Not Normal	13	13

It can be observed that 110 of the blood pressure measurements show an isolated systolic hypertension and 13 shows an isolated diastolic hypertension.

Table 12: Descriptive statistics of Systolic and Diastolic blood pressures across

		Statistic	Systolic	Diastolic
Age	<30	Mean	118.17	76.83
		Medium	113.50	75.00
		Variance	380.167	118.567
		Sta. Dev	19.498	10.889
		Minimum	100	63
		Maximum	150	90
	30-39	Mean	132.12	78.85
		Medium	130.00	80.00
		Variance	600.826	75.335
		Sta. Dev	24.512	8.680
		Minimum	90	60
		Maximum	210	90
	40-49	Mean	135.15	82.45
		Medium	130.00	80.00
		Variance	755.368	152.903
		Sta. Dev	27.484	12.365
		Minimum	80	50
		Maximum	210	110
	50-59	Mean	139.49	83.85
		Medium	140.00	80.00
		Variance	576.000	270.991
		Sta. Dev	24.000	16.462
		Minimum	90	60
		Maximum	222	190
	60-69	Mean	137.92	81.54
		Medium	140.00	80.00
		Variance	444.812	134.265
		Sta. Dev	21.091	11.587
		Minimum	80	50
		Maximum	195	110
	70+	Mean	146.68	81.51
		Medium	150.00	80.00
		Variance	549.909	132.768
		Sta. Dev	23.450	11.522
		Minimum	100	60
		Maximum	220	110

Table 12 confirms the finding of Franklin, Sutton-Tyrrel, Belle, Weber, and Kuller (1997) that systolic blood pressure increases with age whereas diastolic blood pressure rises until 60 years after which it starts to decline.

4.2 Model fitting

In section 4.2 and 4.3, the data was subjected to only the preliminary analysis, but in this section, GLMM and GEE are used in fitting the data. The analysis was carried out using ‘R' software.

4.2.1 Fitting a model using GLMM

The generalized linear mixed model was utilized to access the effects of age, sex, weight, family history, systolic blood pressure, diastolic blood pressure, and visit on a binary FBS. At a particular review, the doctor interest is whether the sugar level fall or rise above a threshold. For this reason, the researcher seeks to use the binary FBS instead of the continuous FBS.

Table 13 gives the output of the parameter estimation of both fixed and random effects using glmer function in R.

Table 13: Parameter estimates using GLMM

Random effects:				
Groups name	Variance	Std. Deviation.		
Subject (Intercept)	4.596	2.144		
Fixed effects:				
	Estimate	Std. Error	z value	P-value
(Intercept)	-1.725	1.969	-0.876	0.380
Age	0.034	0.018	1.931	0.053
Sex Male	-0.011	0.630	-0.019	0.985
Weight	0.032	0.016	1.986	0.046 *
Systolic	-0.018	0.008	-2.212	0.026 *
Diastolic	0.003	0.015	0.241	0.809
FH Yes	1.751	0.492	3.555	0.000 ***
Visit	0.101	0.126	0.806	0.420

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

The variance of the random intercept (4.596) shows large between-patients variability FBS. Weight, Systolic blood pressure, and Family history are statistically significant at 5%

significance level. Age, on the other hand, is significant at 10%. Gender, Diastolic blood pressur e, and Visit are not significant in the model. A patient with higher weight is more likely to have higher FBS level than those with lower weight on a logit scale. In the case of systolic blood pres sure, it has a negative effect on the likelihood of FBS level. The positive value of family history means that it more likely that a person with a positive family history of diabetes to have higher F BS level than those without. The standard errors of the covariates are also quite small.

4.2.2 GLMM model of covariates interaction with visit

Table 14 display the output of GLMM model for the significant covariates in table 13 with interaction with time.

Table 14: Covariates interaction with time under GLMM

Random effects:				
Groups name	Variance	Std. Deviation.		
SN (Intercept)	4.032	2.008		
Fixed effects:				
	Estimate	Std. error	z-value	P-value
Age	0.029	0.027	1.091	0.275
Visit	0.760	1.047	0.726	0.467
Weight	0.030	0.023	1.298	0.194
Systolic	-0.008	0.014	-0.565	0.572
FHYes	2.325	0.725	3.203	0.001**
Age:Visit	0.001	0.009	0.145	0.884
Visit:Weight	0.000	0.008	0.029	0.976
Visit:Systolic	-0.004	0.005	-0.803	0.421
Visit:FHYes	-0.310	0.253	-1.221	0.222

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

The GLMM display demonstrated no significant effect for any of the interaction factors. Individual covariates, which were significant in table 13, tends not significant in the interaction model (Table 14) except family history.

4.2.3 GEE output

The summarized GEE output is presented in the Table 15.

The GEE parameter estimation for the principle impacts in light of three suppositions (Exchangeable, Independent, and Unstructured) with their separate model-based and observational based standard errors assessments and p-values. The autoregressive (AR) working correlation matrix structure assumes that measurements are equally spaced for all subjects. However, the measurements for this study are not equally spaced hence AR (1) is not included in the Table 15.

Apart from gender, the estimates for the other parameters are similar across all the correlation assumptions. The estimates for the intercept, sex, and systolic blood pressure are negative for all the assumptions.

The estimates for age, weight, systolic, and family history are significant at $\alpha = 0.05$ for the exchangeable and unstructured assumption while only age and family history are significant under the independent assumption. This gives a clear indication that there is a correlation in the data for time independent covariates.

Table 15: GEE parameter estimates

	Exchangeable				**Independence**				**Unstructured**			
Parameter	**Estimate**	**N S.E.**	**R S.E.**	***P-value***	**Estimate**	**N S.E.**	**R S.E.**	***P-value***	**Estimate**	**N S.E.**	**R S.E.**	***P-value***
(Intercept)	-0.945	1.186	1.017	*0.353*	-0.927	0.962	1.046	*0.375*	-0.686	1.14	1.01	*0.496*
Age	0.021*	0.01	0.01	*0.048*	0.021*	0.007	0.01	*0.041*	0.021*	0.01	0.01	*0.044*
sexMale	-0.034	0.377	0.379	*0.927*	-0.125	0.274	0.366	*0.732*	-0.071	0.356	0.373	*0.848*
Weight	0.018*	0.009	0.009	*0.044*	0.012	0.007	0.008	*0.152*	0.017*	0.009	0.008	*0.049*
Systolic	-0.011*	0.005	0.005	*0.024*	-0.01	0.005	0.005	*0.067*	-0.011*	0.005	0.004	*0.022*
Diastolic	0.002	0.009	0.008	*0.785*	0.004	0.01	0.009	*0.611*	0.00	0.009	0.007	*0.991*
FHYes	0.995*	0.288	0.297	*0.001*	1.043*	0.21	0.29	*0.00*	0.965*	0.271	0.295	*0.001*
Visit	0.059	0.078	0.068	*0.391*	0.063	0.095	0.075	*0.394*	0.047	0.081	0.068	*0.491*

Where N S.E is ‘Naive’ or model-based estimator standard error and R S.E are Robust estimator standard errors.

Considering the estimates standard errors, they are nearly not far apart from each other for the three models. Choosing a correct correlation structure in modeling GEE is one of the important choices in GEE modeling. All the correlation structure presented in the table give similar standard errors. According to Royall (1986), when the correlation structure is modeled rightly, then the naive and robust estimators will give similar standard errors.

The standard error assess for the empirical estimator and the model-based are around the same for the GEE exchangeable model than the other two models.

In this regard, the exchangeable working correlation will be preferred to the other two structures in fitting the analysis.

Table 16: GEE covariates with Time (visit) interaction

Parameter	Exchangeable				Independence				Unstructured			
	Estimate	**N S.E**	**R S.E.**	***p-value***	**Estimate**	**N S.E**	**R S.E.**	***p-value***	**Estimate**	**N S.E**	**R S.E.**	***p-value***
(Intercept)	-2.001	1.996	1.63	*0.219*	-2.429	*2.213*	1.767	*0.169*	-2.161	2.097	1.735	*0.213*
Age	0.021	0.017	0.014	*0.117*	0.013	*0.018*	0.015	*0.362*	0.025	0.018	0.015	*0.092*
Visit	0.574	0.742	0.622	*0.355*	0.751	*0.898*	0.681	*0.27*	0.708	0.762	0.657	*0.281*
SexMale	0.586	0.605	0.626	*0.349*	0.556	*0.646*	0.639	*0.384*	0.539	0.638	0.606	*0.374*
Weight	0.021	0.016	0.014	*0.152*	0.016	*0.016*	0.014	*0.262*	0.024	0.017	0.014	*0.109*
Systolic	-0.003	0.011	0.01	*0.738*	0.002	*0.013*	0.01	*0.882*	-0.008	0.011	0.01	*0.399*
Diastolic	-0.005	0.021	0.019	*0.778*	0.001	*0.024*	0.021	*0.962*	0.0005	0.022	0.019	*0.981*
FHYes	1.481*	0.455	0.438	*0.0007*	1.466*	*0.491*	0.445	*0.0009*	1.404*	0.482	0.457	*0.002*
Age:Visit	-0.0004	0.006	0.005	*0.921*	0.003	*0.008*	0.006	*0.566*	-0.002	0.006	0.005	*0.741*
Visit:sexMale	-0.291	0.214	0.193	*0.133*	-0.307	*0.259*	0.209	*0.142*	-0.273	0.219	0.184	*0.138*
Visit:Weight	-0.0008	0.005	0.004	*0.844*	-0.002	*0.006*	0.005	*0.71*	-0.003	0.006	0.005	*0.541*
Visit:Systolic	-0.004	0.004	0.004	*0.293*	-0.005	*0.005*	0.004	*0.234*	-0.001	0.005	0.004	*0.707*
Visit:Diastolic	0.003	0.008	0.008	*0.659*	0.003	*0.01*	0.009	*0.859*	-0.0003	0.009	0.008	*0.973*
Visit:FHYes	-0.231(·)	0.162	0.133	*0.0826*	-0.194	*0.197*	0.154	*0.209*	-0.194	0.166	0.138	*0.159*

Where (*) is significance at $\alpha = 0.10$

Most of the interaction terms in the three models have negative estimates (Table 16). This may indicate a decrease in the probability of response. Considering all the correlation assumption, family history interaction with the visit is significant at 10% significance level under the exchangeable assumption. All the other interaction gives non-significant estimates. Only family history is significant among the individual covariates across the three-correlation assumption.

The two standard error estimates for the exchangeable correlation matrix assumption have similar values than the other two correlation matrix assumptions.

There is a marginally significant effect of patients with a positive family history of diabetes being more likely to have FBS greater than 7mmol than those without a family history of diabetes.

Systolic blood pressure has a significant negative effect. However, this is not consistent with literature. Henry, Thomas, Benetos, and Guize (2002) suggested an increase in high blood pressure is likely to increase the fasted blood sugar level and vice versa. The finding of negative effect may be due to measurement or data recording errors.

The fitted logistic model under GEE for the significant covariates will therefore be

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0692 + 0.02\,Age + 0.018Weight - 0.011sbp + 0.995\,fh$$

where, $\hat{p}$ the estimated probability of FBS greater than 7mmol, sbp is systolic blood pressure and fh is a family history of patients.

CHAPTER FIVE

FINDINGS, CONCLUSION AND RECOMMENDATION

Introduction

In chapter five, findings, conclusions and recommendations regarding the significance of the work are made based on the study results.

5.1 Key Findings of the study

Maximum likelihood method was used to determine the effect size of the risk factors of diabetes on the sugar level of the diabetes patient. The results suggested that Age, Weight, Systolic blood pressure and positive family history are significant in determining the sugar level of patients.

Other findings

- The majority of diabetes patients are 50years and above with females more than the males. The mean fasted blood sugar for both males and females falls above 7.0mmol with males slightly above the females.
- Out of 495 sugar level measurements recorded, 339 representing 68.5 percent are above the threshold of diagnosing a patient as a diabetic.
- Most of the patients have no ‘or' educational level below the secondary school.
- More than 50% of the patients are not in any active employment. This may have an effect on their ability to buy the necessary drugs to reduce or maintain the sugar level at a minimum.

- In all the categories of the systolic blood pressure, the number of patients with higher FBS is more than those with the lower FBS. About 177 out of the 495 systolic blood pressure measurements are higher than the normal 140mmHg.
- Some few diastolic blood pressure measurements are below 60mmHg and few above the normal 90mmHg.

The GLMM gives Weight, Systolic blood pressure and Family history as the significant at 5% significance level with Age significant at 10% significance level. The GEE, on the other hand, included Age as being significant at 5% significance level in determining the sugar level of diabetes patients. Gender, Diastolic blood pressure and visit times are not statistically significant and hence play a less important role in determining the sugar level of patients. Time interaction with covariates shows no significant for both GLMM and GEE models. The study, therefore, confirm findings from other researchers that family history of diabetes, overweight, increasing age, and high blood pressure play a very important role in diagnosing a high sugar content in the blood of diabetes patients. Physical inactivity was also seen as major risk factors (Carlsson, Midthjell, Tesfamarian, and Grill, 2007). Family history of diabetes, an increased BMI, and increased levels of blood pressure among others are found out to play a significant role in diagnosing diabetes (Lyssenko et al., 2008)

5.2 Conclusions

This study has presented an outline of a portion of the key ideas required in the investigation of longitudinal information. Two of the most ordinarily utilized measurable techniques for the investigation of longitudinal information were thoroughly stressed: generalized linear mixed model and the technique of generalized estimating equations (GEEs). Diabetes mellitus data

from the Maamobi Polyclinic was used to illustrate the application of these two techniques. When there are missing observations there is a sharp contrast between the responses of GEE and the mixed model (Park, 1993).

GLMM and GEE both gave similar results regarding the significance effect of the risk factors of diabetes of the sugar level of patients. While the mixed effect model includes individual patient effect in the model and cannot be used to generalize the result to other patients, the marginal model allows results be used otherwise.

The major contribution of the study is the analysis of the factors that contribute to the rise and fall of the sugar level of diabetes patients. Health providers can base on these factors in controlling the sugar level of their patients. In controlling the blood pressure of patients, care must be taken by health providers to prevent isolated systolic and diastolic hypertensions. Patients with a family history should be giving much attention since it is highly significant than all the other risk factors considered in this study. Patients should also take into consideration their diet not to overweight since weight play important role in sugar content in the blood.

The number of times a patient visits the hospital for review play a less important role whether the sugar level will decrease of increase. This suggests that patients can maintain low blood sugar when they can control some factors like eating fewer carbohydrates and sugar-free foods without visiting the Health Centre frequently.

5.3. Recommendations

Other risk factors of diabetes like alcohol intake, smoking habits, exercising habits, etc., needs to be incorporated in further studies on their effects on the sugar level of patients. Other models such as Markov model, Bayesian estimation can be used to compare the results obtained by the GEE and GLMM.

Further studies in this area need to address some of the limitations encountered in this study. Due to the repeated measures on the same individual for a period, researchers should do the measurements themselves and not rely on Hospital records to avoid measurements errors that may arise in the records. Regular visit time spacing should use in other to allow other models to be used in the analysis to compare results. Patients from more than one hospital should be considered to know whether there is a group difference in the variation of the sugar levels.

Healthy life style through aerobic exercise and proper diet is recommended for maintaining the sugar level in the blood very low. People with family history of diabetes should frequently test their sugar level since they are at a high risk of developing it.

REFERENCES

Airy, G. B. (1861). *On the algebraical and numerical theory of errors of observations and the combination of observations*. Macmillan and Company.

American Diabetes Association (2015). Standards of medical care in diabetes. *Diabetes care*, *28*(suppl 1), s4-s36.

Assal, J. P., and Groop, L. (1999). Definition, diagnosis and classification of diabetes mellitus and its complications. *World Health Organization*, 1-65.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, *81*(4), 767-775.

Borch-Johnsen, K., Mandrup-Poulsen, T., Zachau-Christiansen, B., Joner, G., Christy, M.,Kastrup, K., and Nerup, J. (1984). Relation between breast-feeding and incidence rates of insulin-dependent diabetes mellitus: a hypothesis. *The Lancet*, *324*(8411), 1083-1086.

Carlsson, S., Midthjell, K., Tesfamarian, M. Y., and Grill, V. (2007). Age, overweight and physical inactivity increase the risk of latent autoimmune diabetes in adults: results from the Nord-Trøndelag health study.*Diabetologia*, *50*(1), 55-58.

Centers for Disease Control and Prevention (2011). National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2010.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data*, 32..

Dahlquist, G. (1998). The aetiology of type 1 diabetes: an epidemiological perspective. *Actapaediatrica*, *87*(s425), 5-10.

Dahlquist, G., Frisk, G., Ivarsson, S. A., Svanberg, L., Forsgren, M., and Diderholm, H. (1995).Indications that maternal coxsackie B virus infection during pregnancy is a risk factor for childhood-onset IDDM. *Diabetologia*,*38*(11), 1371-1373.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002).Analysis of longitudinal data. *Oxford Statistical Science Series*.

Dorman, J. S., Steenkiste, A. R., Burke, J. P., and Songini, M. (2003). Type 1 diabetes and multiple sclerosis. *Diabetes Care*, *26*(11), 3192-3193.

Edwards, L. J. (2000). Modern statistical techniques for the analysis of longitudinal data in biomedical research. *Pediatric pulmonology*, *30*(4), 330-344.

Fisher, R. A. (1918). The correlation among relatives on the supposition of mendelian Inheriatnce. *Aust. J. Agric. Res*, *14*, 742-757.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 22 (5), 700-725.

Fitzmaurice, G. M., and Molenberghs, G. (2008). Advances in longitudinal data analysis: an historical perspective. *Longitudinal Data Analysis*, 3-30.

Franklin, S. S., Sutton-Tyrrell, K., Belle, S. H., Weber, M. A., and Kuller, L. H. (1997). The importance of pulsatile components of hypertension in predicting carotid stenosis in older adults. *Journal of hypertension*, *15*(10), 1143-1150.

Goldstein, H., Baxter-Jones, A., & Helms, P. (1993). Models for analysis of longitudinal data. *European Respiratory Journal*, *6*(9), 1416-1416.

Guariguata, L., Whiting, D. R., Hambleton, I., Beagley, J., Linnenkamp, U., and Shaw, J. E. (2014). Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes research and clinical practice*, *103*(2), 137-149.

Haberman, S. J. (1978). Analysis of Quantitative Data.Volume 1 Introductory Topics.

Henry, P., Thomas, F., Benetos, A., & Guize, L. (2002). Impaired fasting glucose, blood pressure and cardiovascular disease mortality. *Hypertension*, *40*(4), 458-463.

Hirschhorn, J. N. (2003). Genetic epidemiology of type 1 diabetes. *Pediatric diabetes*, *4*(2), 87-100.

Holt, R. I. (2004). Diagnosis, epidemiology and pathogenesis of diabetes mellitus: an update for psychiatrists. *The British Journal of Psychiatry*,*184*(47), s55-s63.

Honeyman, M. C., Coulson, B. S., Stone, N. L., Gellert, S. A., Goldwater, P. N., Steele, C. E., ... and Harrison, L. C. (2000). Association between rotavirus infection and pancreatic islet autoimmunity in children at risk of developing type 1 diabetes. *Diabetes*, *49*(8), 1319-1324.

Hyoty, H., Hiltunen, M., Knip, M., Laakkonen, M., Vahasalo, P., Karjalainen, J., …and Akerblom, H. K. (1995). A prospective study of the role of coxsackie B and other enterovirusinfections in the pathogenesis of IDDM. *Diabetes*,*44*(6), 652-658.

Hyöty, H., Hiltunen, M., Reunanen, A., Leinikki, P., Vesikari, T., Lounamaa, R., ...and Fagerlund, A. (1993). Decline of mumps antibodies in type 1 (insulin-dependent) diabetic children and a plateau in the rising incidence of type 1 diabetes after introduction of the mumps-measles-rubella vaccine in Finland. *Diabetologia*, *36*(12), 1303-1308.

Islam, M. A., and Chowdhury, R. I. (2006). A higher order Markov model for analyzing covariate dependence. *Applied Mathematical Modelling*, *30*(6), 477-488.

Josephat, P., and Ismail, A. (2012). A Logistic Regression Model of Customer Satisfaction of Airline. *International Journal of Human Resource Studies*,*2*(4), 197.

Karvonen, M., Tuomilehto, J., Libman, I., & LaPorte, R. (1993). A review of the recent epidemiological data on the worldwide incidence of type 1 (insulin-dependent) diabetes mellitus. *Diabetologia*, *36*(10), 883-892.

Kolb, H., and Pozzilli, P. (1999). Cow's milk and type I diabetes: the gut immune system deserves attention. *Immunology Today*, *20*(3), 108-110.

Kosorok, M. R. (1993). *A longitudinal mode] for binary response data with time-dependent covariates*. Technical Report 84, University of Wisconsin, Madison, Dept. of Biostatistics

Kyvik, K. O., Nystrom, L., Gorus, F., Songini, M., Oestman, J., Castell, C., ...and Michalkova, D. (2004). The epidemiology of type 1 diabetes mellitus is not the same in young adults as in children. *Diabetologia*, *47*(3), 377-384.

Laird, N. M., and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963-974.

Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Lipsitz, S. R., Kim, K., and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in medicine*, *13*(11), 1149-1163.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1992). A three-stage estimator for studies with repeated and possibly missing binary outcomes. *Applied statistics*, 203-213.

Lyssenko, V., Jonsson, A., Almgren, P., Pulizzi, N., Isomaa, B., Tuomi, T., ...and Groop, L. (2008). Clinical risk factors, DNA variants, and the development of type 2 diabetes. *New England Journal of Medicine*, *359*(21), 2220-2232.

Manning, C. (2007). Generalized Linear Mixed Models (illustrated with R on Bresnan et al.'s datives data).

McCullagh, P., and Nelder, J. A. (1989). *Generalized linear models* (Vol. 37).CRC press.

McIntosh, E. D. G., and Menser, M. A. (1992).A fifty-year follow-up of congenital rubella. *The Lancet*, *340*(8816), 414-415.

Mehta, S. R., Kashyap, A. S., & Das, S. (2009). Diabetes mellitus in India: The modern scourge. *Medical journal armed forces India*, *65*(1), 50-54.

Muenz, L. R., and Rubinstein, L. V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics*, 91-101.

Onkamo, P., Väänänen, S., Karvonen, M., & Tuomilehto, J. (1999). Worldwide increase in incidence of Type I diabetes–the analysis of the data on published incidence trends. Diabetologia, 42(12), 1395-1403.

Pak, C., Mcarthur, R., Eun, H. M., and Yoon, J. W. (1988). Association of cytomegalovirus infection with autoimmune type 1 diabetes. *The Lancet*,*332*(8601), 1-4.

Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine*, *12*(18),1723-1732.

Peer, N., Kengne, A. P., Motala, A. A., and Mbanya, J. C. (2014). Diabetes in the Africa Region: an update. *Diabetes research and clinical practice*, *103*(2), 197-205.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 1033-1048.

Rahman, M. S., & Islam, M. A. (2007). Markov structure based logistic regression for repeated measures: An application to diabetes mellitus data. *Statistical Methodology*, *4*(4), 448-460.

Schlesselman, J. J. (1982). Basic methods of analysis. *Case-control studies: design, conduct, analysis*, 171-220.

Sperling, M. A. (2003). *Type 1 diabetes: etiology and treatment*. Springer Science and Business Media.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 961-971.

Tabachnick, B.G., and Fidell, L. S. (2007). *Experimental designs using ANOVA*. Thomson/Brooks/Cole.

Ware, J. H., Lipsitz, S., and Speizer, F. E. (1988). Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine*, *7*(1-2), 95-107.

Whiting, D. R., Guariguata, L., Weil, C., and Shaw, J. (2011). IDF diabetes atlas: global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes research and clinical practice*, *94*(3), 311-321.

Wild, S., Roglic, G., Green, A., Sicree, R., and King, H. (2004). Global prevalence of diabetes estimates for the year 2000 and projections for 2030.*Diabetes care*, *27*(5), 1047-1053.

World Health Organization. (1999). Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus.

World Health Organization. (2009). *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization.

World Health Organization. (2011). Waist circumference and waist-hip ratio: Report of a WHO expert consultation, Geneva, 8-11 December 2008.

World Health Organization.(2016). *Global report on diabetes*.World Health Organization.

Wu, Y., Ding, Y., Tanaka, Y., and Zhang, W. (2014). Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention.*International journal of medicalsciences, 11*(11), 1185.

Zeger, S. L., and Qaqish, B. (1988). Markov regression models for time series: a quasi-likelihood approach. *Biometrics*, 1019-1031.

Zimmet, P. Z. (1992). Kelly West Lecture 1991 challenges in diabetes epidemiology—from West to the rest. *Diabetes care*, *15*(2), 232-252.

APPENDIX A

DESCRIPTIVE STATISTICS

FBScode * Sex * Visit Cross tabulation

Count

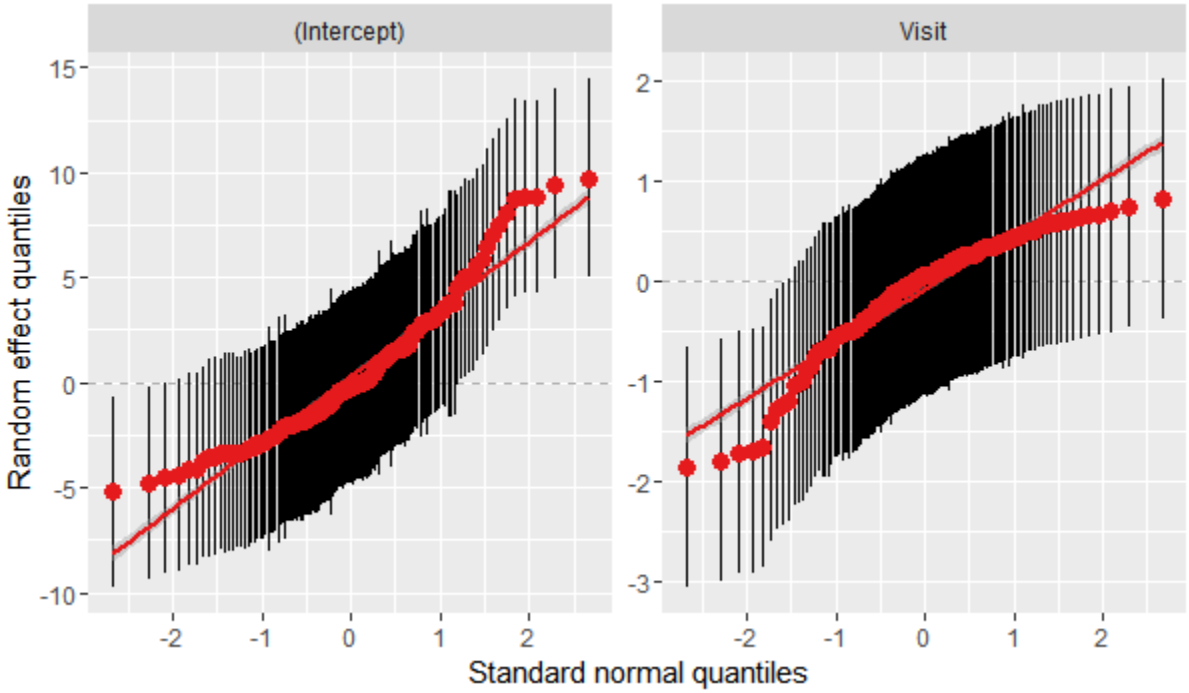
Visit			Sex		Total
			Female	Male	
1	FBScode	FBS<=7,0	40	10	50
		FBS>7.0	88	17	105
	Total		128	27	155
2	FBScode	FBS<=7,0	37	8	45
		FBS>7.0	80	17	97
	Total		117	25	142
3	FBScode	FBS<=7,0	28	9	37
		FBS>7.0	62	10	72
	Total		90	19	109
4	FBScode	FBS<=7,0	19	5	24
		FBS>7.0	58	7	65
	Total		77	12	89
Total	FBScode	FBS<=7,0	124	32	156
		FBS>7.0	288	51	339
	Total		412	83	495

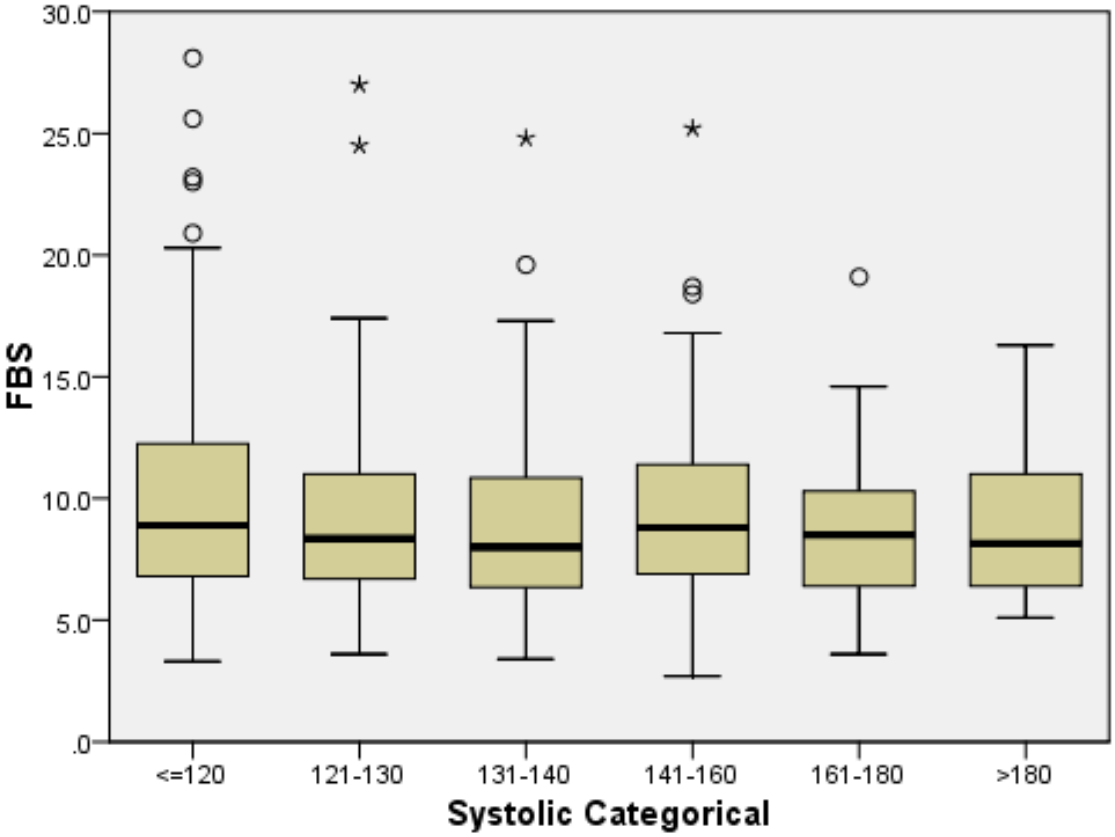
Visit

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	155	31.3	31.3	31.3
	2	142	28.7	28.7	60.0
	3	109	22.0	22.0	82.0
	4	89	18.0	18.0	100.0
	Total	495	100.0	100.0	

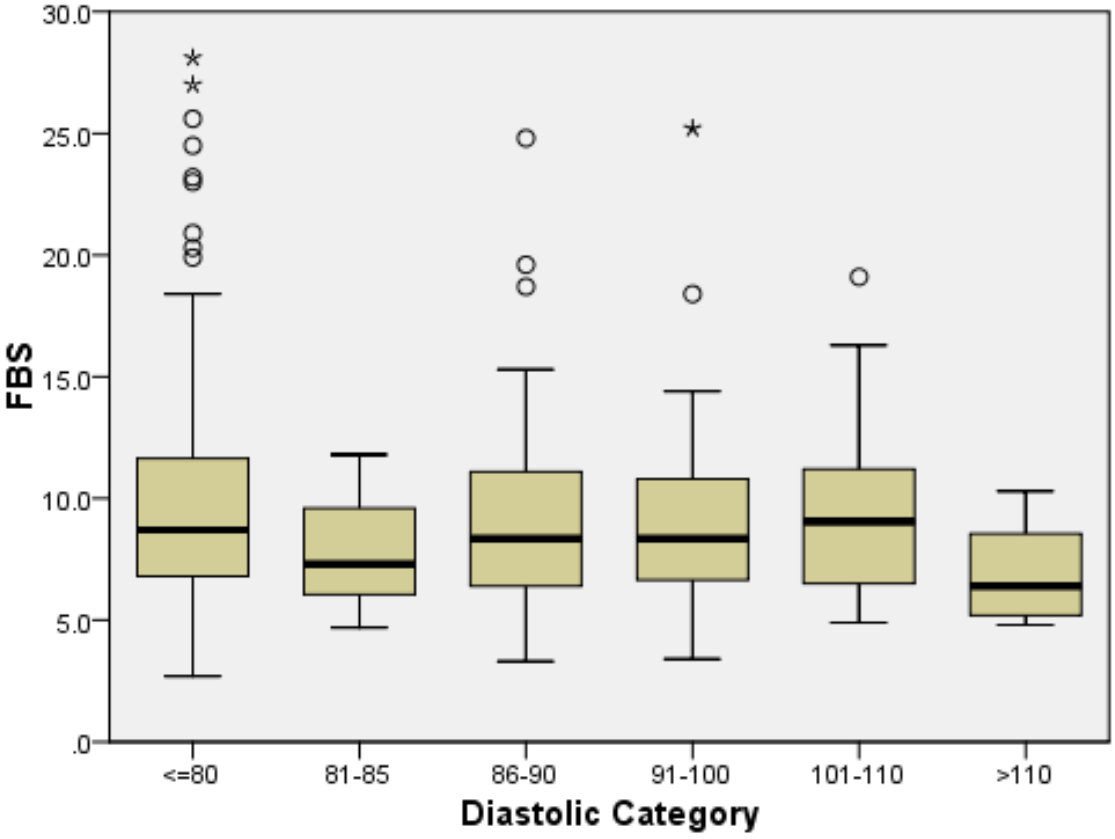
APPENDIX B

GRAPHICAL PRESENTATION OF DATA

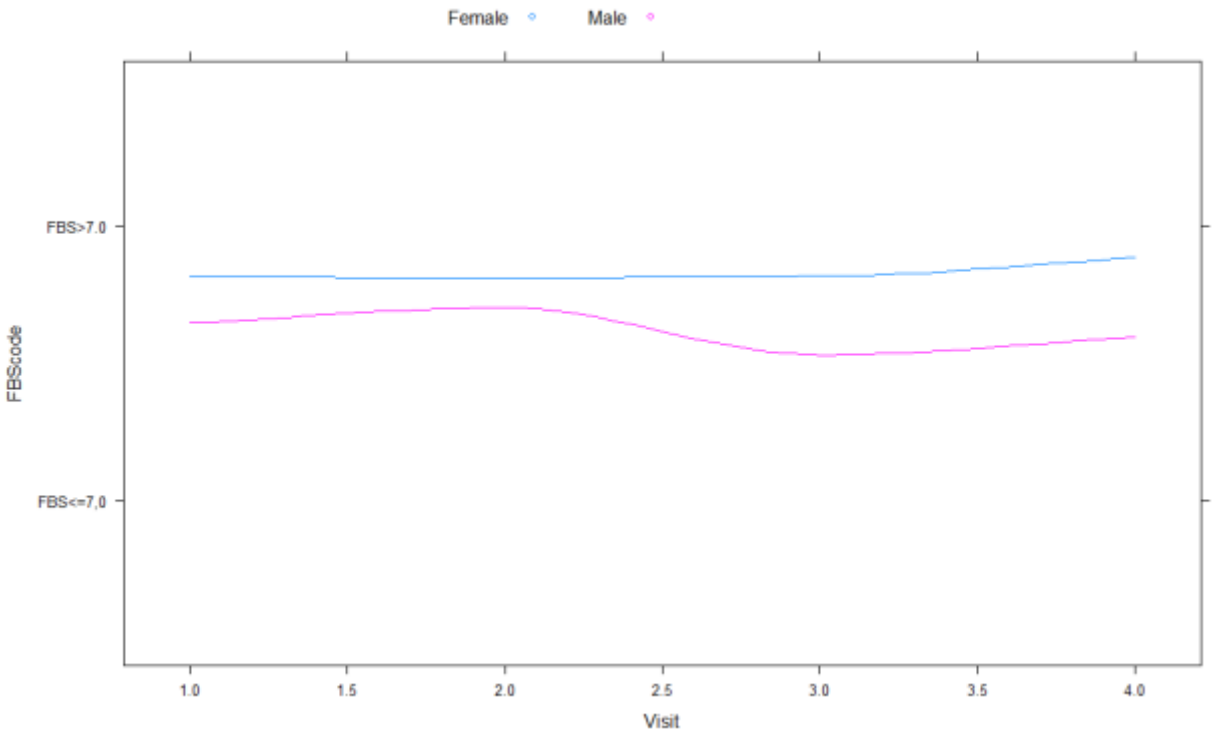


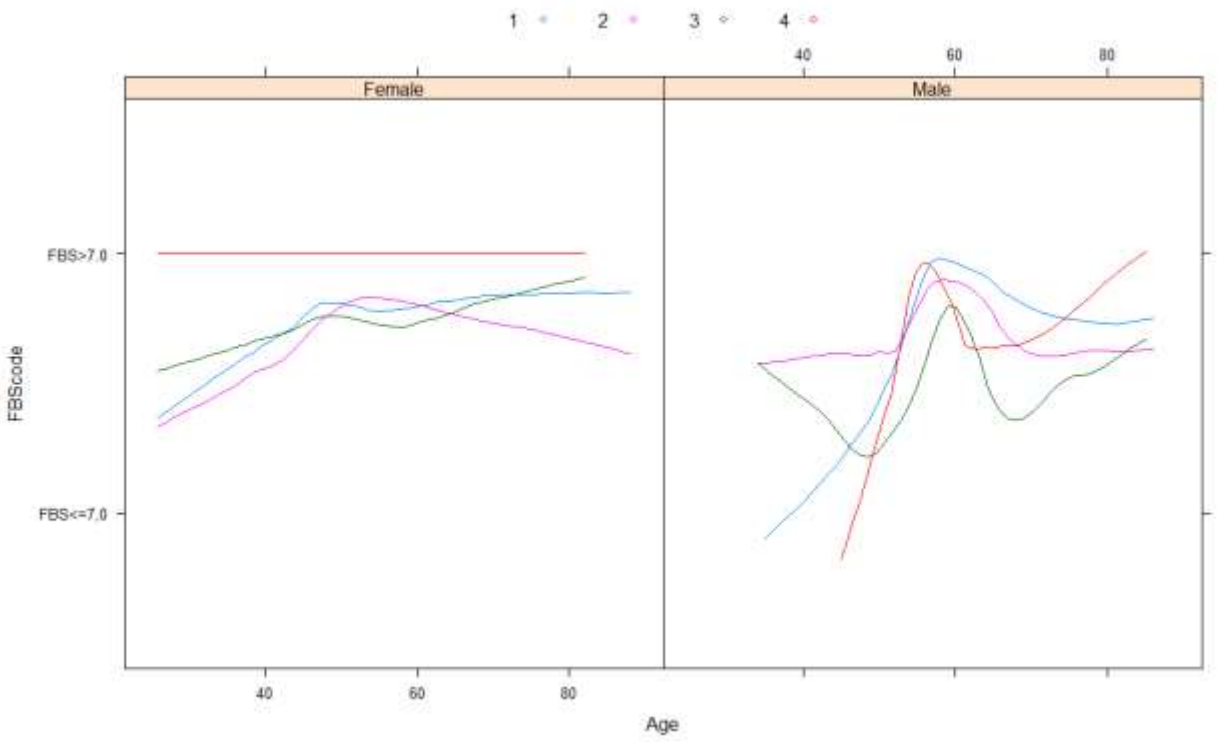


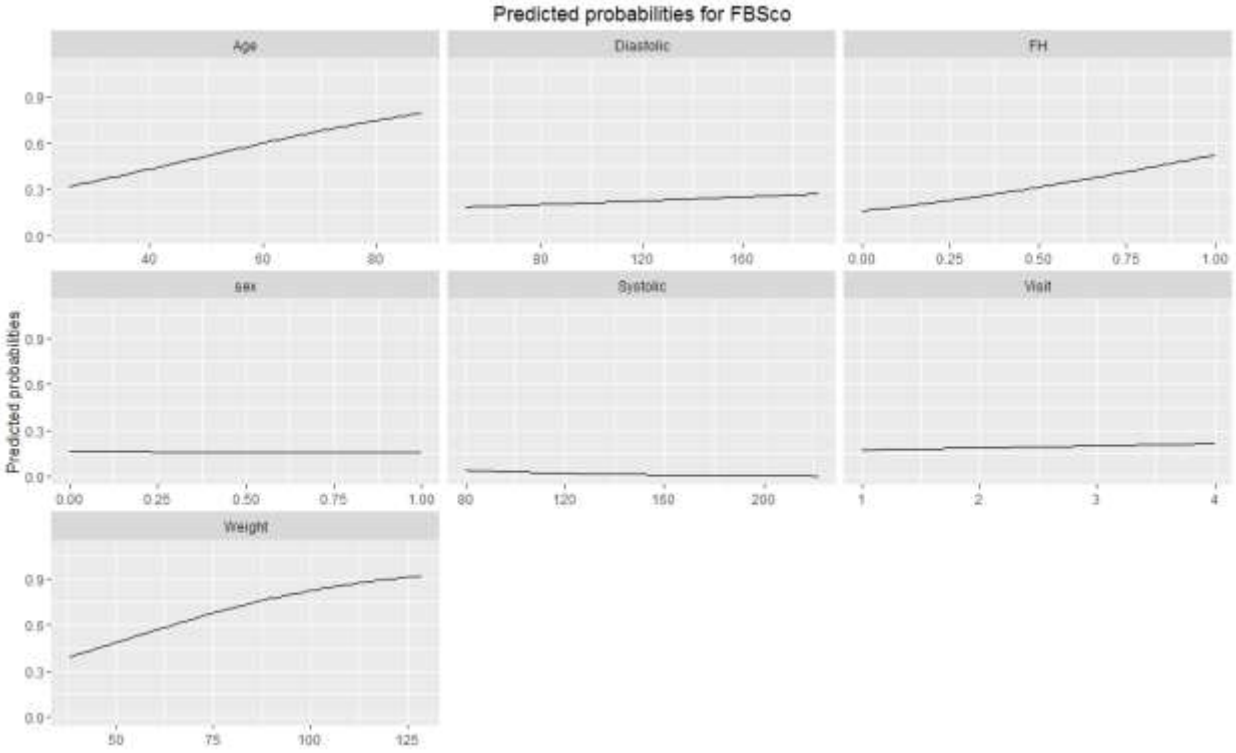
Boxplot of FBS against Systolic blood pressure



Boxplot of FBS against Diastolic blood pressure


Female
Male
FBS>7.0
FBS<=7.0
FBScode
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Visit


1
2
3
4
Female
Male
FBS>7.0
FBS<=7.0
FBScode
40
60
80
Age


Predicted probabilities for FBSco
Age
Diastolic
FH
Sex
Systolic
Visit
Weight
Predicted probabilities
0.0
0.3
0.6
0.9
40
60
80
80
120
160
0.00
0.25
0.50
0.75
1.00
200
1
2
3
4
50
75
100
125

APPENDIX C

R CODES

```
library(foreign)
dab<-read.spss("c:\\Users\\Even\\Desktop\\project data.sav",to.data.frame = TRUE)
head(dab)
Agedata<-dab$Age
AgCat<-dab$AgeCat
Wdata<-dab$Weight
Sdata<-dab$Systolic
Sexdata<-dab$sex
Ddata<-dab$Diastolic
FHdata<-dab$FH
FBSdata<-dab$FBS
Sexcode<-ifelse(Sexdata=="Male",1,0)
FBSco<-ifelse(FBSdata<7.1,0,1)
FHcode<-ifelse(FHdata=="Yes",1,0)
library(Matrix)
library(MASS)
library(nlme)
library(lme4)

library(languageR)

library(optimx)
library(arm)
library(ggplot2)
library("bbmle")
library("gridExtra")
library(gee)
library(car)
##GLMM model
m1<-glmer(FBSco~Age+sex+Systolic+Weight+Diastolic+FH+Visit+(1+Visit|SN),family=bino
mial,data=dab)

summary(m1)
Anova(m1)
plot(m1)

#GEE model with excheangeable correlation structure
m11<-gee(FBSco~Age+sex+Systolic+Weight+Diastolic+FH+Visit,family=binomial,id=SN,data
=dab,corstr = "exchangeable")
summary(m11)

#computing the p-value#s using normal approximation
2*pnorm(abs(coef(summary(m11))[,5]),lower.tail=FALSE)
```

```
#GEE model with independence correlation structure
m12<-gee(FBSco~Age+sex+Systolic+Weight+Diastolic+FH+Visit,family=binomial,id=SN,data
=dab,corstr = "independence")
summary(m12)
#computing the p-value#s using normal approximation
2*pnorm(abs(coef(summary(m12))[,5]),lower.tail=FALSE)

#GEE model with independence correlation structure
m13<-gee(FBSco~Age+sex+Systolic+Weight+Diastolic+FH+Visit,family=binomial,id=SN,data
=dab,corstr = "unstructured")
summary(m13)

#computing the p-value#s using normal approximation
2*pnorm(abs(coef(summary(m13))[,5]),lower.tail=FALSE)

#Finding isolated systolic hypertension
Systoliccode<-ifelse(dab$Systolic>100,"Normal","Not normal")
Diastoliccode<-ifelse(dab$Diastolic>60,"Normal","Not normal")
table(Systoliccode,Diastoliccode)

#Isolated Diastolic hypertension
Syscode<-ifelse(dab$Systolic<140,"Normal","Not normal")
Diascode<-ifelse(dab$Diastolic<90,"Normal","Not normal")
table(Syscode,Diascode)

## random intercept and slope
ggplot(m1,aes(x = Visit,y=FBS, group=SN,colour=SN))+geom_point()+geom_path()

#check normality of the random effects
reStack<- stack(ranef(m1)$SN)
qq<-print(qqmath(~values | ind, reStack, scales = list(relation = "free")))

library(lattice)
xyplot(FBScode ~ Age|sex, group = Visit, data = dab, type = "smooth", auto.key = list(columns
= 4))
xyplot(FBScode ~ Visit, group = sex, data = dab, type = "smooth", auto.key = list(columns = 4))

library(sjPlot)
mod3<-glmer(FBScode~Systolic+Visit+(1+Visit|SN),family=binomial,data=dab)
summary(mod3)
pred<-coef(m1)$SN

#plot fixed effects correlation matrix
sjp.glmer(m1, type = "fe.cor")
```

```
## plotqq-plot of random effects
sjp.glmer(m1, type = "re.qq")

# plot probability curve of fixed effects
sjp.glmer(m1, type = "fe.pc")
```

APPENDIX D

R - OUTPUTS

GLMM

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMo
d']
 Family: binomial  (logit )
Formula: FBSco ~ Age + sex + Systolic + Weight + Diastolic + FH + Visit +      (1 + Visit | SN)
   Data: dab

     AIC      BIC   logLik deviance df.resid
   531.7    578.0   -254.9    509.7      484

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2750 -0.4469  0.2072  0.3762  1.7576

Random effects:
 Groups Name        Variance  Std.Dev. Corr
 SN     (Intercept) 4.7838805 2.18721
        Visit       0.0004909 0.02216  -0.96
Number of obs: 495, groups:  SN, 155

Fixed effects:
             Estimate     Std. Error    z value     Pr(>|z|)
(Intercept) -1.658412    2.026125    -0.819      0.413064
Age          0.034454      0.018038    1.910       0.056125 .
sexMale     -0.027736    0.643365  -0.043      0.965613
Systolic    -0.018861     0.008594  -2.195      0.028190 *
Weight       0.032034     0.016225    1.974      0.048339 *
Diastolic    0.003527     0.015164    0.233      0.816072
FHYes        1.750174     0.503110    3.479      0.000504 ***
Visit        0.088447       0.162353    0.545      0.585903
---
Signif.codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
        (Intr) Age    sexMalSystlc Weight DistlcFHYes
Age         -0.549
sexMale   -0.178 -0.001
Systolic  -0.187 -0.158  0.024
Weight    -0.534  0.139  0.173 -0.134
Diastolic -0.326  0.065 -0.053 -0.431 -0.103
FHYes     -0.151  0.062  0.139 -0.024  0.059 -0.050
```

Visit -0.303 0.068 0.145 0.043 0.067 0.004 0.150

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link: Logit
Variance to Mean Relation: Binomial
Correlation Structure: Exchangeable

Call:
gee(formula = FBSco ~ Age + sex + Systolic + Weight + Diastolic +
FH + Visit, id = SN, data = dab, family = binomial, corstr = "exchangeable")

Summary of Residuals:
Min 1Q Median 3Q Max
-0.8870030 -0.4847071 0.1785297 0.3265787 0.6094144

Coefficients:
Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept) -0.945358707 1.186714548 -0.79661845 1.017074911 -0.92948779
Age 0.020529770 0.010757328 1.90844516 0.010365035 1.98067541
sexMale -0.034935741 0.377166291 -0.09262689 0.379559878 -0.09204277
Systolic -0.011499501 0.005172563 -2.22317271 0.005094233 -2.25735675
Weight 0.018478181 0.009851871 1.87560126 0.009174858 2.01400182
Diastolic 0.002219909 0.009285400 0.23907525 0.008152506 0.27229779
FHYes 0.995341619 0.288455556 3.45058918 0.297494732 3.34574536
Visit 0.059233647 0.078887608 0.75086124 0.068984906 0.85864649

Estimated Scale Parameter: 1.034803
Number of Iterations: 4

Working Correlation
[,1] [,2] [,3] [,4]
[1,] 1.0000000 0.3848591 0.3848591 0.3848591
[2,] 0.3848591 1.0000000 0.3848591 0.3848591
[3,] 0.3848591 0.3848591 1.0000000 0.3848591
[4,] 0.3848591 0.3848591 0.3848591 1.0000000

2*pnorm(abs(coef(summary(m11))[,5]),lower.tail=FALSE)
(Intercept) Age sexMale Systolic Weight Diastolic FHYes
0.3526363525 0.0476276860 0.9266640610 0.0239857918 0.0440093420 0.7853930529 0.0008
206172

```
      Visit
0.3905355823
```

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

```
Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Independent

Call:
gee(formula = FBSco ~ Age + sex + Systolic + Weight + Diastolic +
    FH + Visit, id = SN, data = dab, family = binomial, corstr = "independence")

Summary of Residuals:
      Min        1Q    Median        3Q       Max
-0.8680544 -0.4857960  0.1843237  0.3141567  0.5875627


Coefficients:
                 Estimate  Naive S.E.      Naive z  Robust S.E.   Robust z
(Intercept) -0.927230498 0.962320732 -0.9635358 1.046049721 -0.8864115
Age          0.020597372 0.007903408  2.6061381 0.010120957  2.0351210
sexMale     -0.125650823 0.274491655 -0.4577583 0.366988790 -0.3423833
Systolic    -0.010120826 0.005235613 -1.9330735 0.005533413 -1.8290384
Weight       0.012390136 0.007069768  1.7525520 0.008653379  1.4318264
Diastolic    0.004917097 0.010018575  0.4907981 0.009692205  0.5073249
FHYes        1.043402945 0.210524633  4.9562036 0.290099120  3.5967119
Visit        0.063952610 0.095731873  0.6680389 0.075122315  0.8513131

Estimated Scale Parameter:  1.028583
Number of Iterations:  1

Working Correlation
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
```

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link: Logit
 Variance to Mean Relation: Binomial
 Correlation Structure: Unstructured

Call:
gee(formula = FBSco ~ Age + sex + Systolic + Weight + Diastolic +
   FH + Visit, id = SN, data = dab, family = binomial, corstr = "unstructured")

Summary of Residuals:
      Min   1Q   Median   3Q   Max
-0.8860311 -0.4892956 0.1831702 0.3240065 0.6086345

Coefficients:
                Estimate   Naive S.E.   Naive z   Robust S.E.   Robust z
(Intercept) -0.686610327 1.140963333 -0.601781238 1.010310093 -0.6796036
Age   0.021091163 0.010168903 2.074084415 0.010518672 2.0051166
sexMale   -0.071374399 0.356462871 -0.200229548 0.373409406 -0.1911425
Systolic   -0.011345328 0.005132484 -2.210494795 0.004969548 -2.2829696
Weight   0.017299921 0.009247228 1.870822403 0.008818077 1.9618701
Diastolic   0.000092221 0.009298679 0.009917645 0.007766438 0.0118743
FHYes   0.965376653 0.271768104 3.552207338 0.295668894 3.2650599
Visit   0.047455579 0.081396255 0.583019193 0.068892748 0.6888327

Estimated Scale Parameter: 1.030972
Number of Iterations: 4

Working Correlation
          [,1]   [,2]   [,3]   [,4]
[1,] 1.0000000 0.4063093 0.4124650 0.2754958
[2,] 0.4063093 1.0000000 0.2223296 0.1471088
[3,] 0.4124650 0.2223296 1.0000000 0.2396417
[4,] 0.2754958 0.1471088 0.2396417 1.0000000

APPENDIX E

SPSS CODING SCHEME

Coding scheme for data entry

Variable	**Meaning**	**Codes**
Sex	Female	0
	Male	1
Age Category	<30	0
	30-39	1
	40-49	2
	50-59	3
	60-69	4
	70+	5
Educational level	Below secondary	0
	Secondary and higher	1
Occupation of patients	Unemployed	0
	Informal sector	1
	Formal sector	2
FBS	≤7.0	0
	>7.0	1
Family history of diabetes	Non – positive family history	0
	Positive family history	1
Visit	First visit	1

	Second visit	2
	Third visit	3
	Fourth visit	4
Systolic blood pressure category	≤120	0
	121-130	1
	131-140	2
	141-160	3
	161-180	4
	>180	5
Diastolic blood pressure category	≤80	0
	81-85	1
	86-90	2
	91-100	3
	101-110	4
	>110	5