# Developing Longitudinal Models for Monitoring Chronic Diseases in Computerised General Practice (GP) Records: A case study in Chronic Kidney Disease (CKD)

By

**ZALIHE YARKINER**

**Thesis**

Submitted in partial fulfilment of the requirements

of Kingston University for the award of

**Doctor of Philosophy in Applied Statistics**

**Faculty of Science, Engineering and Computing**

**School of Mathematics**

**Kingston University, London**

September 2015

## Acknowledgement

I would like to express my special thanks to my director of studies Dr. Rosie O'Neil, my second supervisor Dr. Gordon Hunter and my third supervisor Prof. Simon de Lusignan for their valuable suggestions, encouragement, commitment and continual enthusiasm through this research. I would also like to thank to Dr. Penelope Bidgood for her support and advice, particularly during the first year of this project.

In addition, I would like to express my gratitude to my family, specifically to my mum; Pembe Yarkiner, my dad; Ozer Yarkiner, my sister; Ozce Nasiboglu and my brother; Mehmet Nasiboglu for their support and encouragement during this research.

Finally, I would also like to express my appreciations to my friends who have encouraged me during my study.

## Declaration

I declare that this thesis is completely my own work and none of this work has been submitted for a degree at another university.

The data used in this research project was collected from General Practices over England and Wales and the data is provided by Prof. Simon de Lusignan.

## Abstract

Analysis of longitudinal data is a rapidly growing field of statistical analysis, in response to the increasing availability of longitudinal datasets in many disciplines. Longitudinal studies are becoming more popular as they allow investigation of the same individuals over time, and where both within-individual and between-individual differences can be examined. Since the study of change over time is necessary in many areas, longitudinal studies and meaningful analysis of longitudinal data is essential. The health sector is one such area where longitudinal research is playing an increasingly important role.

The aim of this research is to examine statistical methodologies for the analysis of longitudinal medical data, specifically General Practice (GP) records. All General Practices (GPs) in England and Wales are now computerized and routinely record detailed patient information, hence providing a rich longitudinal dataset. This research investigates new techniques and adaptations of existing methodologies to understand and explain patterns of change and the natural development and treatment of chronic diseases within routinely collected GP data. The data used here, although taken from a raw sample of 129 General Practice records, have been subjected to some cleaning and recoding in places, hence it should be considered as a secondary data source. Through out the data driven applications presented, different sub-samples of the original dataset have been used. For the main part the full cleaned sample of 876951 patients is used where possible. Smaller samples ranging between 472 and 58675 patients are used depending on the outcome of interest and the availability of valid observations for the various applications employed.

Mainly regression-based techniques, in two broad categories, were used to analyse the repeated measurements from each patient in our dataset. Firstly, linear and generalized mixed modelling approaches were used, whereas in the second phase of the project, the applications of semi-parametric and non-parametric approaches were investigated. The case study of particular interest in this research project is the incidence and progression of chronic kidney disease (CKD). There is a lack of knowledge and understanding of the natural history of CKD and its progression over time. This project aims to address these issues.

The advanced statistical models used in this research quantify how kidney function, assessed using estimated Glomerular Filtration Rate (eGFR), changes with respect to time and how other factors, including other related medical conditions (known as co-morbidities of CKD), affect kidney function and its change over time. The techniques and approaches used in this study are motivated by mixed model designs. The decline of kidney function as time progresses for typical CKD patients is observed to be non-linear. The type of nonlinear mixed models developed in this project do not assume that the decline of eGFR over time is linear, and hence are better able to model the progression of CKD than more traditional linear models. As a consequence, the proportion of the total variation in the outcome that can be explained by considering the patient level factors is tripled through the use of these non-linear models, showing they have much greater explanatory power than previous, simpler statistical models. The disease under study is Chronic Kidney Disease (CKD) but the methodologies should also be applicable other chronic, progressive diseases.

## Contribution to knowledge

This study combines two research aspects and will contribute to both the relevant medical and statistical areas of research. In the health sector, detailed longitudinal data on an individual patient basis is widely available but analysis of such data up to now has been minimal as examination of large and complex longitudinal datasets can be difficult, expensive and time consuming (Crinson *et al.*, 2010). However there is a very great potential for such data to inform knowledge, understanding and practice within the Health sector. In addition, there is scope to link GP records with other areas of health service and use provisions, such as laboratory tests, hospital admissions, etc. to increase understanding of the nature of diseases, and treatment procedures. In recent years, National Health Service (NHS) managers in the UK and health researchers have recognized the potential benefits of using routinely collected data to improve analysis in determining what factors contribute to success in disease diagnosis, treatment and management (De Lusignan *et al.*, 2010).

More generally and across other disciplines, much research has been conducted using cross-sectional studies and the advent of longitudinal studies is relatively new in statistical research. Therefore, methodologies for analysing longitudinal data are not well established. Developing new statistical modelling frameworks for application to such data is the originality of this research which will enable researchers to interpret such data in a reliable and credible way. The development of such a modelling framework for Chronic Kidney Disease (CKD) will enable the identification of the association of cause, effect and gradual change between renal function, co morbidities and various health indicators. This will allow earlier diagnosis, appropriate and timely treatment, of the disease resulting in slowing down the progression of the disease to the end stage which will extend the lifetime and health of patients. Successful application of this modelling framework will help to increase the efficiency of health organisations. This modelling framework could then also be applied to other chronic diseases such as asthma, arthritis, cardiac failure, diabetes mellitus types 1&2, which are the most common types of chronic diseases in England and Wales.

## Table of Contents

Table of Contents

## List of Tables

## List of Figures

# Glossary of terms

ACE – Angiotension I-Converting Enzyme

ACR – Albumin/Creatinine Ratio

AIC – Akaike Information Criterion

AM – Additive Model

ANCOVA – Analysis of Co-variance

ANOVA – Analysis of Variance

AR-1 – First Order Autoregressive

ARB – Angiotension Receptor Blockers

BayesX – Bayesian Perspective of Generalized Regression

BIC – Bayesian Information Criteria

BMI – Body Mass Index

BP – Blood Pressure

BS – Base Spline

BSS – Between Sums of Squares

CEBVD – Cerebrovascular Disease

CG – Cockcroft and Gault

CHD – Coronary Heart Disease

CI – Confidence Interval

CKD – Chronic Kidney Disease

CVD – Cardiovascular Disease

DBP – Diastolic Blood Pressure

df – degrees of freedom

DIC – Deviance Information Criterion

DPQL – Double Penalized Quasi-Likelihood

EDA – Exploratory Data Analysis

edf – expected degrees of freedom

eGFR – Estimated Glomerular Filtration Rate

EM – Expectation-Maximization

EPO – Erythropoietin

ESRD – End Stage Renal Disease

EXP(B) – Expected Beta Parameter which is interpreted as odds ratio

GAM – Generalized Additive Model

GAMM – Generalized Additive Mixed Models

GCV – Generalized Cross-Validation

GEE – Generalized Estimating Equations

GEM – Generalized Expectation-Maximization

GFR – Glomerular Filtration Rate

GLM – Generalized Linear Model

GLMM – Generalized Linear Mixed Model

GLS – Generalized Least Square

GP – General Practice

Hb – Haemoglobin

HbA1c – Glycated Haemoglobin

HDL – High Density Lipoprotein

HRT – Hormone Replacement Therapy

HSE – Health Survey in England

HUNTII – Health Survey of Nord-Trondelag Country

ICC – Intra-class Correlation Coefficient

IDMS – Isotope Dilution Mass Spectrometry

IHD – Ischemic Heart Disease

KDIGO – Kidney Disease: Improving Global Outcomes

KDOQI – Kidney Disease Outcomes Quality Initiative

LDL – Low Density Lipoprotein

LL – Log-Likelihood

LM – Linear Model

LMM – Linear Mixed Model

LVH – Left Ventricular Hypertrophy

MANOVA – Multivariate Analysis of Variance

MAR – Missing at Random

MCAR – Missing Completely at Random

MCMC – Markov Chain Monte Carlo

MCV – Mean Corpuscular Volume

MDRD – Modification of Diet in Renal Disease

MGCV – Mixed GAM Computation Vehicle

ML – Maximum Likelihood

MLE – Maximum Likelihood Estimation

**Glossary of terms**

MRFIT – Multiple Risk Factors Intervention Trial

NHANES – National Health and Nutrition Examination Survey

NICE – National Institute for Health and Clinical Excellence

NKF – National Kidney Foundation

NR – Newton-Raphson

NSAID – Non-Steroidal Anti-Inflammatory Drugs

OLS – Ordinary Least Square

OR – Odds Ratio

PCR – Protein/Creatinine Ratio

PLS – Penalized Least Square

PQL – Penalized Quasi-Likelihood

ps – P-Spline

PVD – Peripheral Vascular Disease

QOF – Quality Outcomes Framework

QQ-Plots – Quantile-Quantile Plots

R2BayesX – R interface that use estimate structured additive regression models with BayesX methodology

Ref. df – Reference degrees of freedom

REML – Restricted Estimation of Maximum Likelihood

RRT – Renal Replacement Therapy

SBP – Systolic Blood Pressure

SCr – Serum Creatinine

SD – Standard Deviation

## Glossary of terms

SE – Standard Error

Sig – Significance value

SIGN – Scottish Intercollegiate Guidelines Network

SPSS – Statistical Software for Social Sciences

STAR – Structured Additive Regression Models

$t_0$ – Time zero

TG – Serum Triglycerides

TSS – Total Sums of Squares

UK – United Kingdom

UKPDS – United Kingdom Prospective Diabetes Study

USA – United States of America

USRDS – United States Renal Data System

WHR – Waist: Hip Ratio

WLS – Weighted Least Square

WSS – Within Sums of Squares

# Chapter 1 – Introduction

## 1 Introduction

This research study focuses on developing and evaluating longitudinal statistical models to analyse computerised general practice (GP) records. Nowadays, most General Practices (GPs) in the UK use a computerised system in order to keep records of their patients and hence each registered patient should have a historical, detailed and complete record of all visits made to their GP. This means that there is a vast amount of detailed and reliable information about the health of the nation readily available at little or no cost. In this thesis, the research questions are: can we use routinely collected GP records to inform research and practice in the medical field and, subsequently, how can these be achieved and what are the best methodologies to extract meaningful interpretations from the available data? We address these questions by focusing specifically on developing understanding of the progression of chronic diseases presented and treated in the primary care setting in particular, chronic kidney disease (CKD).

The popularity of analysing repeated measurements within data has been increasing since the mid-1980s due to the development of suitable methodologies designed to enable investigations of change observed in an outcome over time, and also the factors affecting this change. However, practical applications of repeated measures techniques exhibit several challenges, such as requiring the data to meet with strict analytical assumptions, for example requiring the response to follow a normal distribution and having equal correlations between every pair of successive repeated measurements. Additionally, there are other factors that occur outside of the control of the researcher which in turn cause potential problems in such research. These problems include patient drop-outs, due to moving home, death or refusal to co-operate, that create missing observations within the data. Furthermore, different patients can have different number of repeated measurements and repeated measurements might not be taken at equal time intervals. These issues restrict the use of available traditional methods for the analysis of repeated measurements, such as ANOVA techniques, which have strict assumptions. Therefore special statistical techniques have been developed to analyse such repeated measures designs.

The main aim of this research is to evaluate the potential of the application of statistical methodologies for the analysis of routinely collected longitudinal health data, specifically

General Practice (GP) records. More specifically the research aims to investigate emerging techniques or adaptations of existing methodologies to understand and explain patterns of change and the influence of key factors on the natural progression of chronic diseases using GP data. The research presented here uses regression based techniques including logistic, linear mixed modelling, semi-parametric and non-parametric modelling approaches to analyse repeated measurements from within GP records of a large sample of patients.

Our case study for this research is based on statistical modelling of the progression of Chronic Kidney Disease (CKD), a disease which is progressive and leads to end stage renal failure. In this area of medicine, there is a lack of knowledge of the natural history of CKD and on the progression of CKD over time. Since there are no obvious early symptoms of the disease, CKD is most commonly diagnosed when in its later more serious stages. Early diagnosis of CKD is currently very rare and it is a clinical area that is recognised as needing further research and understanding.

Data was obtained from all patients registered with 129 GPs throughout England and Wales. Since this data is a sample from the population of the UK, the first question that rises relates to the validity and reliability of this sample. A comparison of the demographics of this data against the UK census data (see chapter 3) confirmed that the data is a good representation of the whole population. The next issue was then to look at how we can apply statistical methods to such data to analyse multiple individual series of repeated measurements to inform about the progression of chronic diseases, and CKD in particular.

Exisiting CKD literature usually assumes that the nature of the natural decline in eGFR (meaning that kidney function gets worse over time) is linear. Preliminary investigations of the data used here suggest that this is not the case. Hence a main objective of this work is to identify and apply appropriate statistical methodologies which are appropriate to the outcomes available (repeated eGFR readings), and allow further understanding of the behaviour of kidney function over time for patients with CKD.The statistical techniques used in this thesis are motivated by mixed model designs which allow for both within subject and between subject variations to be modelled simultaneously.These models are applied to patients' repeated eGFR values (measurements based on a continuous scale), where there are unequal number of repeated measures per patient, recorded at unequal time intervals.

When analysis is restricted to only patients with a CKD diagnosis, the distribution of eGFR values (i.e. the dependent variable) does not follow a normal distribution. To allow for this generalized linear mixed models are applied, transforming the outcome using a gamma distribution with log link function, to eliminate the problems that can be created by violation of normality assumption. However, using such transformations can also make the interpretation of the influence of covariates on the original outcome variable more difficult. Additionally, alternative semi-parametric and full Bayesian models, to analyse such designs from a Bayesian perspective, are considered.

The advanced statistical models applied in this thesis are used to help understand the behaviour of kidney function over time and how it is affected by other factors such as the co-existence of other diseases. The work presented in this thesis shows that by developing nonlinear mixed models, when the decline of eGFR over time is not assumed to be linear, the total variation in the outcome that can be explained by using the patient level factors is tripled. Therefore, it is vital to consider this nonlinear aspect of eGFR decline in order to truly understand the progression of CKD.

Both the theoretical background to the methods and approaches used here are valid means of modelling progression of chronic kidney disease and the results of this project indicate that these techniques would be applicable to other chronic progressive diseases which are diagnosed, treated and managed within the primary care setting. Such applications could provide valuable and beneficial information and improve understanding about the natural histories of progressive diseases, which would be of great value within the health and primary care sector.

In summary the aims of this research are investigation of new and emerging statistical methods to analyse complex longitudinal data such as GP data, illustration of the potential of using statistical methods to extract meaningful interpretations of data and demonstration of the use of such models to explain patterns of change and influence of key factors on this change. Beside of this, the objectives of this research are identification of the appropriate developing methodologies, application of suitable methods and assessing the models formed by using such methods.

## 1.1 Thesis outline

This thesis records the work carried out to achieve the aims and objectives set out above. Chapter 2 provides a review of current knowledge of the two aspects of this research; the clinical condition of CKD, and the ongoing development and growth of longitudinal data analysis from a statistical modelling perspective. Each aspect is reviewed in a separate section.

The clinical section provides details on how the kidney function is working in urinary system, how CKD condition is formed, classified and progress, what factors are affecting the progression of CKD and how the disease can be treated and manged. The main reason of this section is to get current information about CKD, so that the statistical models can be applied efficiently.

The statistical section provides details on broad range of statistical nethodologies, so that the suitable new and emerging statistical methods can be identified to analyse complex longitudinal data such as GP data.

In chapter 3, a detailed account of both the dataset used in the research and the process of data preparation and exploratory data analysis is given. The process of data preparation includes details of an intense data cleaning procedure which shows how the data is prepared so that the statistical methodologies can be applied on the data to build a realible statistical model. In chapter 3, once the data is prepared for the analysis, exploratory data analysis is carried out on the dataset, so that the findings from preliminary basic descriptive statistical analyses, provide an initial understanding of the content of the dataset. This chapter helps in understanding what the data is showing and also provides a validation of the data as a suitable and reliable sample of the population before deeper, more sophisticated, analyses are performed.

Chapter 4 examines the sizes of effects for the covariates believed to be the major influencing co-morbidities of CKD based on the literature review carried out in chapter 2. Initially, considering the outcome as having the condition of CKD or not, Logistic regression techniques are employed to evaluate the factors influencing the prevalence of CKD. Then, using the outcome as having a rapid decline or not based on the two definitions of rapid decline of CKD according to NICE guidelines (these definitions also being stated in chapter 4), Logistic regression techniques are used to identify which factors are affecting the progression of CKD. In this way, factors affecting the prevalence of CKD and factors affecting the progression of CKD can be identified separately. Additionally in chapter 4, the differences on the factors affecting the progression of CKD based on two definitions of rapid decline are identified.

In chapter 5, instead of the various binary outcomes used in the models presented in chapter 4, the outcome is now taken as the repeated eGFR measurements over time and parametric models are used to assess the progression of CKD. In chapter 5, modern, complex and emerging longitudinal models are used instead of classical approaches, so that both within and between-subject variations can be investigated simultenously, methodologies can be applied on unbalanced and incomplete datasets, allowing modelling of continuous covariates and investigation of time-dependent covariates. Initially when the response is assumed to follow a Gaussian distribution and the progression of CKD is assumed to be a linear progression, linear mixed models are analysed and developed as advanced approaches, starting with a very simple linear mixed model and, from this, building more complex models to increase the quality of model fit to the data and to further explain the variation observed in the outcome. Furthermore, more complex approaches are used when the response is assumed to follow a non-Gausian distribution but still assuming the progressiong of CKD to be a linear progression. These models include linear mixed models using a transformed outcome, generalized linear mixed models, linear mixed models using the gamma distribution with log link function and polynomial linear mixed models as a first step to take nonlinearity of the progression of CKD over time into account.

In chapter 6, since the progression of CKD is not always linear, to account for nonlinear progression of CKD, nonlinear mixed modelling is performed using generalized additive mixed models (GAMM) where, in these types of models, spline functions are used to model the data instead of assuming linearity. In this way, more precise models are obtained that explain greater proportions (about triple) of the variation in the outcome. In nonlinear mixed models, in addition to various co-morbidities, the effects of age and gender are also considered by changing the dependent variable into repeated serum creatinine measurements that are independent of age and gender, and hence the standard relationship between eGFR and serum creatinine is used to interpret the results obtained from nonlinear models in terms of the original outcome (i.e. eGFR). In this way, the effects of age and gender on the progression of CKD can be identified. Furthermore, in order to investigate the progression of CKD from a Bayesian perspective and to open up a further research, an experimental work is carried out by using alternative semi-parametric and full Bayesian models that are presented in chapter 6.

Finally, the thesis is concluded in chapter 7, where an overview of the main research findings of the project is presented and ideas for future research are suggested.

## 1.2 Computing Requirements

Models developed in this thesis are computed using Statistical software Package for Social Sciences (SPSS). IBM SPSS Statistics version 21 is a comprehensive system of analysing data. Statistics Base package was used as add-on enhancement to the full IBM SPSS Statistics system for descriptive statistics for the preliminary data analysis (in chapter 3) and Advanced Statistics package was used as add-on enhancement to the full IBM SPSS Statistics system used for the regression models of chapter 4 and 5. For the models of chapter 6, three different packages in R were used namely; package 'gamm4', package 'mgcv' and package 'R2BayesX'. Generalized additive mixed models using mgcv and lme4 (the pakage 'gamm4') was used to fit generalized additive mixed models through a version of mgcv's gamm function, using lme4 for parameter estimation (Wood and Scheipl, 2015). Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation (the pakage 'mgcv') was used in GAMs, GAMMs and other generalized ridge regression with multiple smoothing parameter estimation by REML (Wood, 2015). Estimate Structured Additive Regression Models with BayesXpackages (the package 'R2BayesX') which is an R interface to estimate structured additive regression (STAR) models with BayesX was used to fit full bayes model (Belitz *et al.*, 2012) (Umlauf *et al.*, 2015).

# 2 Background Information and Literature Review

## 2.1 Literature Review on Medical Background

### 2.1.1 Kidney Function and Urinary System

The urinary system comprises the kidneys, the ureters and the bladder. The two kidneys are situated underneath the ribs (Figure 2.1).

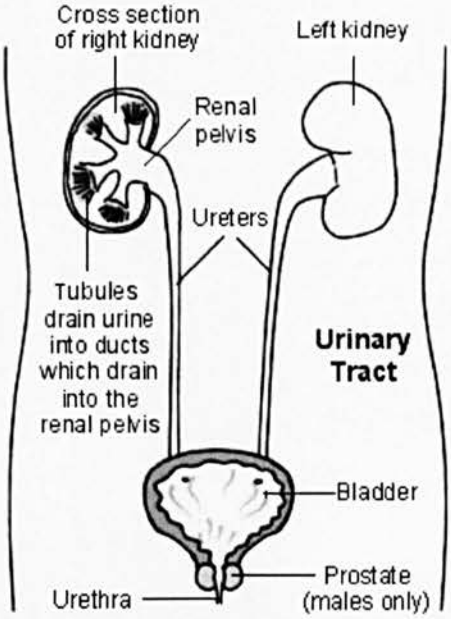

Figure 2.1: Autonomy of Kidneys and Urinary System

The main purposes of the kidneys are to remove the waste materials from the blood, and excess liquid as urine. Moreover, they excrete the existing drugs in the body, regulate salt, potassium and acid substances in the fluid content, and produce vitamin D necessary for healthier and stronger bones. Furthermore, they produce hormones, aiding the function of other organs, such as the erythropoietin hormone that is used to generate red blood cells and other hormones that are important for stabilising blood pressure and calcium metabolism.

Urea is removed from the body via filtering units called nephrons, consisting of small capillaries existing inside a ball of individual nephrons called the glomerulus. Each nephron also consists of a minor tube called a renal tubule. Urine is formed when urea mixes with water and additional materials which is then delivered into the renal tube via the aid of nephrons.

As shown in Figure 2.1, there are two ureters; the thin tubes which transport the urine from the kidneys to the bladder. Every 10 to 15 seconds, a small volume of urine is drained from the ureters into the bladder. In the case the urine is retained in the ureters, kidney infection can be observed. The bladder as an organ with a triangular shape positioned in the lower abdomen, stores the urine drained into it through the urethra. The bladder of a healthy person can store approximately two cups of urine for two to five hours. There are two sphincter muscles located in the opening of the bladder. These muscles support the outflow of urine by fitting closely near the opening of the bladder. The bladder is innervated, signalling a person when it is the time to drain the urine.

The urethra as a narrow tube enables urine to be excreted from the body. A normal urination process occurs when signals from the bladder are received by the brain, which signal to the sphincter muscles to loosen up so that the urine can pass from the urethra. Usually, kidneys filter around 200 quarts of fluid in every 24 hours and from those 200 quarts, approximately two quarts are emptied as urine and around 198 quarts are restored. The urine that is drained from the bladder can be stored in the bladder from 1 up to 8 hours.

### 2.1.2 Chronic Kidney Disease (CKD)

According to the Kidney Disease Outcomes Quality Initiative (KDOQI), chronic kidney disease (CKD) is defined as a gradual and usually permanent loss of kidney function.

The glomerular filtration rate (GFR) is defined as the total amount of fluid filtered by all of the glomeruli in a minute. Therefore, the GFR is directly proportional to the number of nephrons and to the size of the glomeruli. For children and small adults, because there will be fewer nephrons and the size of the glomeruli will be smaller compared to larger adults, a typical GFR value is expected to be lower (Rowe *et al.*, 1976). In clinical practice, GFR is used to determine the efficiency of the kidney function and is usually normalised according to body surface area e.g. by dividing the GFR by 1.73 m$^2$ in adults. This constant factor normalises GFR to body surface area and is calculated using an equation according to Du Bois and Du Bois (1996), which takes into account height and gender only. The reason for this normalisation is to take into account the differences of the height of patients. A normal GFR in young adult women or men is in the range of 100-120 mL/min (Du Bois and Du Bois, 1916).

Since direct measurement of the GFR is not possible, indirect measurement is available by using the clearance of various exogenous markers such as inulin (a fructose polysaccharide), iothalamate, and ioexhol (Australia and New Zealand Dialysis and Transplant Registry, 2011). An exogenous marker that has been accepted as the "gold standard" is inulin. However, the indirect measurement of the GFR is not commonly used because it is expensive, inconvenient and can be subject to the radioactivity or iodinated differences between these exogenous markers. Thus, in the past 35 years, a number of equations have been developed to estimate GFR (eGFR). Examples of equations used to evaluate eGFR are shown in Table 2.1, which are mostly based on age, gender and ethnicity. The endogenous marker used to evaluate eGFR in all of these equations is serum creatinine (SCr). Serum creatinine with a low molecular weight (113 Daltons) is produced as a result of a non-enzymatic reaction where creatinine in muscles is broken down at a constant rate (Levey *et al.*, 1988). The amount of creatinine resulting from this process primarily depends on the person's muscle mass, which in turn depends on body weight (Porbes and Bruining, 1976). It is established, for a given specific body weight, males normally have more muscle mass than females and additionally, African Americans and Afro-Caribbean's have more muscle mass compared with Caucasians (Cockcroft and Gault, 1976). Age is another important factor that affects the muscle mass, as in adulthood it decreases (Gallagher *et al.*, 1997). From an age of approximately 25, serum creatinine can also be affected by dietary intake factors such as meat intake or the use of a protein supplement (Levey *et al.*, 1988). Since creatinine is filtered by the glomeruli and produced by the tubular cells, total creatinine clearance is obtained when the amount of GFR and tubular secretion are combined. However, tubular secretion also depends on the kidney function. Therefore, the measurement of SCr overestimates GFR. SCr has a negative association with GFR, meaning that when the GFR is high, the SCr percentage is low (10-40% volume of blood) and when the GFR is low, the SCr percentage is high (50-60% volume of blood) (Baver *et al.*, 1982).

### 2.1.3 Estimation of Glomerular Filtration Rate (GFR)

There are two ways to estimate GFR; either by calculating the creatinine clearance or by estimating the creatinine clearance or eGFR.

Creatinine clearance can be calculated by calculating the creatinine level in a 24 hour urine specimen. However, this method used to calculate creatinine clearance is not preferred since it is burdensome.

The equations shown in Table 2.1 are estimated GFR or creatinine clearance using serum creatinine, age, gender, ethnicity and/or weight.

Table 2.1: Various equations to estimate GFR

| Name | IDMS Traceable | Equation |
|---|---|---|
| Crockcroft and Gault formula (1976) (Cr Cl) | NO | $(140 - age) \times (Wt\ in\ kg) \times (0.85\ if\ female)/$<br>$(72 \times SCr\ in\ mg/dL)$<br>(Cockcroft and Gault, 1997) |
| Modification of Diet in Renal Disease (MDRD) formula (1999) (eGFR) | YES | $eGFR = 175 \times (0.011312 \times SCr)^{-1.154} \times$<br>$age^{-0.203} \times 0.742(if\ female) \times$<br>$1.212(if\ black)$<br>(Levey *et al.*, 2006) |
| Modification of Diet in Renal Disease (MDRD) formula (1999) (eGFR) | NO | $eGFR = 186 \times (0.011312 \times SCr)^{-1.154} \times$<br>$age^{-0.203} \times 0.742(if\ female) \times$<br>$1.212(if\ black)$<br>(Levey *et al.*, 1999) |
| CKD-EPI formula (2009) (eGFR) | YES | |
| White/Other and Female | | If SCr ≤62 µmol/L<br>$eGFR = 144 \times \left(\frac{SCr}{0.7}\right)^{-0.329} \times (0.993)^{age}$<br>If SCr > 62 µmol/L<br>$eGFR = 144 \times \left(\frac{SCr}{0.7}\right)^{-1.209} \times (0.993)^{age}$ |
| White/Other and Male | | If SCr ≤80 µmol/L<br>$eGFR = 141 \times \left(\frac{SCr}{0.9}\right)^{-0.411} \times (0.993)^{age}$<br>If SCr > 80 µmol/L |

| | |
|---|---|
| | $eGFR = 141 \times \left(\frac{SCr}{0.9}\right)^{-1.209} \times (0.993)^{age}$ |
| Black and Female | If SCr ≤62 µmol/L<br>$eGFR = 166 \times \left(\frac{SCr}{0.7}\right)^{-0.329} \times (0.993)^{age}$<br>If SCr > 62 µmol/L<br>$eGFR = 166 \times \left(\frac{SCr}{0.7}\right)^{-1.209} \times (0.993)^{age}$ |
| Black and Male | If SCr ≤80 µmol/L<br>$eGFR = 163 \times \left(\frac{SCr}{0.9}\right)^{-0.411} \times (0.993)^{age}$<br>If SCr > 80 µmol/L<br>$eGFR = 163 \times \left(\frac{SCr}{0.9}\right)^{-1.209} \times (0.993)^{age}$<br>(Levey *et al.*, 2009) |

The estimation of creatinine clearance can be achieved by using the original Cockcroft and Gault formula (Cockcroft and Gault, 1976). This formula overestimates GFR due to the contribution of the tabular secretion and was not normalised based on body surface area.

The Modification of Diet in Renal Disease (MDRD) formula is an alternative way of directly estimating GFR as it has shown to give accurate estimates of GFR when the eGFR value is below 60 mL/min/1.73m$^2$. By using the MDRD formula, GFR was found to be 26% lower in women than in men and about 18% lower in Caucasians rather than in African Americans. This is because African Americans tend to have higher muscle mass and hence higher serum creatinine levels, than Caucasians. Therefore, this ethnicity effect causes African Americans to have lower eGFR compared to Caucasians. The MDRD equation being normalised to body surface area by including a factor of 1.73m$^2$ on the denominator has an advantage over the

Cockcroft and Gault formula. However, GFR was underestimated and hence the accuracy decreased when the MDRD equation was used for patients where eGFR values were above 60 mL/min/1.73m$^2$.

More recently, the CKD-EPI formula was developed consisting of eight different sub-formulas shown in Table 2.1. The choice of sub-formula to apply depends on the gender, ethnicity and whether the serum creatinine value is lower or greater than a specified threshold. The threshold value used depends on both gender and ethnicity. It has been demonstrated that both the MDRD equation as well as the CKD-EPI formula provide very similar estimates of GFR when the actual GFR is below 50 mL/min/1.73m$^2$. However, when the actual GFR is above 50 mL/min/1.73m$^2$, the CKD-EPI formula provides a better estimate of GFR than the MDRD formula. Although a reduction in bias was obtained, estimates of GFR remained imprecise when the CKD-EPI formula was used instead of the MDRD formula. Hitherto, it is unknown whether replacing the MDRD formula with the CKD-EPI formula will create any changes in clinical detection and management of the CKD and additionally, the CKD-EPI formula has not been validated by the Kidney Disease Education Programme to make a recommendation on the general application of this equation. Therefore, clinicians are continuing to use the MDRD formula instead of the CKD-EPI equation.

Cells renew and propagate generating Cystatin C (a 13 K Dalton protein) at a constant rate. This protein is easily filtered by the glomeruli and is reabsorbed by the tubes to be metabolised instead of secreted (Randers and Erlandsen, 1999). Therefore, it is considered as an alternative marker of GFR (endogenous marker). Since the concentration of this protein does not depend on age, gender and muscle mass, it is a very good estimator of the GFR for certain groups of people (Filler *et al.*, 2005). However, it is not used in clinical practice because of the high cost of its measurement and the un-standardised assays of its measurement (i.e. using a different laboratory procedure to obtain the measurement of Cystatin C for each person).

All of the prediction equations detailed in Table 2.1 assume that the muscle mass is estimated using demographic information such as the age, gender and ethnicity of patients. However, as the estimation of muscle mass is used by only taking account of demographic factors rather than its precise measure; all the equations will give an imprecise estimate of the actual GFR.

At the time  a biomarker is measured, either serum creatinine or Cystatin C, all of the prediction equations assume that the kidney function is stable; hence, it is inappropriate to use these equations to measure kidney function if there is an acute kidney injury.

In terms of obesity and underweight conditions, when the patients are at extreme conditions, either having a very low body mass index (< 18.5 kg/m$^2$) or a very high body mass index (>30 kg/m$^2$), problems occur in the estimation of GFR using these prediction equations. For underweight patients, both the MDRD and the CG formulas have a tendency to overestimate GFR. On the other hand, for obese patients, creatinine-based estimations of GFR are likely to create problems in the estimation of GFR (Lamb *et al.*, 2005). In obese patients, the CG formula tends to overestimate clearance and hence overestimates GFR using the actual body weight (Froissart *et al.*, 2005).

In the United Kingdom, serum creatinine is measured in laboratories either using a colorimetric (by observing the colour change in the presence of a chemical compound) or an enzymatic (by observing enzyme reactions) assay, then an isotope dilution mass spectrometry (IDMS) traceable value (that is the value calibrated against National Institute of Standards) is calculated and hence GFR is estimated using the MDRD formula (Estimating GFR, 2005) (MacKenzie, 2006). However, the interpretation of the results should be carefully examined since this formula is not appropriate for use with obese patients and patients under the age of 18. It has not been validated to be used for patients aged above 75 either, although it can be used to provide an indicative result.

### 2.1.4 Classification of CKD

Chronic Kidney Disease defined as either by kidney damage or in the case the GFR value drops below 60 is classified as *chronic* or acute according to whether these structural or functional abnormalities exist for longer than 90 days. Kidney damage is identified by pathological abnormalities (detected by a biopsy) or with the identification of markers of kidney damage where the patient needs to receive a kidney transplant (Vassaloti *et al.*, 2007). Markers of kidney damage are classified into two categories; the urinary abnormalities such as proteinuria and/or haematuria and blood abnormalities (renal tabular syndromes); such abnormalities are detected by scanning and kidney transplantation. After the existence of CKD was determined, CKD was classified into stages according to the National Kidney Foundation-Kidney Disease Outcome Quality Initiative (NKF-KDOQI) established in 2002, and the same classification was adopted in Kidney Disease: Improving Global Outcomes (KDIGO) in 2004 (KDOQI, 2002) (Levey *et al.*, 2005). The international staging system of describing Chronic Kidney Disease detailed in Table 2.2, is authorised in the UK by the Scottish Intercollegiate Guidelines Network (SIGN), the National Institute for Health and Clinical Excellence (NICE) and the Joint Specialty Committee on Renal Disease (Scottish Intercollegiate Guidelines Network, 2008) (Burden *et al.*, 2005).

Table 2.2: International Staging System of Chronic Kidney Disease

| Stage | Definition | eGFR (mL/min/1.73m$^2$) |
|---|---|---|
| 1 | Presence of kidney damage, with normal or raised GFR | $\geq 90$ |
| 2 | Presence of kidney damage, with mildly reduced GFR | 60-89 |
| 3 | Moderately reduced GFR | 30-59 |
| 4 | Severely reduced GFR | 15-29 |
| 5 | End-stage kidney disease | $< 15$ |

(KDOQI, 2002).

As more predictive evidence of CKD became available, an international controversies conference on definition, classification, and prognosis in CKD organized by Kidney Disease Improving Global Outcomes (KDIGO), 2009 decided to make some modification to the classification system established by KDOQI in 2002 (Gansevoort *et al.*, 2009). The three modifications of the CKD classification system suggested and approved. First is the addition of the cause of CKD to the stage, if known. Furthermore, stage 3 was subdivided into two parts; namely the stage 3A in the case $45\leq$ eGFR$\leq 59$ and the stage 3B in the case $30\leq$ eGFR$\leq 44$ (mL/min/1.73m$^2$). Finally, the level of albuminuria was added to the stage of CKD based on eGFR. Currently, KDIGO has developed new guidelines for CKD, where these modifications are included to the classification in the form of a CGA (cause, eGFR and albuminuria) classification (Levey *et al.*, 2011).

Chronic kidney disease is a broad term which includes different kinds of particular renal diseases such as diabetic nephropathy, glomerulonephritis, and hypertensive nephropathy. Therefore, each patient diagnosed with CKD exhibits different renal pathologies, as well as separate natural histories and prognosis. Hence, such diverse natural history results in a unique pattern of decline in eGFR for each type of underlying renal disease (e.g. an adult having polycystic kidney disease would be expected to have a linear decline in eGFR).

### 2.1.5 Symptoms of CKD

In mild to moderate stages (between stages 1 to 3) of CKD, no serious symptoms are observed. However, as the disease progresses into later stages (e.g. stage 4), then symptoms like tiredness and lower energy levels than normal are observed. As the disease develops (stage 5), additional symptoms are observed such as loss of appetite a struggle in clear thinking, an observed reduction in weight, muscle cramps, dry and itchy skin, frequent urination, having pompousness near the eyes, demanding to hold more fluids possibly resulting in the swelling of feet and ankles and a pale complexion as a result of anaemia (UK National Kidney Federation, 2010).

### 2.1.6 Importance of CKD

Over the past 50 years, improvements in technology and immunology of dialysis and transplantation have increased the accessibility of renal replacement therapy (RRT). Hence, the management of patients with end stage renal disease (ESRD) has been developed. However, when compared to the general population, the quality of life of the patients who receive dialysis is significantly poorer (Merkus *et al.*, 1997). Hence, identification of CKD at earlier stages is vital as it will reduce the rate of progression to ESRD.

Organisations in the UK and around the world, such as the Scottish Renal Registry, UK Renal Registry, European Renal Registry, Australia and New Zealand Dialysis and Transplant Registry and the United States Renal Data System have noted that the incidence and the prevalence of patients who are on kidney dialysis treatment is well-reported. However, the incidence and the prevalence of patients having CKD are poorly reported (Scottish Renal Registry Report, 2010) (UK Renal Registry, 2010) (ERA EDTA Registry, 2011) (Australia and New Zealand Dialysis and Transplant Registry, 2011) (US Renal Data System, 2011). Poor reporting of incidence and prevalence of patients having CKD might be due to the underestimation or the overestimation of cases which occurs for several reasons. An overestimation can be a result of a study using a single creatinine measurement to define CKD rather than two eGFR estimations taken at least 90 days apart. On the other hand, underestimation can be the result of a study using records of a population collected during routine clinical care, where it is assumed that any patient who is not involved in routine clinical care does not have the disease.

The first study on the prevalence of CKD in USA was reported in 1999 by the National Health and Nutrition Examination Survey (NHANES), which analysed the data collected between 1999 and 2004 from adults aged ≥20. The study reported the prevalence of CKD in stages 1-4 in the USA as 13.1% (Coresh *et al.*, 2007). Subsequently, the AusDiab study conducted in 2003 in Australia for adults aged ≥25 reported as the prevalence of CKD at stages 3-5 in Australia being 11.2% (Chadban *et al.*, 2003). In 2006, the Health Survey of Nord-Trondelag County (HUNT II) in Norway found that the prevalence of CKD at stages 1-5 to be 4.3% and at stages 1-5 to be 10.2% (Hallan *et al.*, 2006). The latest information in Europe according to the ERA-EDTA Registry Annual Report in 2010, gives the highest incidence rate of CKD in the French speaking part of Belgium (179.8 per million people) followed by Spain, the Canary Islands and the Dutch speaking part of Belgium respectively, with the lowest incidence rate of 74.9 per million people found in Finland. On the other hand, based on the U.S. Renal Data System, Annual Data Report, USRDS (2011), the highest incidence of CKD worldwide is found in Mexico with 597 cases per million people, followed by USA, Taiwan and Japan with rates of 371, 347 and 278 per million people respectively.

In the UK, an older study, namely the NEORICA project, gathered information from 1998 until 2003 from laboratory databases using primary care computer records. The study reported the prevalence of CKD (stages 3 to 5) as 10.6% for females and as 5.8% for males based on all adults in the study older than 18 who had a valid creatinine record (Stevens *et al.*, 2007). However, since only 30% of adults (aged ≥18) had valid creatinine records, selection bias occurred in this study and the study resulted in a biased outcome in terms of identifying CKD patients. In a more recent study carried out by the Health Survey in England (HSE) in 2009-2010, the prevalence of CKD at stages 1-5 was found as 13% (Roth *et al.*, 2011). Patients aged ≥16 were included in this study and a single sample of blood or urine was used for the creatinine measurement. Another study namely the QI CKD was performed in 2010 in the UK which reported the prevalence of CKD in the UK as 6.7% when patients aged ≥18 are taken into account. All of these previous studies in England such as the HSE and NEORICA, have indicated that about 5-10% of adults (aged ≥18) in the UK have moderate to severe CKD (at stages 3-5) and that prevalence is higher in females than in males.

To sum up, prevalence of CKD varies considerably between different countries. In general, since CKD is a growing health problem world-wide, prevalence of CKD is expected to rise especially in UK and USA. Furthermore, the demographic structure of a particular region affects the actual prevalence of CKD, given that the actual prevalence of CKD differs extensively based on the age and the gender profile of the population. Therefore, understanding the natural history of CKD and the way kidney function declines over time is important. The early diagnosis of CKD and suitable treatment and management can prevent the progression of renal function into end-stage renal disease (ESRD).

### 2.1.7 Progression of CKD

There is no single cause of CKD, as there are many factors which can lead to the permanent damage of kidneys and development of CKD. In the UK, three out of four cases of CKD in adults occur due to diabetes, high blood pressure or naturally ageing kidneys. About nine out of ten people with CKD at stages 3-5 have high blood pressure (UK National Kidney Federation, 2010). Therefore individuals with diabetes or high blood pressure have a higher risk of developing CKD (de Lusignan *et al.*, 2009). In such cases, a routine blood test is recommended at regular intervals in order to monitor kidney function. At stages 1 to 3, where the disease is at a mild to a moderate level, CKD does not normally show any symptoms. Due to the asymptomatic nature of the disease, CKD is usually undiagnosed at earlier stages (Król *et al.*, 2008).

Renal disease usually results in the patient having a renal replacement therapy (RRT). Since it is easy to define patients who require RRT, documentation on the patients taking RRT are reported accurately. However, RRT is a treatment and not a clinical stage. Therefore, identifying CKD cases by using RRT does not take into account the patients who are not receiving RRT and hence cannot be used as a measure of the progression of CKD in patients who are not receiving RRT. Moreover, repeated measurements cannot be obtained using RRT due to the outcome measure (i.e. eGFR) being dependent on the treatment received in RRT. The doubling of the concentration of serum creatinine is the second milestone, used to monitor the progression of CKD. The doubling of serum creatinine is associated with a decline in eGFR by 50%. The third measure that can be used to quantify the progression of CKD is the eGFR slope (i.e. the rate of change of eGFR over time). The results obtained from both the eGFR slope and the doubling of serum creatinine can be inappropriate in patients having acute kidney injury and have to be used with caution (Maudsley and Williams, 1996). The cause of death of CKD patients is not directly attributable to the kidney disease. Patients are reported to die from other related illness such as cardiovascular diseases. Cardiovascular mortality is found to have strong association with renal disease. However, this mortality might be over-exaggerated (Maudsley and Williams, 1996).

### 2.1.8 Factors Affecting CKD

According to Amaresan and Geetha, (2008), glomerulonephritis and chronic interstitial nephritis were common causes of CKD in developing countries in the past but in recent years, the situation has changed due to the increasing incidence of diseases such as diabetes mellitus and hypertension which have become the major co-morbidities of CKD. De Lusignan *et al.* (2005) and Bruch *et al.*, (2007) state that cardiovascular diseases and chronic use of Non-Steroidal Anti-inflammatory Drugs (NSAID) (i.e. aspirin) are common causes of CKD. A research carried out in China by Chen *et al.*, (2009) suggested that age, and central obesity (excessive abdominal fat), hypertension, diabetes, anaemia, hyperuricaemia and nephrolithiasis contribute to the progression of CKD.

Ageing is a natural process and cannot be adjusted to a specific level, so it has been defined as an uncontrollable factor for the progression of CKD (Chen *et al.*, 2009). Ageing causes renal function deterioration and increases the risk of developing other diseases such as diabetes, hypertension and atherosclerotic vascular disease (Graves, 2008). Older people also have a higher chance of requiring the administration of nephrotoxic drugs for treatment of cancer, infections and coronary artery disease (Graves, 2008) (Glassock and Winearls, 2009). Therefore, aging will tend to result in the progression of CKD (Chen *et al.*, 2009). Barri (2006) have reported that micro-albuminuria is a predictor of hypertensive renal disease. Amaresan and Geetha, (2008) have reported that glomerular diseases, proteinuria, anaemia, a very high protein intake, smoking, obesity, being above the age of 60 and a family history of CKD also increase the risk of CKD. CKD has been determined as a disease which coexists with other conditions such as dyslipidemia, hypertension, smoking and diabetes having adverse effect on the kidneys. These factors are considered to increase the absolute risk of progression of CKD (Tonelli *et al.*, 2006). Later of this section will consider the impact of each of these co-morbidities on CKD.

### Cardiovascular Disease (CVD)

Cardiovascular Disease as a broad term of different heart and blood vessel diseases CVD includes three main categories, the coronary heart diseases (CHD), stroke and peripheral arterial (vascular) disease (PVD) according to PH25 NICE Guidelines (2010). CHD include various heart conditions; such as ischemic heart disease and hypertension. Therefore, the definition for the diagnosis of CVD in this study is taken as having either hypertension, IHD or PVD. Since stroke is a condition occurring due to the shortage of blood supply in the brain, it is separated from the definition of CVD.

The occurrence of cardiovascular disease in patients having CKD and who require kidney dialysis is 20-100 times greater than the incidence of cardiovascular disease in the general population. Therefore, a high proportion of such patients die from CVD (Foley *et al.*, 1998).

Hypertension is also known as high blood pressure and it is a condition which occurs when the blood pressure in the arteries is constantly higher than normal. It has been demonstrated that hypertension increases the risk of cardiovascular disease in the general population. When compared to the general population, the age and gender adjusted odds ratios show that in a patient with CKD, hypertension is 2 to 1 times more common compared to a patient without CKD (with the 95% confidence interval being 2.0-2.2) (Stevens *et al.*, 2007). In patients with CKD, hypertension is more common, and the risk of cardiovascular disease is higher than in the general population (Kannel, 1996). Furthermore, uncontrolled hypertension causes proteinuria and hence faster progression of CKD which in turn will result in a greater risk of cardiovascular disease. Therefore, in patients with CKD, the control of hypertension is very important.

Hypertension causes damage to other organs such as the heart, blood vessels, the brain and the eyes. This damage may lead to a gradual progression of cardiovascular and renal diseases (Gallaghter *et al.*, 2010). Therefore, hypertension can be seen as an independent risk factor for cardiovascular and renal diseases which in turn increase mortality rates (Barri, 2006).

There are two components in blood pressure measurement, namely the systolic measurement, which represents the peak pressure in the arteries, and the diastolic measurement, which characterises the minimum pressure in the arteries. Guidelines recommend that patients with CKD should maintain their systolic blood pressure (SBP) at 120-139 mmHg, and diastolic blood pressure (DBP) below 90 mmHg (National Collaborating Centre for Chronic Conditions, 2008). In patients with diabetes mellitus, SBP should be kept between 120-129 mmHg while DBP should be kept below 80 mmHg (Scottish Intercollegiate Guidelines Network, 2008) (National Collaborating Centre for Chronic Conditions, 2008). However, in the case the SBP is reduced below 100-110 mmHg, then this might be harmful (Gordon *et al.*, 2011).

A patient suffering from hypertension shows high blood pressure in their arteries, affecting the glomeruli in their kidneys and leading to glomerular hypertension, which in turn damages the kidneys (Brenner *et al.*, 1982). Existence of high blood pressure in glomerular capillaries results in a more rapid decline of kidney function (Jafar *et al.*, 2003). Furthermore, hypertension is a major factor contributing to development of ESRD. This is particularly notable for people of black (African American or Afro-Caribbean) ethnicity, who are six times more likely than white people to progress to ESRD from hypertension due to factors such as medical care, socioeconomic status, education level, alcohol and drug use and genetic predisposition. Therefore, in patients with CKD, blood pressure should be strictly controlled in order to slow down the progression of renal disease. In high risk patients, optimum blood pressure (SBP/DBP) should be maintained below 130/80 mm Hg.

### Diabetes Mellitus

Diabetes mellitus, or simply diabetes results from high blood sugar levels (i.e. glucose level). There are three main types of diabetes, namely; type I diabetes, type II diabetes and gestational diabetes. Type I diabetes occurs as a result of the failure of the pancreas to produce sufficient insulin whereas type II diabetes results when the body cannot use the secreted insulin properly, which is known as insulin resistance. Gestational diabetes arises in women when the blood glucose level is raised during pregnancy. It has already been found that cardiovascular disease is a major cause of diabetes mellitus. Therefore, when a patient has diabetes, the patient has three times more risk for death due to myocardial infarction, which is also known as a heart attack, than the general population (Kannel *et al.*, 1985). Several types of cardiovascular diseases often exist in patients with diabetes (especially patients with type II diabetes) and additionally, the existence of hyperglycaemia (a condition of very high glucose level) results in macrovascular disease (disease of large blood vessels) since these diseases (i.e. hyperglycaemia and macrovascular disease) are directly related (Milicevic, 2008). If diabetic nephropathy (with albuminuria) is present, this increases the risk of death. Thus, control of blood pressure is very important in patients with diabetes in order to reduce the development and progression of cardiovascular diseases. Similarly, in patients with diabetes, glycaemic control (control of blood sugar level) is very important in order to prevent microvascular complications, as the problems occurring due to damage of the small blood vessels (UK Prospective Diabetes Study Group, 1998). If a patient with diabetes is identified as having renal failure, then kidney transplantation is advised to reduce the risk of cardiovascular mortality (Reddy *et al.*, 2003).

### Anaemia

Haemoglobin is the iron-based metalloprotein in red blood cells responsible for oxygen transport. A normal haemoglobin concentration for non-pregnant women above 15 years old is 7.4 mmol/l, while the normal concentration for a man above the age of 15 is 8.1 mmol/l.

Anaemia as an iron deficiency condition caused by a significant haemoglobin level reduction such as below the normal value. Erythropoietin (EPO) as the hormone that controls the production of red blood cells is excreted from the peritubular cells which are located in the peritubular capillaries in the renal cortex. The malfunction of the peritubular cells results in the loss of production of the EPO hormone, resulting in anaemia; a condition thus related with CKD. Anaemia causes insufficient oxygen supply to the tissues, called hypoxia. Anaemia also encourages the formation of excessive connective tissue (fibrosis) (Drüeke *et al.*, 2006; Singh *et al.*, 2006; Kraut *et al.*, 2005).

The reduction of the number of nephrons and tubular cells in kidneys caused by ageing and the decreased renal blood flow and EPO secretion is another cause of anaemia. Hence, anaemia is more common in the elderly population (above 60). In elderly people, anaemia increases the risk of mortality and other co-morbidities such as CKD. Since the haemoglobin concentration is an independent measure of deterioration of the renal function, lowering the haemoglobin concentration below 14 g/dl will accelerate the development of CKD to ESRD specifically in patients with type II diabetes (Lee *et al.*, 2008).

The impact of anaemia on patients with CKD was demonstrated in several studies (Levin *et al.*, 1999; Foley *et al.*, 1996)). Anaemia increases the risk of left ventricular hypertrophy (LVH) (a condition resulting from the increase in pressure or volume overload in the heart capillaries due to the reduction of blood flow) and fibrosis. The association between anaemia and LVH is reported in a study that for every 0.5 g/dL decrease in haemoglobin, the risk for the patient having LVH increases by 32% (Levin *et al.*, 1999). Another study performed by Foley et al (1996) has shown that there is a strong correlation between anaemia and cardiac abnormalities in patients having dialysis therapy. Foley and colleagues (1996) evaluated the risk of left ventricular dilation as 46%, the risk of poor left ventricular ejection fraction as 55% and the risk of death after the RRT as 14% higher than a healthy patient with a 1 g/dL decrease in the patient's haemoglobin concentration (Foley *et al.*, 1996). Therefore, the treatment of anaemia is very important in patients with CKD and anaemia.

## Proteinuria

According to the current literature, several studies have shown that there is a strong correlation between proteinuria (as the presence of protein in the urine) and cardiovascular disease (Kannel and Culleton, 2000; Wahi *et al.*, 1997; Tonelli *et al.*, 2006). According to the Framingham heart study (Kannel and Culleton, 2000), proteinuria (diagnosed with a dipstick test on the general population cohort of Framingham) was shown as an independent risk factor of cardiovascular disease and that mortality was due to other causes (Kannel and Culleton, 2000). Another study carried out by the Multiple Risk Factor Intervention Trial (MRFIT) cohort, (including only men under the risk of cardiovascular disease and using the same measurement to diagnose proteinuria as in the Framingham study and the MRFIT study) conceded that proteinuria is an independent predictor of cardiovascular disease and mortality was due to all causes (Wahi *et al.*, 1997). The study carried out by Tonelli *et al.*, (2006) again using a dipstick test to detect proteinuria showed that risk of all-cause mortality due to proteinuria and kidney dysfunction increased as the patient deteriorated from CKD stage 3. On the other hand, a study conducted in the UK taking account 13,177 community residence (that shares a common value) adults aged over 75 concluded that proteinuria diagnosed by dipstick test is an independent risk factor for all-cause mortality but not cardiovascular mortality in particular (Roderick, 2009).

Many studies were performed using detection of albuminuria in the urine to diagnose proteinuria (Yudkin *et al.*, 1988; Damsgaard *et al.*, 1990; Perkovic *et al.*, 2008; Hemmelgarn *et al.*, 2010). Albuminuria is a specific type of proteinuria, detected by collecting a 24-hour urine sample, and measuring the albumin concentration. The Alberta Kidney Disease Network evaluated the relationship between albuminuria and cardiovascular disease using extensive regional laboratory-collected data where the albuminuria was measured based on the result of 24-hour urine sample. The study concluded that the albumin/creatinine ratio (ACR) is an independent predictor of heart attack (myocardial infraction) and all-cause mortality (Hemmelgarn *et al.*, 2010). In the same study, performed by the Alberta Kidney Disease Network, the albumin/creatinine ratio was determined using a dipstick test to diagnose proteinuria and the same correlation was reported. Several other studies performed using a 24-hour urine sample to detect albuminuria have concluded the same relationship between ACR and cardiovascular disease (Yudkin *et al.*, 1988), which is also verified on the elderly group (age 60-74) (Damsgaard *et al.*, 1990). A meta-analysis including 26 studies on proteinuria and coronary disease, concluded that the risk posed to a patient of developing coronary disease is 1.47 times more if the patient has proteinuria with the 95% Confidence Interval (CI) being 1.23-1.74 (Perkovic *et al.*, 2008). The same meta-analysis also concluded that there is a significant albuminuria dose-dependent effect (Perkovic *et al.*, 2008). Therefore the risk of developing coronary disease increases if the patient has macroalbuminuria (occurring in high levels of albumin in the urine) rather than microalbuminuria (occuring in low albumin levels in the urine).

Over the recent years, the relationship between albuminuria and eGFR was taken into account by the CKD Prognosis Consortium. The meta-analysis published by CKD Prognosis Consortium included 22 studies and over 1.2 million patients with over 100,000 urine ACR measurements (Astor *et al.*, 2011). Furthermore, for some patients dipstick measurement was also used. This meta-analysis was conducted to evaluate the correlation of albuminuria and eGFR on the all-cause and cardiovascular mortality for the general population. The meta-analysis, reported a linear relationship between albuminuria and the risk of cardiovascular disease (Astor *et al.*, 2011). However, no significant interaction was found between albuminuria and eGFR. Another meta-analysis carried out by the CKD Prognosis Consortium on a cohort of patients with CKD suggested the same linear relationship between albuminuria and the risk of cardiovascular disease. A further study including 1056 patients with type II diabetes and 1375 non-diabetic patients validated the association between the concentration of total protein in the urine and cardiovascular and all-cause mortality.

### Obesity

Obesity is an independent factor in relation to other co-morbidities that affects the progression of CKD (Rutter, 2011). According to current reports there is a debate about the best measure of obesity. Currently, two different types of measurements are used in the clinical research namely; the body mass index (BMI) and the waist: hip ratio (WHR). A patient is classified as obese if their BMI is greater than 30 or if their WHR is above 0.85 for females or 0.90 for males. A higher BMI or WHR increases the risk of progression of CKD. Weight loss either by surgical or non-surgical methods improves the control of blood pressure, reduces proteinuria and reduces hyper-filtration in the kidneys (Ibrahim *et al.*, 2006).

Moreover, obesity increases the risk of development of diabetes, hypertension and dyslipidemia (a condition of having abnormal amounts of lipids) which are also associated with a higher risk of developing CKD (Navaneethan *et al.*, 2009). However, obesity relates to excess body weight, and this increases the risk of CKD independently of blood pressure and the existence of type II diabetes (Goncalves *et al.*, 2009). Patients maintaining their weight at normal levels or achieving weight loss by non-surgical interventions, will typically decrease their BMI and have an improvement in GFR and also maintaining weight at normal levels will create no change in their plasma glucose and triacylglycerol levels that will prevent the increase in risk of getting diabetes and cardiovascular diseases. Additionally, a decrease in BMI will decrease the systolic blood pressure as well as the total cholesterol (Goncalves *et al.*, 2009). According to Navaneethan *et al.*, (2009), a decline in BMI also results in a decrease in proteinuria. In the case of an extremely obese patient, surgical intervention is used for weight loss helping to regulate GFR, and resulting in a decline in systolic blood pressure and microalbuminuria.

Microalbuminuria is reduced as a result of lowering the urinary excretion, which is in turn, is reduced by lowering the blood pressure and the blood lipid levels (Navaneethan *et al.*, 2009). Weight-gain results in an increase in BMI and typically reduced GFR levels while increasing in plasma glucose and triacylglycerol levels (Goncalves *et al.*, 2009).

### Low Protein Diet

The standard protein intake for an adult is considered as ≥0.8 g/kg per day. Moderate to severely limited protein intake is restricted to 0.3-0.6 g/kg per day (Garrett, 2007). Fouque *et al.* (2007) have suggested that patients with CKD, a moderate to severely limited protein intake tends to postpone the progression of CKD to ESRD or death (Fouque *et al.*, 2006; Fouque *et al.*, 1992; Fouque and Aparicio, 2007; El Nahas *et al.*, 1984).

## Alcohol Consumption

Alcohol consumption of less than 10 g per day is considered as light, an alcohol intake between 10 g and 30 g per day is classified as moderate alcohol consumption whereas heavy alcohol consumption is defined as being in excess of 30 g of alcohol per day (White *et al.*, 2009). Heavy alcohol consumption is an indirect risk factor of CKD, due to its increasing risk of hypertension and cardiovascular diseases and cerebrovascular diseases such as stroke. Comparing light alcohol consumption with moderate to heavy alcohol consumptions, both moderate and high alcohol consumptions increase the risk of albuminuria, due to the increase in urinary albumin excretion, even after an adjustment for age, gender and baseline kidney function. Recent statistics have shown that individuals aged 25 to 44 are current drinkers while people above the age of 65 are mostly non-drinkers (White *et al.*, 2009).

## Nephrology Referral

In order to determine the impact of a nephrology referral (timing to go to nephrologist), the pre and post referral slopes for the GFR dependence on time were calculated. The slope of GFR was broadly categorised into two types; the "*progressive slope*" which indicates development of CKD having a GFR slope ≤-1 and the "*non-progressive slope*" which was determined from GFR slope >-1 (Jones *et al.*, 2006).

The GFR slope was further sub-categorised as:

- The disease is progressing fast if the GFR slope ≤-5,
- The disease is progressing slowly if the GFR slope is between -5 and -1,
- The disease is stable if the GFR slope is between -1 and +1,
- The disease is progressing moderately if the GFR slope is between +1 and +5 and
- The disease is highly non-progressing if the GFR slope is ≥ +5 (Jones *et al.*, 2006).

Jones *et al.* (2006) suggested that GFR decline slowed significantly after a nephrology referral. Using different methods such as Jaffe assay or Enzymatic creatinine assay of measuring serum creatinine before and after the referral could be a reason for a reduction of GFR decline (Jones *et al.*, 2006). Reduction of blood pressure (BP) was also observed after a nephrology referral also suggesting that a good BP control early in the disease process is very important since it can lead to a slower decline in GFR and therefore lower mortality rates (Jones *et al.*, 2006).

### 2.1.9 Treatment of CKD

Hypertension is classified as being the most significant factor causing cardiovascular diseases such as heart failure. Cardiovascular disease being one of the major risk factors for CKD, it is important to treat hypertension to reduce the risk of developing cardiovascular disease and hence CKD (Miller *et al.*, 2006). High blood pressure causes the heart of the patient to work harder and pump the blood to the whole body and hence damaging the blood vessels including the glomeruli. Damaged blood vessels in kidneys lead to the protein to transfer from the blood to the urine. Therefore, hypertension has a strong correlation with urine protein loss and lower blood pressure targets should be met in order to control hypertension and hence to minimise the transfer of protein from the blood to the urine.

The main risk factor of renal disease progression is the increase in systolic blood pressure levels (hypertension), which if it is not controlled, it can lead to microalbuminuria and CKD. Angiotensin I-converting enzyme (ACE) inhibitors are prescribed for the treatment of high blood pressure (Barri, 2006). ACE inhibitors and angiotensin receptor blockers (ARBs) are the two main agents prescribed for the treatment of hypertension (Berger *et al.*, 2007). Patients having hypertension together with other co-morbidities such as heart failure or diabetes, ACEIs and ARBs also exhibit benefit in treating these co-morbidities conditions (Miller *et al.*, 2006). ACE inhibitors are the preferred agents over ARBs which are only prescribed as an alternative if ACE inhibitors are not suitable for the patient (Miller *et al.*, 2006). Both ACE inhibitors and ARBs treat diabetic nephropathy (progressive kidney disease caused by diabetes). Captopril is a kind of ACE inhibitor which is the standard treatment of type I diabetes, whereas there is no standard ARB that can be used in the same treatment. On the other hand, irbesartan and losartan are the two kinds of ARBs which are standard for the same role, as treatment for type II diabetes whereas there is no ACE inhibitor that is accepted (Miller *et al.*, 2006). ACEIs are established to slow down the development of type I diabetes and in patients with type II diabetes, ACEIs are prescribed to slow the development of micro-macroalbuminuria (Miller *et al.*, 2006) (Berger *et al.*, 2007).

According to the United Kingdom Prospective Diabetes Study (UKPDS), it was demonstrated that, patients with type II diabetes, controlling the blood pressure decreases the possibility of developing damage or disease of a kidney (Miller *et al.*, 2006). However, in order to reduce the risk of nephropathy in patients with diabetes, glucose control was also significant (Miller *et al.*, 2006).

Apart from hypertension, high levels of proteinuria (heavy proteinuria) were also an important factor which needs to be considered. In patients with chronic nephropathies (kidney diseases), heavy proteinuria accelerates progression towards ESRD. Heavy proteinuria can still be observed even when ACEIs are prescribed as treatment (Ruggenenti *et al.*, 2008). Therefore, Ruggenenti *et al.* (2008) applied a multimodal intervention to achieve optimum urinary protein levels. This intervention resulted in a significant decrease in proteinuria. The study also showed that proteinuria reduction independently results in a slower rate of decline of GFR and hence slows the deterioration of kidney function. In order to normalize proteinuria and to decrease the decline of renal function in patients without type II diabetes but with chronic nephropathies, the study used a multidrug treatment technique. The response to the treatment depended on each patient's underlying condition (Ruggenenti *et al.*, 2008).

Two studies published in 2006, namely the CREATE study and the CHOIR study have examined the treatment of anaemia in CKD patients. The CREATE study included patients with CKD at stages 3 and 4. The main aim of this study was to distinguish between high and low targets of haemoglobin. At the end of the study, a high haemoglobin level was found to be 13.4 g/dL and a low haemoglobin level was found to be 11.5 g/dL. However, no difference was observed in the outcome (having a cardiovascular event) of high versus low haemoglobin groups (Drüeke *et al.*, 2006). On the other hand in the CHOIR study, the design of the study was the same as the CREATE study but a high haemoglobin level was determined as 13 g/dL and the low haemoglobin level was determined as 11.3 g/dL. However, the result of the CHOIR study was different from the CREATE study in that it was concluded in the CHOIR study that patients having high haemoglobin values also have increased rates of cardiovascular events (Singh *et al.*, 2006). Moreover, it has been found that patients in this group have higher incidence of hypertension and thus require higher doses of EPO (i.e. drug containing erythropoietin hormone that controls red blood cell production) to prevent undesirable cardiovascular events.

Recently, a study examining the effect of darbepoetin (which causes the production of erythropoietin hormone) treatment was published. This study included patients with CKD who were not under dialysis treatment and who were not diagnosed with anaemia and type II diabetes. In this study, darbepoetin was used to target two groups, one with haemoglobin level at about 13 g/dL and other with haemoglobin level below 9 g/dL. As a result, the study concluded that the control of haemoglobin had not improved the primary end point (which was death, myocardial infraction, unstable angina, heart failure, stroke or asymptomatic hyperuricaemia) (Pfeffer *et al.*, 2009).

### 2.1.10 Management of CKD

The prevalence of CKD is high worldwide, and in the UK in particular. Therefore, the management of CKD is very important (Burden *et al.*, 2005). Guidelines have been produced to make recommendations for the identification, management and referral in primary care of patients with CKD (Burden *et al.*, 2005) (UK Renal Association, 2011). In the UK, CKD is usually identified as stage 3, where an eGFR value was recorded alongside with a creatinine measurement for each patient. In the UK, the patients diagnosed with CKD at stage 3 are normally managed by a general practice team of doctors and nurses. Early identification of CKD was encouraged by the Quality Outcomes Framework (QOF) indicators which were firstly introduced in 2006. The QOF of the General Medical Services contract in the UK developed a payment system for primary care practices based on the number of chronic illnesses treated. QOF has benefitted CKD patients through earlier diagnosis of CKD, creating a registration system for the CKD patients in order to ensure that a regular monitoring was performed, following-up kidney function, proteinuria and blood pressure regularly for those patients with CKD and providing management of blood pressure to achieve specific targets.

The development of cardiovascular diseases as a result of CKD usually results in death (de Lusignan *et al.*, 2009), hence, to decrease mortality, it is very important to diagnose CKD at the earlier stages to prevent or slow down the progression of the disease by prescribing a suitable treatment. Good control of blood pressure and of glucose level in diabetes, medication to lower cholesterol levels, a healthy diet, regular physical activity and keeping weight under control are used to prevent cardiovascular disease (UK National Kidney Federation, 2010).

Dietary protein restriction is one of the factors which need to be considered to delay the progression of CKD. Protein restriction terminates the increase in glomerular disease and slows the progression to end-stage renal disease (Amaresan and Geetha, 2008). Patients typically need to start protein restriction at the early stages of CKD when the GFR was around 60 mL/min/1.73m$^2$. If the patient's GFR value is between 25-55 mL/min/1.73m$^2$, then their protein intake should restricted to be 0.8 g/kg/day. If the GFR figure is less than 25 mL/min/1.73m$^2$, then protein intake should be restricted to 0.6 g/kg/day (Amaresan and Geetha, 2008) (Black *et al.*, 2010).

As proteinuria is important screening analysis which can indicate glomerular disease, reduction in proteinuria has to be considered. Reduction in proteinuria delays the progression of CKD (Andrews, 2008). Both the REIN study and the MDRD study carried out by Amaresan and Geetha (2008) validate this argument. Hypertension in conjunction with CKD increases the risk of cardiovascular disease and increases both the mortality (death) and morbidity (having the disease) rates. Control of hypertension is achieved by controlling the blood pressure (BP). All patients with CKD should maintain a BP lower than 130/80 mmHg, and patients having proteinuric renal disease should retain a BP below 120/75 mmHg. In 1993, Jafar *et al.* reported that in order to reduce the development of CKD and proteiuria, ACEIs are more successful than other anti-hypertensive treatments (Amaresan and Geetha, 2008). Two large prospective randomised trials carried out by Amaresan and Geetha (2008) suggested that ARBs are also effective in reducing progression of CKD in type II diabetic patients. Coughing as a side effect is very rare from prescription of ARBs whereas it is common in patients taking ACEIs (about 40%) (Burden and Tomson, 2005). The prescription of ACEIs in combination with trandolapril (an ACEI drug used for the treatment of high BP) and ARBs in combination with losartan (an ARB drug used for the treatment of high BP), caused greater success in reducing the development of CKD and urinary protein excretion than using the individual drugs (ACEIs or ARBs) (Amaresan and Geetha, 2008) (Black *et al.*, 2010).

Stringent glycemic control is achieved by reducing the patient's blood sugar level. The target blood sugar level should be less than 100mg/dl under fasting conditions and less than 130mg/dl two hours after eating a meal slowing the progression of CKD in diabetics (Crowe *et al.*, 2008). In patients with CKD, anaemia is caused by the reduction of the erythropoietin during the early stages of the disease. Controlling anaemia is important in CKD because anaemia can produce adverse effects, on the heart and vascular system which can lead to death. In order to correct anaemia, as well as managing erythropoietin levels, iron supplementation in the diet is recommended. Correction of anaemia will improve a patient's quality of life (Amaresan and Geetha, 2008; Andrews, 2008).

Hyperlipidemia (high levels of lipids in the blood) increases the development of CKD as well as the cardiovascular morbidity and mortality. The MDRD study carried out by Amaresan and Geetha (2008) suggested that lower levels of serum high density lipoprotein (HDL) cause a rapid decrease in GFR. The same study also suggested that low density lipoprotein (LDL) should be kept under 100mg/dl. Therefore, statins are prescribed as a treatment to decrease LDL and increase HDL. Statins control hypertension by decreasing peripheral arterial resistance. Statins also decrease the acuteness of proteinuria (Amaresan and Geetha, 2008).

Treatment of underlying diseases such as hypertension, anaemia and diabetes, or of infections and earlier referral to nephrologists were recommended to avoid the deterioration of kidney function. Excessive fluid volume due to obesity causes hyper filtration which in turn leads to microalbuminuria and glomerulosclerosis (a condition of hardening of glomerulus in kidney). Therefore, treatment of obesity and control of salt intake, with a suitable nutritional diet and regular exercise is recommended to the patient. Smoking causes vascular problems and systemic hypertension increasing the risk of development of CKD or causes an increase in progression if already present. Therefore, patients aiming to control CKD must give up smoking entirely (Amaresan and Geetha, 2008; Andrews, 2008; Frankel *et al.*, 2005).

## 2.2 Literature Review on Statistical Background

### 2.2.1 Longitudinal Studies

Longitudinal data is becoming widely available across many fields of study and from varied resources, and shares common characteristics with multivariate data and time series data, For example, like with multivariate data the same subject is measured over time at various time points, resulting in repeated measurements rather than measuring different features of the same individual at a particular time point; and as with time series data measurements are obtained from the same subject about a particular feature which are ordered in time. The main difference between longitudinal data and time series data is the number of subjects measured and the number of repeated measurements obtained from each subject. While large number of subjects are measured in longitudinal data with each subject having small number of repeated measurements, in time series data, a single subject is measured with large number of repeated measurements. (Fitzmaurice *et al.*, 2004).

Correlated data is generated when the data contains either multivariate observations, repeated measurements, time series, clustered data or spatially correlated data. Longitudinal data is defined as a special case of such (correlated data) when a series of measurements are taken from the same subject over long periods of time (Diggle *et al.*, 1994, Molenberghs and Verbeke, 2005). Longitudinal data is naturally generated in many fields including medical research, epidemiology and public health, biological science, environmental science and the social sciences (Fitzmaurice, 2009). The focus of this research project is in the usability of longitudinal data within the medical field. The term '*longitudinal*' in this instance then, indicates data collected over relatively long periods of time and the potential for understanding changes in outcomes over time using this data.

Modern longitudinal studies first emerged in the early 1970s (Hedeker *et al.*, 2006). But it only more recently, that large and complex longitudinal studies have started to grow in popularity. In the last 20 years there has been a steady increase in the interest and use of longitudinal data for research puposes. Researchers have realized the significance of analysis of longitudinal data in biomedical and health-care applications, for example, when they are interested in understanding the progression of a disease over time and detecting the influence of various factors such as age, gender and presence of other diseases on this progression. Therefore, a general aim of longitudinal data analysis is to examine the change in a response over time and also to find factors that impact on this particular change (Fitzmaurice and Molenberghs, 2009). This can be achieved by analysing data taken from measuring the same individual repeatedly over time (Davis, 2002). In such studies where the behaviour of the subject is investigated over time, the relationship between behaviour and time is called *diachronic;* as opposed to '*synchronic*' where the behaviour of the subject is examined at one particular point of time (Ruspini, 2002). An early example of longitudinal study used in medical research was in 1981 in Italy where diachronic relationships were initiated to observe the effect of a treatment on individuals at the end of the study period.

### 2.2.2 Types of longitudinal studies

A single cross-sectional study is where a single observation is observed from each subject. The data for cross-sectional studies can be obtained in two ways; either at the micro (individual) level or at the macro (population) level (Ruspini, 2002). In cross-sectional studies data is observed at a single time point, and hence these studies are easier to set up and cheaper to run than longitudinal studies. Cross-sectional studies investigate a wide sample of the population at a defined time and allow the researcher to achieve results almost immediately (Ruspini, 2002). When a series of repeated cross-sectional studies (also called trend studies) are collected by using different samples of subjects, the information about the response is observed over a long period of time, resulting a longitudinal study (Ruspini, 2002). Repeated cross-sectional studies often obtain the data retrospectively, where the data is collected from examination of past events (Diggle *et al.*, 1994). Therefore the data is available almost immediately. Hence, this feature of repeated cross-sectional studies generates an advantage over a true longitudinal study that usually requires a long period of time for collection if the event is observed prospectively (forward in time). Repeated cross-sectional studies are useful for examining changes at the macro level and can be quite straightforward in this respect. The main disadvantage is that the measurements are not based on the same subject, which can introduce great effects due to between subject variations.

In most cases, including the medical area, longitudinal data are created from accurately controlled experiments such as clinical trials which are collected prospectively (Diggle *et al.*, 1994). Prospective longitudinal studies (also called panel studies) are the preferred type of longitudinal study as observations are repeated regularly at fixed time intervals over time using the same sample. This approach also reduces the scope for error due to retrospective recollection (backwards in time), where there is a danger that the subject will not remember information correctly (Diggle *et al.*, 1994). In these studies, information is gathered when subjects are present in the study and data is collected through future observations (Ruspini, 2002). These studies usually investigate the change at the micro level.

Panel studies are a type of prospective longitudinal study where the data is collected at frequent intervals and is used to determine the stability or fluctuation of opinions and attitudes (Ruspini, 2002). Again data is collected by observing the same individuals regularly over fixed intervals. From the literature, it can be concluded that the shorter the time interval between observations, easier it is to obtain a higher percentage of responses due to the relationship built up between individual/household and the survey or over time (Ruspini, 2002) Panel studies are commonly consumer based or household based.

Rotating panel studies are slightly different in that the sample changes as the population changes. The purpose of this type of panel study is to keep the sample up to date. In this way, the characteristics of the original sample are updated and this decreases the possible bias that can occur in the study due to the characteristics of the sample. Observations are collected from individual only once and never again. Therefore, a rotating sample can form a control group as the sample is not exposed to possible effects of contributing in the survey.

Split panel studies are a further development where data is obtained from investigating rotating samples together with another set of subjects observed over a long period of time. Both rotating and split panels are a combination of panel studies and repeated cross-sectional studies (Ruspini, 2002).

A cohort panel study is designed so that rather than observing data from the same individual, the information is collected from a random sample of individuals who are having the same life-event within the same time interval, such as British birth cohort studies that include repeated surveys to investigate same group of people from birth through their lives. When the same sample is observed over time, a cohort study is effectively a series of panel studies.

Linked or Administrative panel studies gather data from public administration processes. Often in such studies, data from different sources are combined together using unique personal identifier. For example, registration data is joined together with census data by using a unique identifier and hence, a large dataset is obtained by linking administrative data with census data (Ruspini, 2002).

In retrospective longitudinal studies, observations are gathered retrospectively meaning that the data is obtained by considering past events, going backwards in time. Information is collected in relation to repeated events at different time points in the past. However, a drawback of both repeated cross-sectional and longitudinal studies is that information is collected at discrete time points (Ruspini, 2002). Therefore, the time variable in the data set is not the actual time and instead it is the observation point where the response is obtained. In cases when the response is obtained at unequal intervals, analysing such data sets by using discrete time points can produce a bias results. Hence, continuous time points should be created to investigate the actual effect of time on the response, specifically if the repeated measurements are obtained at unequal time intervals.

### 2.2.3 Advantages and Disadvantages of Longitudinal Studies

Longitudinal studies have many advantages over traditional data collection methods, Perhaps the most obvious of these being that in general fewer individual subjects are needed to attain similar statistical power (than say cross-sectional studies) (Hedeker *et al.*, 2006). The reason for this, is that in longitudinal studies, more information is achieved from multiple measurements on the same subject. Therefore, when there are equal numbers of subjects and the same outcome is measured, longitudinal studies can provide better estimators than cross sectional studies.

Further, longitudinal studies are commonly used in analysing changes in a response over time by separating the two types of heterogeneity (variation), namely between subject differences and within subject differences. This is not possible in cross-sectional designs since only a single measurement is observed from each subject (Diggle *et al.*, 1996; Hedeker and Gibbons, 2006). Additionally, a longitudinal study investigation can also take account of the unmeasured subject-specific variability (random variations) in the response. Unmeasured subject-specific factors are components such as genetic, environmental, social or behavioural factors that cannot be measured but which affect the response variable and should be considered in the analysis (Fitzmaurice *et al.*, 2004). Taking these factors into consideration in the analysis of longitudinal data, should lead to an estimation of the response with a better accuracy compared with that provided by cross-sectional studies (Diggle *et al.*, 1996).

For example, in applications to the social sciences, there are three different kinds of effects that can influence outcomes in longitudinal data. These are ageing effect, cohort effect and period effect. Ageing effect results due to the flowing of time, cohort effect represents dependency of the outcome on those subjects born in the same year and period effect is the dependency of the outcome on the time interval of the study. When any two effects are known, the third can also be determined. An advantage of using longitudinal data in such applications is that we can differentiate the ageing effect from cohort effects, again this cannot be achieved with cross sectional data (Hedeker *et al.*, 2006).

In applications to the medical field, an example of the use of longitudinal data is in clinical trials where interest lies in drawing conclusions about the effect of different kinds of treatments over a particular time period. In clinical trials, groups of subjects are selected to participate in different kinds of treatments over a particular time period. In this way, these studies supply information about within individual change and allowing the investigator to control for individual heterogeneity (Hedeker *et al.*, 2006) and refer to changes to individuals' characteristics over time (Fitzmaurice *et al.*, 2004).

The advantages of longitudinal data suggest that longitudinal studies appear to be more powerful than cross-sectional studies (Diggle *et al.*, 1996). Additionally, since longitudinal data is already clustered as the data is obtained from taking repeated observations from a single individual at different times, exact estimates of change can be achieved. This is due to observations in clustered data usually displaying a positive correlation over time meaning that measurements closer together in time show higher correlation than measurements further apart in time (Fitzmaurice *et al.*, 2004). Longitudinal data is a special case of clustered data where the orders of the measurements have a significant effect due to the positive correlation between the measurements taken from the same subject (Fitzmaurice *et al.*, 2004). The problems of correlation can be overcome by taking account of within individual change using a suitable methodology.

Longitudinal studies also have some disadvantages. The vital assumption of independence between observations of the response which underpins most common statistical techniques cannot be taken in longitudinal data. The assumption is violated because observations taken from the same individual are dependent on each other and cannot be assumed to be independent. (Hedeker *et al.*, 2006).

Furthermore, the measurements taken from the same individual are observed in an ordered time sequence and even the time interval between the two measurements does not have to be equally spaced, obtaining these measurements repeatedly over time results in the data being correlated. Therefore according to Diggle *et al.*, (1996), if this correlation is disregarded, at least three main problems arise including the interpretation of regression parameters in the usual way being invalid, estimation of regression parameters being impractical and bias being incurred due to inadequate description for missing data.

The correlation between the repeated measurements is caused by three sources of correlation, namely between subject heterogeneity, serial correlation and measurement error (Diggle *et al.*, 1996; Fitzmaurice *et al.*, 2004; Molenberghs and Verbeke, 2005). Between-subject heterogeneity is the variation occurring in the response due to considering different subjects. The response obtained from each subject under the same controllable conditions can be different due to the unmeasured factors (random factors) mentioned earlier. Serial correlation results when the repeated measurements observed from the same individual are dependent on the time interval between the two adjacent measurements (Fitzmaurice *et al.*, 2004). There is a negative relationship between the time interval of two adjacent measurements observed from a single individual and the correlation between those two measurements. This means that two repeated measurements taken from the same individual close together in time have a stronger correlation compared to two repeated measurements from the same individual taken further apart in time (Diggle *et al.*, 1996). In such cases, this correlation between the repeated measurements within an individual results from carryover effect. The *carryover effect*, which is also called as *sequence effect*, occurs when different measurements of the same response are obtained from the same individual at different time points. In this way, the order of the time points might have impact on the response achieved at the end of the study. For instance, in clinical trials where the individual is exposed to different treatments at various time points, the response from one treatment might be conditional on the response from the previous treatments. Therefore, dealing with the issue of carryover effect is important and requires the researcher to use suitable methodologies to conclude strong statistical inferences (Fitzmaurice *et al.*, 2004). However, analytical methods are not developed well enough in the area of longitudinal studies and there is a lack of availability in computer software to deal with longitudinal data (Hedeker *et al.*, 2006).

In most studies, besides between-subject heterogeneity, within subject differences are observed as a result of combination of time-dependent serial correlation and measurement error. Since the availability of data is usually limited, the two separate sources of within subject variability cannot be analysed separately and therefore this type of variability is considered to be a combination of serial correlation and measurement error. In this way, studies can take account two different sources of heterogeneity as being within individual differences and between individual differences. While the capability of longitudinal studies to allow for within subject correlation is an advantage, this correlation violates the usual independence assumption. Hence, more sophisticated statistical methods must be used in the analysis of longitudinal data sets. A further difficulty in analysing such data increases when the data is unbalanced (Gibbons *et al.*, 2010). Unbalanced data usually arise if the measurement times are not the same for all subjects or if an unbalanced structure due to missing observations for some subjects occurs.

Missing data clearly leads to problems in the analysis of longitudinal data and occurs when an individual drops out of the study before the endpoint, this is commonly referred to as attrition. Attrition is usually the main reason for missing values (Hedeker *et al.*, 2006) and can be due to many reasons. For example, in clinical trials experiencing a detrimental side effect from the applied treatment, death of the patient, gaining full benefit from treatment and believing that they will not add any additional benefit to the study, etc. may be causes of attrition.

(Goldstein, 2009). Further sources of missing data, referred to as *incomplete data* occur when an observation is not recorded, for instance if an individual misses the planned appointment time, which results in having mistimed measurements. In this situation, data will be *unbalanced* over time (Fitzmaurice *et al.*, 2004). In longitudinal studies, even in data from randomized and well-controlled clinical trials, the problem of missing values is still highly relevant. According to Fitzmaurice *et al.*, (2004), having a balanced and complete longitudinal dataset is an uncommon case in the health sciences.

Some relevant examples of approaches to dealing with missing data include *completer analysis* which only takes account of the subjects who have completed the study. However, when the completer analysis is used, the size of the subject sample at the end of the study will not be the same as it was initially and will also be biased towards *survivors* (i.e. will ignore those who have died). An alternative approach used to mitigate the effects of attrition is called *Last Observation Carried Forward (LOCF)* (Hedeker *et al.*, 2006). At the point of any subject's discontinuation, the LOCF approach estimates of the subsequent measurements. However, in this way, the LOCF approach introduces some problems. For instance, it assumes that every individual subject is same and will have the same influence to the treatment over the study period. Furthermore, the study ignores that if the subject had continued in study, the actual response of that individual subject might be very different from the response value just being carried forward (i.e. constant) from last observation at the point of discontinuation (Hedeker *et al.*, 2006).

An additional problem commonly occurring in longitudinal studies is the problem of time-varying covariates. This means that when any individual is measured over time, change may be observed in the value of some of the predictors, which are the independent variables, as well as in the value of outcome. For instance, a patient's weight and/or blood pressure will tend to vary over time. The aim of the study is to estimate the correlation between the predictors and outcome variables. However, this dynamic relationship takes place within individuals and therefore may differ from individual to individual. This variation adds complexity to the statistical model and requires more sophisticated analysis than cross sectional studies (Fitzmaurice *et al.*, 2004).

### 2.2.4 Approaches to analyse longitudinal data

One of the simple univariate analysis methods that can be used to analyse repeated measurements is *derived variables analysis* (Diggle *et al.*, 1996; Fitzmaurice *et al.*, 2004; Hedeker and Gibbons, 2006). This is a simple analysis that can be effective in certain cases. The approach is used to combine repeated measurements into summary measures by techniques such as either taking the mean response over time, measuring the linear trend over time, using the last observation carried forward method, using the changes in the measurements or by evaluating the area under the response curve (Hedeker and Gibbons, 2006). Other traditional techniques that are used in the analysis of cross-sectional data, such as Analysis of Variance (ANOVA), can be used to analyse the data by measuring the difference between the group means or by examining the effect of covariate on the response (Diggle *et al.*, 1996). The effect of covariate on the response is investigated by ANOVA when there are two repeated measurements where the difference is taken as an outcome measure (Hedeker and Gibbons, 2006). On the other hand, derived variable analysis cannot deal with unbalanced or incomplete data and cannot model time-varying covariates due to the heteroscedascity (i.e. having different amount of information per subject) caused by the structure of the data.

According to Fitzmaurice *et al.*, (2004), two classical approaches exist to analyse Gaussian (Normal) longitudinal data, namely; repeated measures analysis of variance (ANOVA) which is also known as univariate or mixed model ANOVA, and multivariate repeated measures analysis of variance (MANOVA).

ANOVA models allow correlation to exist between repeated measurements from the same subject. In these models, a random intercept is used. This means that the models allow each individual to have different initial values of the dependent variable. However, exclusion of random slopes from the model, restricts the model. In this way, the model use a compound symmetry covariance structure, which assumes constant variances and covariances over time. Therefore, the slope is assumed to be constant for any one subject. These models do not take serial correlation into account, cannot deal with incomplete and unbalanced datasets and cannot handle continuous covariates (Hedeker and Gibbons, 2006). Therefore, time has to be included in the model as a categorical variable.

MANOVA is an alternative technique which expands the ANOVA method, so that the approach can be used on multivariate response data (Fitzmaurice *et al.*, 2004). This approach can also be used to analyse longitudinal data and will be more applicable compared to ANOVA models because MANOVA allows the modelling of covariance structure in a more adaptable manner (Gibbons *et al.*, 2010). For instance, modelling the covariance structure as unstructured rather than compound symmetry. However, the major restriction required for this approach is that the data have to be complete. Hence, MANOVA is not applicable for use on unbalanced data and cannot model continuous covariates (Fitzmaurice *et al.*, 2004). As ANOVA based techniques have too many constraints and hence cannot address the main aims of longitudinal data analysis, they are not the preferred methods used to analyse such data sets.

#### 2.2.4.1 Regression Models for Gaussian Data

Since the simple approaches described above are not really applicable for use in the analysis of longitudinal data, more sophisticated regression models have been developed. In order for an advanced model to be appropriate to use in real-life practice, it firstly needs to deal with incomplete and unbalanced data in a natural way, then the model should also be able to model time-varying covariates. Furthermore, the model should consider three sources of variation, namely; between subject variation, serial correlation and measurement error to model the covariance structure in a flexible manner and, finally, the model should be able to evaluate the within individual differences.

If the underlying distribution of the data is assumed to be Gaussian (Normal), statistical models that fulfil the criteria mentioned above and are suitable for the analysis of Gaussian longitudinal data include full multivariate models and linear random effect models. These models were first studied by Laird and Ware (1982). The differences between these two approaches arise when the correlation between the repeated measurements is taken into account. The full multivariate models are also known as marginal multivariate models. On the other hand, various random effects models are also known as multi-stage random effects models which were studied by Laird and Ware (1982) and Diggle *et al.*, (1996), hierarchical linear models which were studied by Davidian and Giltinan (1995), subject-specific models which were studied by

Molenberghs and Verbeke (2005), linear mixed models studied by Molenberghs and Verbeke (2005) or mixed-effects regression models which were investigated by Hedeker and Gibbons (2006).

When a deeper investigation is performed on how the two approaches handle correlation between repeated measurements, it can be concluded that full multivariate model and random-effect model are quite similar. In both models, the response is assumed to follow a multivariate normal distribution. In the full multivariate model, the mean response vector and covariance matrix are modelled separately by only looking at the relationship between repeated measurements within a subject, whereas in the random effect model, covariance is modelled by including random terms that can account for subject-specific differences occurring due to unmeasured factors. In this way, in a full multivariate model, the response vector is assumed to have the same form for all subjects, while in random-effect models, regression coefficients are allowed to be subject-specific. Both models can deal with unbalanced and incomplete data. In the case when missing values are assumed to be missing completely at random (MCAR) or missing at random (MAR), than both models are robust to missing data. Both models allow the inclusion of time varying covariates and modelling continuous variables. However, both models have certain weaknesses. The drawback of full multivariate model is the computational complexity, because analytical evaluation depends on the dimension of the covariance matrix. However, when full multivariate models are converted to semi-parametric models, this burden of computational complexity can be overcome and hence robust estimations can be produced. However, such semi-parametric models are only adequate if the emphasis of the researcher is to concentrate on the mean parameters over the sample. Even if the interest is only on the mean parameters, using these methods will reduce the precision and can incur a bias for incomplete data. Furthermore, when the interest lies in estimation of subject-specific responses, full multivariate methods will not be applicable as they cannot decompose the total heterogeneity into within subject differences and between-subject differences. On the other hand, random effect models do separate the total heterogeneity and quantifies these between-subject differences and within subject differences. When random effect models are compared with full multivariate models, random effect models usually require fewer parameters and hence are computationally more favourable than full multivariate methods. Despite the differences between these two methods, they can often result the same fixed effect parameters since the difference between them lies in the covariance matrix formulation and only affects the random effect part of the model.

#### 2.2.4.2 Mixed Models

When the response is only observed from one measurement (i.e. not repeated over time), simple linear models can produce similar results to linear mixed models. This is because in the case of one response measurement, there are no multiple observations from the same subject available to investigate the correlation between the measurements within a subject. Hence, the error term in the model is not divided into two separate random effect and measurement error terms, instead the error term only includes the measurement error in both these models in such situations. However, when the response contains multiple measurements for each subject, the error term is analysed as two separate terms and random effects are added to the linear models in order to form linear mixed models. In both linear and linear mixed models, the response is assumed to follow a Gaussian (normal) distribution. In linear mixed models, in addition to the assumption on the distribution of the response variable, the response is assumed to be linearly related to the other covariates, subjects are assumed to be independent, and both random effects and random errors are assumed to have mean zero and constant variance. It is also assumed that random effects and random errors are uncorrelated. The second main difference between these two types of models is that linear mixed models allow the modelling of the hierarchy involved in the structure of the data. In linear mixed models, because the response will contain multiple measurements from each subject, multiple repeated measurements are considered to be nested by subject. Hence, a hierarchy is involved when considered as individual measurements at the lower level and the subjects at higher level. In certain cases, more than two level structures are also possible for instance in social sciences when students are nested in classes and classes are nested in schools or in primary health care when measurements are nested in subjects and subjects are nested in practices. Therefore, inclusion of the hierarchical structure prevents measurements from being independent and hence simple linear models with the independence assumption cannot be used to model such datasets. Linear mixed effects models take account of this correlation between the measurements by incorporating the random effects into the model.

There are two different types of linear mixed effects models namely; random intercept models where the random effects vector only has a random intercept and a fixed slope and a random intercept and random slope models where both the intercept and slope can vary between subjects. Including a random intercept in the model will allow each subject to have different initial values, so each subject will have its' own deviation from the group mean. Additionally, including a random slope in the model will allow each subject to follow a different linear trend over time from the others. In this way, each subject can have its' own deviation from the group trend over time.

In most cases, the individual subjects are not the main interest. Knowing the individual deviations from the group mean and group trend over time will help in estimation of the heterogeneity between subjects and will affect the evaluation of fixed effects parameters. Fixed effects parameters are estimated either by (restricted) maximum likelihood approach or Bayesian approaches (Verbeke & Molenberghs, 2000). These models consider the correlation between the repeated measurements taken from the same individual in the case of multiple response measurements, but different subjects are still assumed to be independent from each other. The theory of these models will be explained in chapter 5 together with the application to the dataset used in this project.

#### 2.2.4.3 Regression Models for Non-Gaussian Data

Full multivariate and random effects models can be extended to analyse non-Gaussian longitudinal data, where the response from each subject is not normally distributed but has a known distribution which belongs to the exponential family such as the Gamma distribution. These extended models are called generalized linear models (GLMs), and they were first studied by McCullagh and Nelden (1989). As for longitudinal Gaussian data, there are several sub-types of generalized linear models such as marginal models, subject-specific models and conditional models.

Marginal models are the generalisation of full multivariate models which are used in the Gaussian case. Therefore, marginal models also model the covariance structure and measurement error separately while examining the correlation between repeated measurements for an individual subject. In marginal models, the *maximum likelihood approach* (maximisation of the actual log likelihood function of the data) is used as the parametric approach to model the covariance matrix. On the other hand, semi-parametric approaches that rely on specification of the first two moments can be used to model the covariance structure and such models include *generalized estimated equations (GEE)* (methods that focus on *population-average* effects and hence cannot be investigated in this study) and *pseudo-maximum likelihood methods* (also known as quasi-likelihood methods that maximise a function that is related to log likelihood of the data). GEE methods were studied by Liang and Zeger (1986) and pseudo-maximum likelihood methods were studied by Molenberghs and Verbeke (2005).

Subject-specific models are the generalized version of random effects models, where the correlation between repeated measurements can be modelled in the same way as in the random effects models in the Gaussian case. Subject-specific models are called random effects models, generalized linear mixed models (GLMMs) or hierarchical models.

Conditional models were first studied by Diggle *et al.*, (1996). In conditional models, the association between the predictor variables and the response is investigated, how the response depends on predictor variables is examined and both the association and the dependence are modelled simultaneously in the same equation using the previous responses. As well as non-Gaussian data, conditional models can also be applied on Gaussian data.

Comparison of these three models (conditional models, subject-specific models and marginal models) when applied to both Gaussian and non-Gaussian data sets, shows the difference in interpretation of fixed effect parameters in both cases (Diggle *et al.*, 1996). When applied to Gaussian data, all three models result the same interpretation of fixed effects parameters. However, in non-Gaussian data, the interpretation of fixed effect is different in each of these three types of models due to use of various different assumptions about the source of correlation between the predictor variables and response. In marginal models, fixed effects measure population-averaged effects of predictor variables on the mean response (i.e. grand mean), while in random effects model, fixed effects measure the effect of the independent variables on the mean response per subject (i.e. group mean) but with subject-specific random effects. On the other hand, in conditional models, fixed effects measure the effects of the predictor variables on the mean response that is conditional on the previous responses of all subjects. Since the interpretation of fixed effects parameters in conditional models are more difficult than for marginal and random effects models, marginal and random effects models are more widely used in practice than conditional models (Molenberghs and Verbeke, 2005).

#### 2.2.4.4 The differences between Generalized Linear Mixed Models (GLMMs), Generalized Linear Models (GLMs) and Linear Models (LMs)

Generalized linear models (GLMs) are the extension of simple ordinary linear models (LMs) where (in the former), the response variable does not have to follow a Gaussian (normal) distribution as is the case in the latter type of model. However, instead of a Gaussian distribution, the response should belong to one of the exponential family of distributions.

Linear mixed models (LMMs) are the extension of simple linear models where the distribution of the response is assumed to follow a Gaussian distribution as in the simple linear models. The main difference between LMs and LMMs is that the latter includes random effects into the model formulation in addition to former model.

Generalized linear mixed models (GLMMs) are the extension of generalized linear models where the assumption about the distribution of the response variable is same as for generalized linear models. The difference between generalized linear models and generalized linear mixed models is that in GLMM, random effects are included in the model, as in linear mixed effects models (LMMs). Therefore, in GLMMs, similar to LMMs, there are two parts, namely; fixed effects and random effects. The error term also has two parts, namely random error and measurement error. Random error is included in order to take account of the error due to unobservable but significant factors such as genetic factors. Since GLMMs make the same assumption as GLMs about the assumption of the response variable, the former models also assume that variance of the response can be defined as a function of the mean and assume that the response has been transformed by using some link function. In models discussed in this section, the relationship between the response and the independent variables is assumed to be linear. However, in practice this relationship is not always linear. In such cases where the relationship is non-linear, the response is first transformed, so that the association between the mean transformed response and the independent variables become linear. This transformation is achieved by using a suitable link function and then GLMMs can be used. Assumptions about the random effect and random errors are still kept the same in GLMM as for LMM and subjects are still assumed to be independent. GLMMs which are the extensions of GLMs, can be treated as a special case of mixed models, containing random effects in addition to fixed effects. As for LMMs, in GLMMs, fixed effects parameters are usually estimated by the maximum likelihood approach by integrating over the random effects. If any missing cases are involved in the data, these are assumed to be either missing at random or missing completely at random.

Furthermore, if the research question can only be answered by modelling the original data or if the association between the mean transformed response and the independent variables is still non-linear after transformation, then marginal and random effects models are no longer effective methods for use in the analysis. For this reason, in order to model this non-linear relationship, marginal and random effect models are extended and generalized. The extensions of random effects models can be more widely used and are called non-linear mixed models. These non-linear mixed model were studied by Davidian and Giltinan (1993, 1995, 2003), by Vonesh and Chinchili (1997) and by Molenberghs and Verbeke (2005). Several authors have identified the effectiveness of such models in health care such as in the pharma kinetic field when analysing the rate of clearance of a drug and in the analysis of growth or decay rates.

#### 2.2.4.5 Generalized Additive Models (GAMs) and Generalized Additive Mixed Models (GAMMs) as non-linear mixed models

Generalized additive models (GAMs) are also another extension of generalized linear models. The model assumptions are the same as for generalized linear models, except for the linearity assumption being removed. This is the main difference between generalized linear models and generalized additive models. The link function which was in the form of linear predictor in generalized linear model is replaced by sum of smooth functions of the predictors in generalized additive models. These smooth functions of the predictors allow a flexible approach to the relationship between the predictor variables and the response. In this way, the linearity assumption became a nonparametric version generalized linear models. The smooth functions are evaluated using the same data as for the simpler models. Therefore, in order to obtain reliable estimates of parameters, a large number of data points are required by generalized additive models, which also increases the intensity in computation of such models.

In GLMs, a single coefficient is estimated for each independent predictor. However, in GAMs, an unspecified, nonparametric function is estimated for each predictor by using a combination of smooth functions. Smooth functions estimate the non-linear relationship between the predictor and the response by using a large number of data points. As a result, a nonparametric function is obtained to describe this relationship. In generalized additive models, the fixed effects parameters are modelled using entirely parametric models as in GLMMs; whereas when smooth functions are used in some covariates, those covariates are modelled using the unspecified, nonparametric function obtained from smooth functions. These unspecified, nonparametric functions are associated with the response variable through a link function (Hastie and Tibshirani, 2009). According to Wood (2006a), GAMs can be estimated by using penalized regression spline methods. In this way, violating the linearity assumption and taking account of the non-linear relationship will increase the accuracy of prediction of the response.

Generalized additive mixed models (GAMMs) are the extension of generalized linear mixed models (GLMMs). In GAMMs, as in GLMMs, the response variable is allowed to have any distribution that belongs to the exponential family. This means that the response can follow a non-Gaussian distribution given that this distribution is from the exponential family such as gamma distribution. In GAMs, fixed effects are modelled by the parametric component. However, the addition of random effect terms to a generalized additive model turns the model into a generalized additive mixed model and in GAMMs, the fixed effects are modelled in a non-parametric way by using a combination of smooth functions. In GAMMs, the response variable is no longer assumed to be linearly dependent on the predictor variables. Another advantage of GAMMs is that these models allow a flexible covariance structure for the random effects. In this way, these models can be seen as desirable for the analysis of longitudinal datasets having a clustered, hierarchical or spatially correlated data structure. In GAMMs, as in GAMs, smoothing functions can only be estimated for continuous predictors.

Using the frequentist approach, in GAMMs, estimation of non-parametric functions and smoothing parameters are carried out using *restricted maximum likelihood (REML)* approach (that uses likelihood function calculated from transformed dataset, so nuisance parameters such as mean and variance will not have an effect) if the response follows a Gaussian distribution. On the other hand, if the response follow a non-Gaussian distribution, *penalized quasi-likelihood (PQL)* (estimates based on 1$^{st}$ order Taylor series) or *double penalized quasi-likelihood (DPQL)* (involving two different types of estimates; one for random effect and one for smoothing parameter) approaches (where in both PQL and DPQL approaches, relationship between mean and variance is specified as a function of mean by likelihood function) are used in the estimation of the non-parametric functions and smoothing parameters. In terms of link functions, an identity link is assumed for Gaussian responses. The reason for choosing REML over ML estimation for a Gaussian response is because a ML estimator is usually biased when used in the estimation of variance whereas a REML estimator is reliable and asymptotically normal even if the response variable does not follow a Gaussian distribution. In most cases where the response is non-Gaussian, the likelihood function cannot in practice be obtained by using a REML estimator due to the computational burden. Therefore, the likelihood is approximated using the *Laplace approximation* (technique used to approximate the integral) and the approximated likelihood is assumed to be the true likelihood. This estimated likelihood is called penalized quasi-likelihood (PQL) and in non-Gaussian responses, either penalized quasi-likelihood or double penalized quasi-likelihood (DPQL) is used in the process of estimating the smoothing spline.

Figure 2.2 shows the different methodologies explained above (see section 2.2) where the highlighted approaches are selected and the data in this research are analysed by using those highlighted approaches. The reasons for choosing these approaches are explained in detail in chapters five and six.

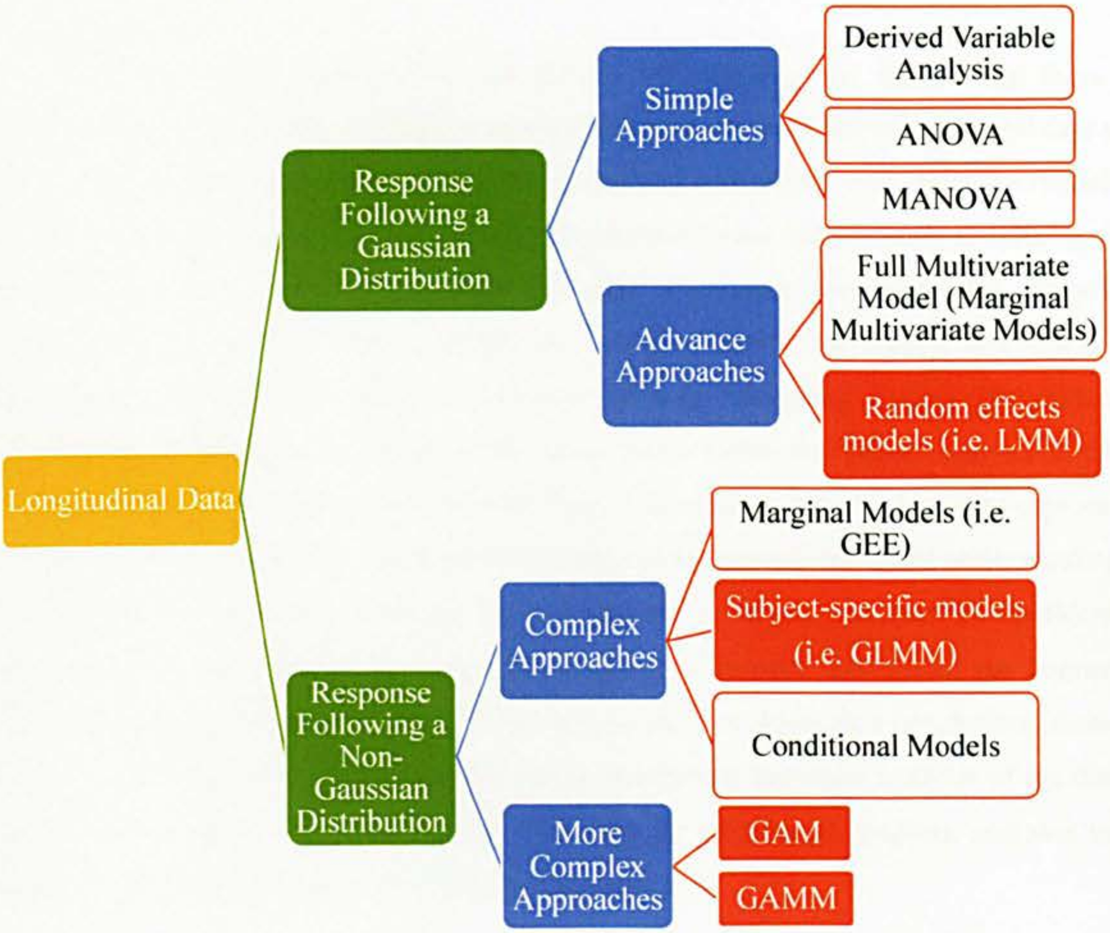
Longitudinal Data
Response Following a Gaussian Distribution
Simple Approaches
Derived Variable Analysis
ANOVA
MANOVA
Advance Approaches
Full Multivariate Model (Marginal Multivariate Models)
Random effects models (i.e. LMM)
Response Following a Non-Gaussian Distribution
Complex Approaches
Marginal Models (i.e. GEE)
Subject-specific models (i.e. GLMM)
Conditional Models
More Complex Approaches
GAM
GAMM

Figure 2.2: Selecting different methodologies to analyse longitudinal data.

# 3 Description of Data and Preliminary Analysis

## 3.1 Introduction

There are several stages within the process of analysing any dataset and these are particularly important in the investigation of data from real life and routinely collected data such as the GP records used here. As the data was not collected specifically with statistical modelling (or even basic statistical analyses) in mind, it is prone to many pitfalls such as missing data, incomplete records, dropout etc. When the data is cleaned and ready for analysis a developmental approach is taken whereby we initially examine the dataset and assess any limitations and/ or problems within it which may affect the modelling process. The next stage is to build up a gradual understanding of the associations within the dataset, relevant to CKD, before proceeding to complex statistical modelling; this is usually achieved via data exploration and examination of some basic descriptive statistics which provide an initial understanding of the relationships that exist within the data. From this point we can then be confident in proceeding to more complex investigations which can broadly fall under the banner of inferential statistics which are used to derive deeper, more sophisticated conclusions from the data. In this chapter, the process of investigative descriptive statistical analysis of the data is presented along with a description of the data used. In subsequent chapters, analyses using inferential statistics and the results thereof are reported.

## 3.2 Data

The data used in this thesis was collected from a sample of 129 UK primary care based General Practices. The UK primary care setting is a registration-based system were patients should be registered with only one general practice. General practices are computerized and use electronic patient record (EPR) systems. Quality of the routinely collected data in practices has greatly improved since the UK Primary care P4P reward scheme was introduced in 2004. CKD measures were added to the P4P scheme in 2006.

Different practices use various brands of EPR systems, and a common data extraction platform, MIQUEST- Morbidity Information Query and Export Syntax was used to combine data from different practices. The data used in this thesis was extracted from general practice

EPR systems using MIQUEST (Michalakidis *et al.*, 2010) and aggregated using well established methods (van Vlymen *et al.*, 2005). After the data was combined from different practices, an online dictionary was created to provide details for each variable in the dataset (Clinical Informatics, 2010). For inclusion in this dataset, practices had to have been using the same EPR system for at least 5 years. The data used in this research comes from the first round of follow-up data collection for the Quality Improvement in CKD (QICKD) trial study. The first round of follow-up data for the QICKD trial study was collected from 1 June 2008 to 26 February 2009. Out of 138 practices interested for the data collection, 4 practices subsequently withdrew from the study where 3 of those practices had concern about the capacity of their server to cope with running complex MIQUEST queries and its potential impact in clinical services. As well as this, 2 practices could not be included due to late consent and further 3 practices could not be taken in this study due to practices being outside the primary care trust where research and development office had not agreed to involve in the study. In total 129 general practices participated in the data colletion. Demographic and other pateints details as well as diagnoses of selected co-morbidities were identified and labelled prior the aggregation of data from the different general practices. Subsequently the MDRD equation was used to evaluate eGFR readings and a diagnosis of CKD is recorded where eGFR<60mL/min/1.73m$^2$ based on two readings which are at least 90 days apart. It is this derived dataset that is used for the analyses presented in this thesis.

The availability and importance of routinely collected data is growing as data quality awareness and technological progress enable the improved processing of large datasets. . There are now several large scale health based databases available within the UK, including the Clinical Practice Research Datalink, CPRD, (which includes GPRD (General Practice Research Database). These provide a huge array of data and information that can be extracted for medical studies. However, accessing the CPRD database can be very costly and hence there was no provision for such cost in this research project. The dataset used in this thesis was available free of charge but the structure and information derived from it would easily be replicated using other General practice resources. Furthermore, the statistical applications applied to this data would be equally applicable to data derived from the other availbe sources. One advantage of the data used here is that at the outset it provided much fresher data such as data from 2003/2004-2008/2009 than could have been obtained by CPRD.

## 3.3 Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) is carried out by obtaining some descriptive statistics from our large and complex GP records data set using mostly graphical representations for examination of the data to reveal the underlying structure of the data, explore significant variables, and to discover outliers and abnormalities. EDA also allows testing of any underlying assumptions made about the data set, leading towards the building of appropriate models and aiding decision making in regard to the best parameter settings for the modelling process.

Variables of interest in this dataset are a combination of continuous and categorical variables, and hence a variety of location, dispersion and shape statistics (means, medians, modes, standard deviations, skewness and kurtosis) are used for exploration.

## 3.4 Data Cleaning and Descriptive Statistics

The longitudinal data set used in this research is set of 129 practices in England and Wales containing 1099262 (about 1.1 million) patients in total. The initial dataset consists of all patients who were registered with and have GP records information from those 129 practices. All available information about patients including; general information about that patient, diseases that they have, prescribed medications, results of relevant blood tests such as serum creatinine, and lifestyle factors such as smoking habits were taken. Data relating to these factors is available for a period of approximately 10 years from 1998/1999-2008/2009, although the full time period is not available for all patients. However the data is a relaiable sample for the period 2003/2004-2008/2009.

As with any real world dataset, the data must be cleaned and validated before any analyses because selecting and cleaning of variables are essential in such huge datasets to ensure that the data used for the analysis is as precise and accurate as possible, so that the data can represent the general population and the results found from the analysis if the data can be interpreted to general population as a whole.

The statistical analysis is based on sample which consists of a group of individuals taken from the whole population. Firstly, variables of interest are identified and some preliminary descriptive statistics are computed for those in order to validate that the sample dataset was a good representation of the whole population of England and Wales. These statistics provide detailed information about the sample taken from the population. If the sample reflects the same patterns as the population, information obtained from summary statistics should also be applicable to use regarding the population as a whole.

For validation purposes, the first step is to compare the "population pyramids" of the GP dataset with 2011 census population of England and Wales in terms of demographic trends such as age, gender and ethnicity, represented in Figure 3.1, Figure 3.2 and Figure 3.3 respectively.

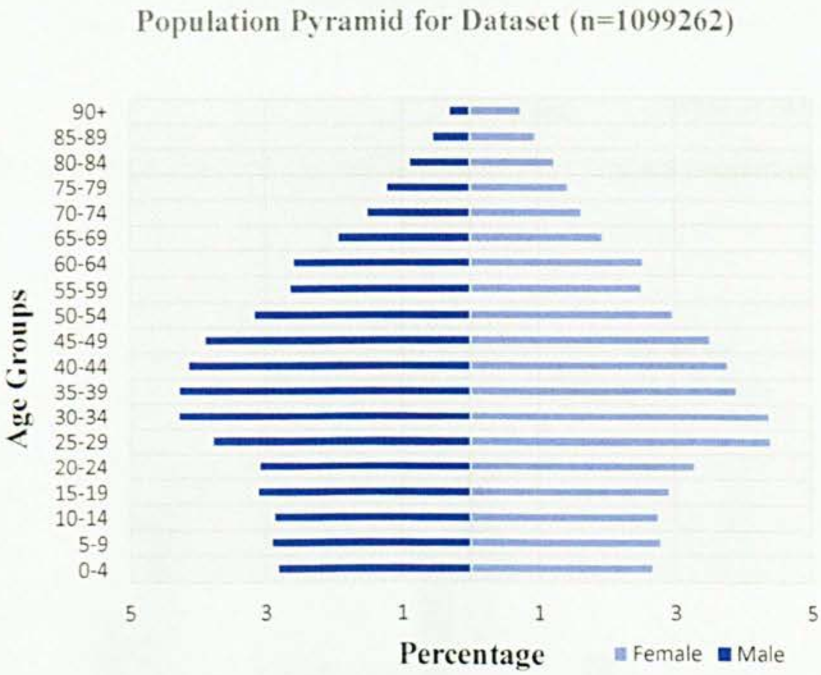

Figure 3.1: Population Pyramid by age and gender of GP dataset

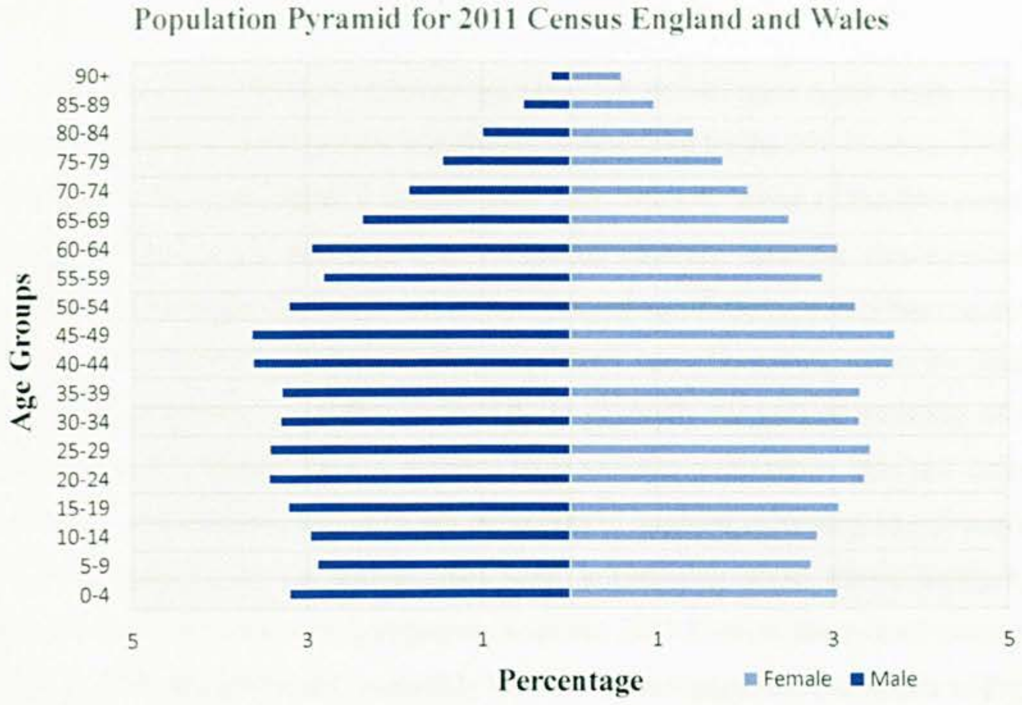

Figure 3.2: Population pyramid by age and gender of 2011 Census population of England and Wales

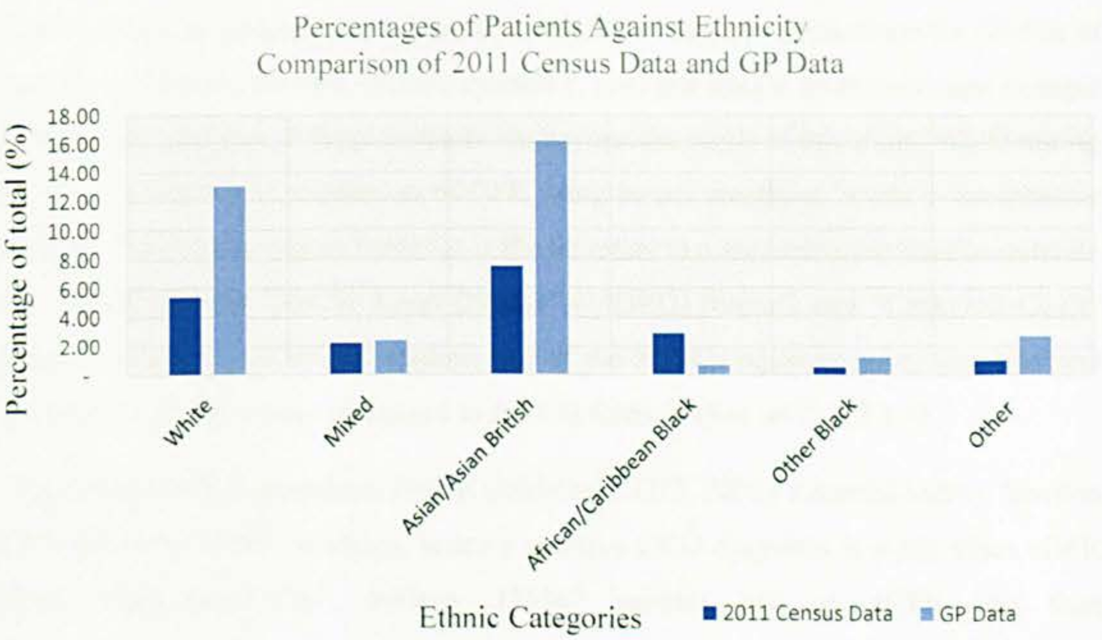

Figure 3.3: Percentages of Patients by Ethnicity Comparison of 2011 Census Data and GP Data[1]

[1]This chart excludes White British (80.5 % of population from Census 2011 Data and 63.89 % of GP Dataset.

The dataset extracted from routinely collected GP records used in this study reflects the age - gender distribution of the general population as described by the 2011 Census for England and Wales quite well. (See Figure 3.1 and Figure 3.2) The UK is one of the few countries in Europe where there are no restrictions on collecting ethnicity data for service-monitoring reasons, unlike for example, Germany and France. Recording of ethnicity data has recently been added to routine coding of GP patient data in primary care. However, while the quality of routinely collected primary care data in the UK is generally recognised as being of a high quality, recording of ethnicity data at practice level remains ineffective. This is a recognised issue in primary care datasets, and there are initiatives to improve recording on a practice level for monitoring reasons. In the dataset used here, information about ethnic background is recorded for 43.8% of patients. In comparison with the 2011 Census, the overall demographic structure of the GP data compares favourably with the known population structure with respect to age gender and ethnic group, and thus provides a reliable sample of the UK population from which to examine and interpret health indications and trends (See Figures 3.1, 3.2 and 3.3).

Selection of the sub-population with CKD was achieved by flagging all patients with a positive CKD diagnosis as described in chapter 2, based on results of a blood test for GFR or of a urine test. Other markers for GFR include cystatin c, urea and insulin levels and radio isotopic methods. However, analyses of these methods are beyond the scope of this study. While testing the actual GFR is expensive, estimation of GFR using serum creatinine levels is inexpensive and acceptable. There are two main formulas in the literature that are commonly used to estimate GFR, the "Modification in Diet in Renal Disease" (MDRD) formula and "Cockcroft-Gault" (CG) formula. According to several studies, use of the MDRD equation gives less bias and greater accuracy in eGFR values compared to the CG formula (See section 2.1.3)

According to NICE guidelines (NICE Guideline CG73, 2008), a normal kidney function has eGFR 90mL/min/1.73m$^2$ or above, hence a positive CKD diagnosis is made when eGFR falls below 90mL/min/1.73m$^2$. Initially 325367 patients had an eGFR less than

90ml/min/1.73m$^2$ whereas 58675 patients were found to have eGFR below 60ml/min/1.73m$^2$, which correspond to the most common point of CKD diagnosis (i.e stages 3-5).

In order to select study participants, it is necessary to know CKD diagnosis information for each patient. In order to keep the sample size at maximum, CKD diagnosis information is created by including both the GP diagnosis information and diagnosis using the most recent eGFR value. Some patients have only one of these type of information and some have both in their records. When the patients have both kind of information, the most recent information is used to determine the diagnosis result. The value used for diagnosis is either an eGFR value calculated by GP from serum creatinine (SCr) value using MDRD formula or an eGFR value obtained from the routine laboratory analysis for that patient which is calculated by using MDRD formula. For the purpose of the study, only the patients that have been diagnosed to have CKD at stage 3 to 5 are considered in the analysis. The initial population of the dataset was about 1.1 million (exactly 1099262) patients. However, as explained below, data cleaning reduced this number to 876951 patients as a study population, and the figure dropped to 30490 patients after taking only CKD patients at stages 3-5 whose have at least two repeated measures of eGFR values during the period of the study.

As CKD is primarily a disease of adulthood, the sample was restricted to patients over the age 18 and since the maximum age for a UK resident to date is 105, an upper age limit of 105 was adopted. Patients outside of this age range were removed from the sample. Data is also cleaned on the main risk factor measures for CKD; for example obesity, measured by body mass index (BMI). BMI ranges are validated against national standards, and outliers removed. Where possible, missing BMIs are imputed using (cleaned) records of heights and weights. Similarly, coding systems used to indicate variables such as ethnicity and smoking are reviewed and corrected. Some continuous variables relating to disease status are transformed to categorical variables based on the national guidelines such as using systolic and di-systolic blood pressure values to generate categories corresponding to hypertension status of the patient. These categories can be either nominal or ordinal format based on the particular variable such as having an ordinal categorical variable in hypertension status of the patient.

For all variables measured on a continuous scale (and retained as such), the following 4 step cleaning and validating procedure is employed. In some datasets, the units used might not be clearly defined and consistently used. Therefore, before starting this procedure, it is essential to decide which unit is to be used for that variable, (considering the mean, mode, median, maximum and minimum values) is a vital stage to follow.

1- Finding the reportable range for each variable for the specific unit by using laboratory manuals for the equipment measuring those variables in the appropriate test conditions.

2- Eliminating values outside the reportable range and re-calculate mean, mode and median values. For instance, a height of a patient might be recorded as 0.0173 m or as 183 m where it is clearly seen that the decimal point is recorded in the wrong place. Therefore such values must be eliminated from the study.

3- Finding outliers by using a Mean±4 standard deviation (SD) criterion. (See range of values and units used for particular variables in Table 3.1).

Table 3.1: Ranges used in final cleaning of data from blood tests where outliers are selected according to criteria of Mean±4SD from Reportable Range values which are taken from published laboratory manuals.

| Test | Range Used | Unit Used |
|---|---|---|
| Serum Creatinine (SCr) | 7.2538-155.5980 | µmol/L |
| Haemoglobin (Hb) | 7.4972-19.8896 | g/dL |
| Serum Plasma Albumin | 28.2309-58.5521 | g/L |
| Serum Triglycerides (TG) | NO LOWER LIMIT-4.9817 | mmol/L |
| Total Cholesterol | 0.7559-9.3747 | mmol/L |
| High Density Lipoprotein Cholesterol (HDL-C) | NO LOWER LIMIT-3.0837 | mmol/L |

| | | |
|---|---|---|
| Low Density Lipoprotein Cholesterol (LDL-C) | NO LOWER LIMIT-6.8410 | mmol/L |
| Glycated Haemoglobin (HbA1c) | 0.5868-12.8160 | % of Total Hb |
| Blood Glucose (Fasting) | NO LOWER LIMIT-13.5469 | mmol/L |
| Mean Corpuscular Volume (MCV) | 64.9002-113.1070 | $\mu m^3$ or fL |

4- Compare the percentage loss of data between the figures before cleaning and after cleaning.

The final step (step 4) makes sure that only the extreme values are eliminated and the amount of data lost is kept minimal. After the cleaning procedure, the distributions of continuous variables are re-created and can be observed to be closer to normal distributions. Overall, categorical variables are re-classified, continuous variables are cleaned and binary variables which are used for gender and disease diagnosis stay as before cleaning.

Summary characteristics (age and gender) of patients in the GP dataset are given in Table 3.2.

Table 3.2: Gender and age profile of the study population (n=876951).

| | | Research GP Records Database (n=876951) |
|---|---|---|
| | | n (% of total) or mean ± SD |
| Gender | Female | 443293 (50.55%) |
| | Male | 433658 (49.45%) |
| Mean age (years) | Total | 46.08 ± 18.423 |
| | Female | 46.66 ± 19.205 |
| | Male | 45.49 ± 17.569 |

| | | |
|---|---|---|
| Age groups | <20 years | 27995 (3.19%) |
| | 20-24 years | 69860 (7.97%) |
| | 25-29 years | 89776 (10.24%) |
| | 30-34 years | 94934 (10.83%) |
| | 35-39 years | 89662 (10.22%) |
| | 40-44 years | 86682 (9.87%) |
| | 45-49 years | 81309 (9.27%) |
| | 50-54 years | 67311 (7.68%) |
| | 55-59 years | 56590 (6.45%) |
| | 60-64 years | 56242 (6.41%) |
| | 65-69 years | 42420 (4.84%) |
| | 70-74 years | 34422 (3.93%) |
| | 75-79 years | 29030 (3.31%) |
| | 80-84 years | 23135 (2.64%) |
| | 85-89 years | 16375 (1.87%) |
| | ≥90 | 11208 (1.28%) |

From Table 3.2, it can be seen that in this dataset, there are about 1.1% more females than males. The overall mean age of the study population is around 46 years and the mean age of females are slightly above this average age whereas the mean age of males are slightly below this average age. From this, it can be observed that on average females are about 1.17 years older than males in this dataset. This is probably mainly due to higher number of females in the very old categories (over 80 years).

If the dataset is split into three broad categories of people according to age, where the categories are classified as young people, mature people and elderly people, then, it can be concluded that the highest proportion of the dataset lies in mature people category with 54.4% of the study population consisting of people aged between 30 and 59. On the other hand, the young people category contains people aged between 18 and 29, representing 21.4% of the whole dataset, and the elderly people category contains people aged between 60 and 90 and above, with 24.2% of the whole dataset. Thus, by studying Table 3.2, it can be stated that there are greater number of mature and elderly people in this dataset compared to the number of young people.

The GP data set was first subdivided into two separate groups, based on whether or not each patient has been diagnosed to have CKD at any time either before or during the 10 year study period which was between 1998/1999-2008/2009. The prevalence of CKD is then reported in Table 3.3 with respect to gender and age groups.

Table 3.3: The prevalence of CKD with respect to gender and age groups.

| | Total number of subjects | Subjects with CKD |
|---|---|---|
| Gender | | N (% of gender group) |
| Female | 443293 | 39619<br>(8.94% of females) |
| Male | 433658 | 19056<br>(4.39% of males) |
| Age groups | | N (% of subjects with CKD) |
| <20 | 27995 | 11 (0.02) |
| 20-24 | 69860 | 72 (0.12) |
| 25-29 | 89776 | 210 (0.36) |

| | | |
|---|---|---|
| 30-34 | 94934 | 472 (0.80) |
| 35-39 | 89662 | 730 (1.24) |
| 40-44 | 86682 | 1271 (2.17) |
| 45-49 | 81309 | 2169 (3.70) |
| 50-54 | 67311 | 2732 (4.66) |
| 55-59 | 56590 | 3306 (5.63) |
| 60-64 | 56242 | 5058 (8.62) |
| 65-69 | 42420 | 5636 (9.61) |
| 70-74 | 34422 | 6719 (11.45) |
| 75-79 | 29030 | 8417 (14.35) |
| 80-84 | 23135 | 8610 (14.67) |
| 85-89 | 16375 | 7570 (12.90) |
| ≥90 | 11208 | 5692 (9.70) |

The prevalence of CKD shows the total number of cases in this study population who have been diagnosed to have CKD. When the prevalence of CKD is compared between genders (Table 3.3) and across different age groups (Table 3.3), it can be seen that the prevalence of CKD is greater in females (8.94% of females) compared to males (4.39% of males) and results in Table 3.3 also confirm that the prevalence of CKD is greater in the older age groups.

Table 3.4: Ethnicity information of study participants

| Ethnic Groups | Research GP Records Database (n=876951)<br>Number (% of total) |
|---|---|
| White | 303462 (34.60) |
| Mixed | 9562 (1.09) |
| Asian/Asian British | 63858 (7.28) |
| Caribbean/African Black | 2622 (0.30) |
| Other Black | 4659 (0.53) |
| Other | 10588 (1.21) |
| Not Stated | 30749 (3.51) |
| Ethnicity Not Recorded | 451451 (51.48) |

The UK is one of the few countries in Europe where there are no restrictions on collecting ethnicity data for service-monitoring reasons, unlike (for example), Germany and France. Recording of ethnicity data has recently been added to the routine coding of patient data in primary care. However, while the quality of this routinely collected (primary care) data in the UK is generally recognised as being of a high quality, recording of ethnicity data at general practice level remains ineffective. This is a recognised issue in primary care datasets and there are initiatives to improve recording at the practice level for monitoring reasons. As can be seen from Table 3.4, in the dataset used here, information about ethnic background is recorded for 48.52% of total patients and from those patients where the ethnic information is recorded (i.e. from those 48.52% of total), in around 9.71% of patients, explicit information of their ethnic backgrounds cannot be obtained due to the information recorded being as not stated or other ethnic group. In this way, information known about the ethnicity of patients is limited to 43.8% of the dataset which makes it more difficult to make inferences based on ethnicity.

Table 3.5: Smoking habits of study participants

| Smoking Status | Research GP Records Database (n=876951)<br>Number (% of total) |
|---|---|
| Non-Smoker | 497120 (56.69) |
| Ex-Smoker | 129779 (14.79) |
| Smoker | 180009 (20.53) |
| Smoking Not Recorded | 70043 (7.99) |

In terms of lifestyle factors, smoking habits are taken into consideration and descriptive statistics are computed. Results from this analysis are reported in Table 3.5 and it is concluded that highest proportion of participants are non-smokers.

Table 3.6: Number of patients in the dataset with records of specified measurements and the corresponding mean measurements.

| Clinical Values | Research GP Records Database (n=876951)<br>Number (% of total) or Mean ± SD |
|---|---|
| Records of Systolic Blood Pressure | 709984 (80.96) |
| Mean Systolic Blood Pressure ± SD | 126.53 ± 16.410 |
| Records of Body Mass Index | 618303 (70.51) |
| Mean Body Mass Index ± SD | 34.7151 ± 1621.10847 |
| Records of Serum Creatinine (SCr) | 212176 (24.19) |
| Mean SCr ± SD | 81.1016 ± 14.89977 |

| | |
|---|---|
| Records of Haemoglobin (Hb) Value | 480274 (54.77) |
| Mean Haemoglobin Value ± SD | 13.6993 ± 1.53623 |
| Records of Albumin | 427116 (48.70) |
| Mean Albumin ± SD | 43.4398 ± 3.67513 |
| Records of Triglycerides | 326954 (37.28) |
| Mean Triglycerides ± SD | 1.4020 ± 0.74338 |
| Records of Cholesterol | 352124 (40.15) |
| Mean Cholesterol ± SD | 5.0606 ± 1.06494 |
| Records of High Density Lipoprotein (HDL) Cholesterol | 319268 (36.41) |
| Mean HDL ± SD | 7.4255 ± 0.41246 |
| Records of Low Density Lipoprotein (LDL) Cholesterol | 209413 (23.88) |
| Mean LDL ± SD | 3.0196 ± 0.94547 |
| Records of Glycated Haemoglobin (HbA1C) | 80066 (9.13) |
| Mean HbA1C ± SD | 6.6526 ± 1.40902 |
| Records of Blood Glucose | 299480 (34.15) |
| Mean Blood Glucose ± SD | 5.3552 ± 1.32227 |
| Records of Mean Corpuscular Volume (MCV) | 487872 (55.63) |
| Mean MCV ± SD | 89.1078 ± 5.68967 |

From Table 3.6, it can be seen that 80.96% of the patients had their systolic blood pressure recorded and since this covers the majority of the patients, it is reliable to state that an average patient has a systolic blood pressure value of 126.53, which corresponds to a normal systolic blood pressure. On the other hand, another conclusion can be made regarding BMI values, where 70.51% of the patients have records of their BMI values, and it is shown that an average patient is obese (BMI > 30). According to the results displayed in Table 3.6, comparing the mean values of blood tests against the normal reference ranges, an average patient is expected to have good levels of haemoglobin, serum creatinine, albumin, and triglycerides, HDL, LDL and MCV. On the other hand, an average patient is expected to have high values of cholesterol (which is defined to be higher than the upper limit of the normal ranges). This might be due to usage of certain drugs such as oral corticosteroids, beta blockers, oral contraceptives, thiazide diuretics, oral retinoid or phenytoin which can result in an increase in cholesterol level. When a patient has high cholesterol levels in the blood, this means that cholesterol can accumulate on the walls of blood vessels, leading to narrowing or even blocking of blood vessels and resulting in arthrosclerosis, which is hardening of the blood vessels, eventually increasing the risk of heart diseases and strokes.

An average patient in this dataset is also expected to have a blood glucose values greater than the upper limit of the normal range. This might indicate that such a patient has diabetes. High blood glucose levels can also be result from other diseases such as acromegaly or acute stress which can result in trauma, heart attack, and stroke, long-term kidney disease, Cushing's syndrome, hyperthyroidism, pancreatic cancer and pancreatitis. Usage of some drugs such as corticosteroids, tricyclic antidepressants, diuretics, adrenaline, oral contraceptives and hormone replacement therapy [HRT] drugs containing osterogen hormone, lithium, phenytoin (Dilantin) and aspirin can also increase blood sugar levels. Excessive food intake shortly before the blood test might be another cause of elevated blood glucose level.

According to Table 3.6, in this dataset, an average patient is expected to have values of glycolhemoglobin (HbA1c) higher than the upper limit of the normal reference ranges. This might be due to having high glucose levels, insufficient diabetic monitoring or iron deficiency. Therefore, high HbA1c levels typically mean that the patient is not in good control of diabetes and might consider changing their treatment plan because the patient could have a risk in developing eye disease, kidney disease or nerve damage.

Table 3.7: Prevalence of CKD and its co-morbidities within the study participants

| Clinical Condition | Research GP Records Database (n=876951)<br>Number (% of total) |
|---|---|
| Diagnosis of CKD (At Stages 3-5) | 58675 (6.69) |
| Diagnosis of Diabetes | 45355 (5.17) |
| Diagnosis of Hypertension<br>(stage 1, stage 2 and hypertensive crisis) | 155161 (17.69) |
| Diagnosis of Cardiovascular Diseases (CVD) | 246055 (28.06) |
| Diagnosis of Ischaemic Heart Disease (IHD) | 32036 (3.65) |
| Diagnosis of Peripheral Vascular Disease (PVD) | 6568 (0.75) |
| Diagnosis of Cerebrovascular Disease (CEBVD) | 19643 (2.24) |
| Diagnosis of Heart Failure | 7584 (0.86) |
| Diagnosis of Stroke | 19548 (2.23) |
| Diagnosis of Anaemia | 21294 (2.43) |
| Diagnosis of Obesity<br>(obese class I, obese class II and III) | 121755 (13.88) |
| Diagnosis of Proteinuria | 33291 (3.80) |

| | |
|---|---|
| Diagnosis of Significant Proteinuria | 3047 (0.35) |
| Diagnosis of Positive Urine | 31457 (3.59) |
| Records of Albumin-Creatinine Ratio (ACR) | 1822 (0.21) |
| Records of Protein-Creatinine Ratio (PCR) | 1756 (0.20) |

For each of the co-morbidities listed in Table 3.7, diagnosis of the clinical condition is designed to be a binary outcome where the outcome has only two possibilities; 0 if the disease is not present and 1 if the disease has been diagnosed. If the information about the diagnosis of any particular disease is missing, then it is assumed that the patient has not been diagnosed to have that specific disease.

Only CKD diagnoses between stages 3 and 5 are considered in this study. Recent studies including NEORICA (2000), Health Survey for England (2009) and QI CKD (2010) indicate that about 5-10% of adults in the UK have moderate to severe CKD (stages 3-5) and the diagnosis of CKD in this dataset is found to be 6.7% which agrees with the existing literature, so this GP dataset is appropriate for further statistical analysis regarding CKD.

Diagnoses of heart-related diseases, including cardiovascular diseases, ischaemic heart disease, peripheral vascular disease, cerebrovascular disease, heart failure and hypertension, covered the highest proportion of the study participants who had co-morbidities of CKD whereas only 5.17% of the study participants were diagnosed to have diabetes. Within these heart-related diseases mentioned earlier, cardiovascular diseases and hypertension are the most commonly diagnosed in this dataset. 13.88% of the study participants had been diagnosed to have obesity (class I-III). This suggests that the prevalence of obesity is high in this dataset compared to the diagnosis of chronic diseases such as CKD.

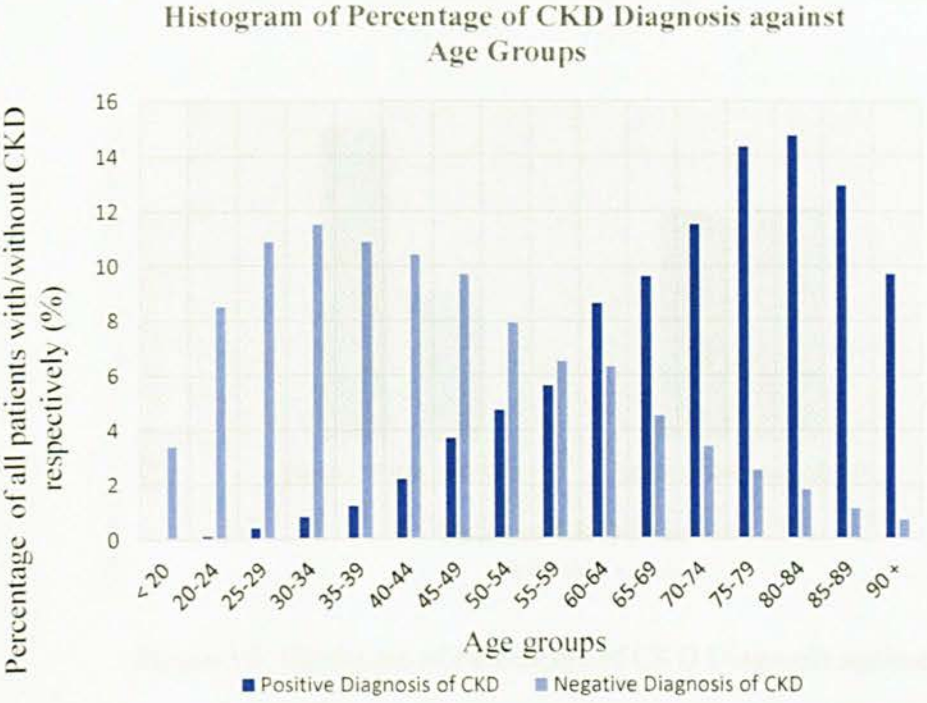

Figure 3.4: Histogram of Percentage of CKD Diagnosis against Age Groups

As can be seen from Figure 3.4, considering the age distributions across the two groups, the one containing the patients diagnosed to have CKD is concentrated in the age range over 55 years and peaking around age 80, whilst the group compromising the patients with no diagnosis of CKD is predominantly distributed across younger age groups, peaking around the age of 30. This means that CKD is a disease predominantly occurring in elderly people.

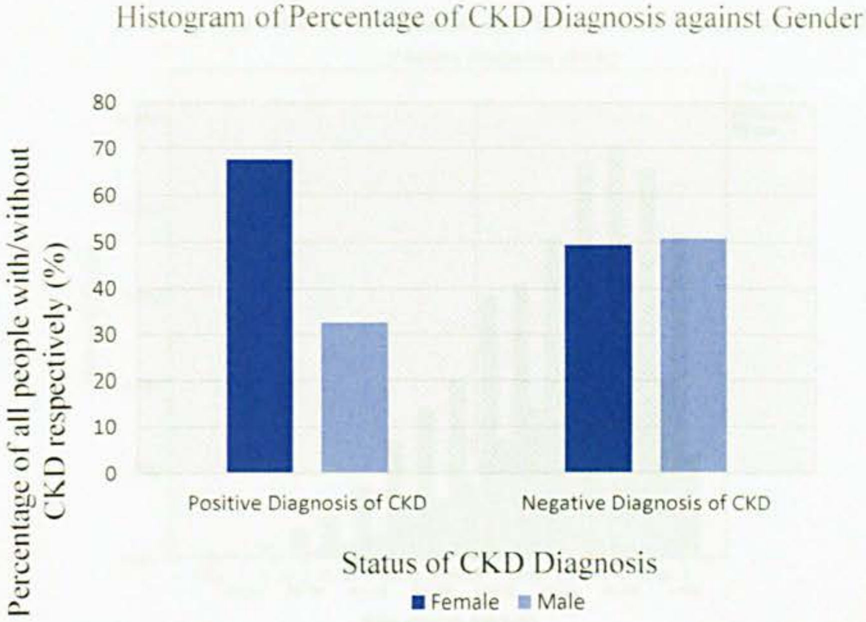

Figure 3.5: Histogram of Percentage of CKD Diagnosis against Gender

The histogram in Figure 3.5 shows that when the two groups (i.e positive diagnosis of CKD and negative diagnosis of CKD) are compared by their gender balance, the group not diagnosed to have CKD had a distribution almost 50% female and 50% male, whereas in the group with positive CKD diagnosis, the percentage of females was approximately 68% and the consequent proportion of males was around 32%. This means that CKD is a disease that is expected to be seen more in females than males.

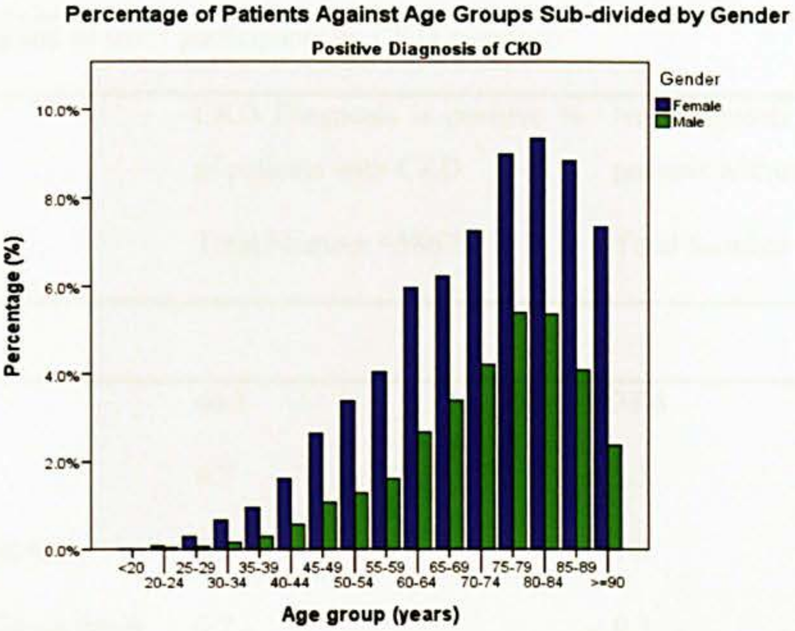

Figure 3.6: Percentage of Patients against Age Groups Sub-divided by Gender

As a result of the comparisons of two groups relating to the positive or negative diagnosis of CKD with respect to age and gender, the distribution of age and gender together can be seen in Figure 3.6 for the patients who have been diagnosed with CKD. From Figure 3.6, it can be concluded that older patients and female patients are more likely to have CKD than are others. Demographic results obtained from our EDA of this GP data set hence indicated similar results to previous studies in this area.

Table 3.6: Ethnic Description of study participants

| Ethnic description of study participants by CKD diagnosis | | |
|---|---|---|
| | CKD Diagnosis is positive % of patients with CKD<br>Total Number =58675 | No Diagnosis of CKD % of patients without CKD<br>Total Number = 818276 |
| Ethnicity | | |
| White | 46.1 | 33.8 |
| Mixed | 0.7 | 1.1 |
| Asian/Asian British | 4.2 | 7.5 |
| Caribbean/African Black | 0.2 | 0.3 |
| Other Black | 0.3 | 0.5 |
| Other | 0.5 | 1.3 |
| Ethnicity not stated | 4.1 | 3.5 |
| Ethnicity information is unknown | 43.8 | 52 |

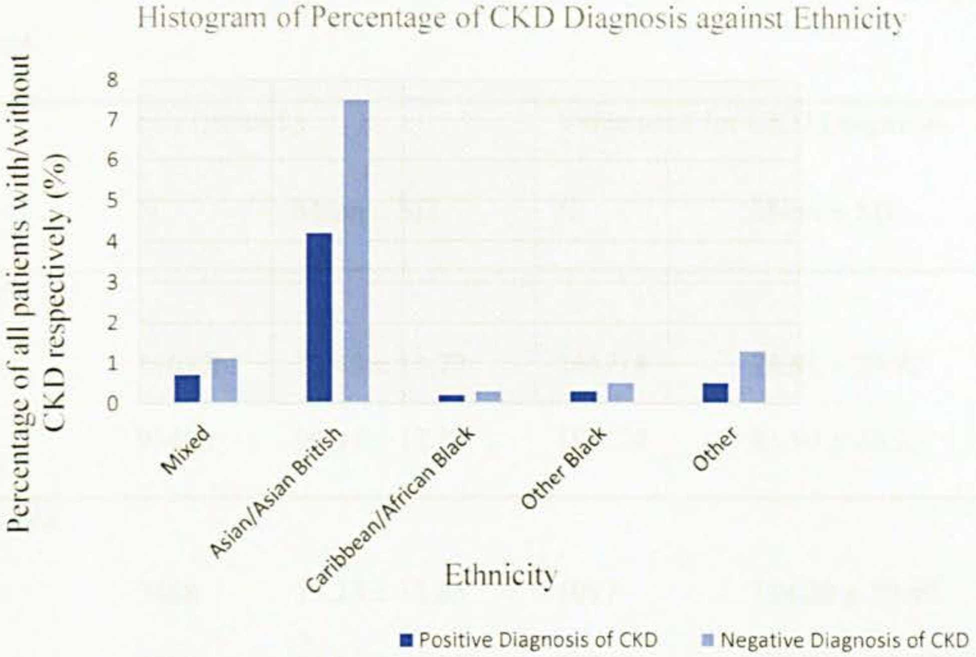

Figure 3.7: Histogram of Percentage of CKD Diagnosis against Ethnicity

White British are excluded from Figure 3.7. Table 3.8 and Figure 3.7 are used to study the ethnic groupings of the patients in this research dataset. In the UK, due to the reasons explained before, ethnicity is not consistently recorded. However, use of summary census-based information rather than a direct measure of ethnicity would introduce bias into the research. In this way, routinely collected data can be used to explain different prevalence rates of major chronic diseases (e.g. CKD) between different ethnic groups, to monitor ethnic differences in disease management and to assist equality in service provision. However, a suitable recording system has first to be established and the benefits of collecting ethnicity data, which is currently poorly recorded in medical records has to be emphasised. It is found that in this GP data set, only around 49% of patients have their ethnicity recorded and of those, 92% have not been diagnosed to have CKD while the other 8% have been diagnosed with CKD.

Values reported from Table 3.8 and histogram from Figure 3.7 shows that when the diagnosis of CKD is considered, the proportions of patients in each ethnic group follow a similar pattern in each category of outcome (i.e. when CKD is present or not). However, a higher proportion of Asian patients are diagnosed with CKD than of other ethnic groups.

Table 3.7: Mean serum creatinine and value used for CKD diagnosis with respect to gender and age groups.

| | SCr (μmol/L) | | Value used for CKD Diagnosis | |
|---|---|---|---|---|
| | N | Mean ± SD | N | Mean ± SD |
| Gender | | | | |
| Female | 116776 | 73.48 ± 11.72 | 248718 | 78.81 ± 20.92 |
| Male | 95400 | 90.44 ± 12.93 | 197124 | 81.90 ± 20.23 |
| Age groups | | | | |
| <20 | 3488 | 73.73 ± 13.85 | 5087 | 114.38 ± 29.40 |
| 20-24 | 9929 | 76.91 ± 14.18 | 16291 | 102.89 ± 23.70 |
| 25-29 | 13031 | 79.00 ± 14.71 | 22740 | 94.71 ± 20.27 |
| 30-34 | 15890 | 79.35 ± 14.95 | 27786 | 90.81 ± 19.08 |
| 35-39 | 18828 | 80.00 ± 14.94 | 32841 | 87.94 ± 18.15 |
| 40-44 | 21974 | 80.93 ± 14.80 | 39667 | 85.12 ± 17.29 |
| 45-49 | 23944 | 81.82 ± 14.77 | 44395 | 82.55 ± 16.84 |
| 50-54 | 21695 | 81.74 ± 14.57 | 42322 | 81.01 ± 16.74 |
| 55-59 | 18887 | 81.95 ± 14.44 | 38983 | 79.36 ± 16.78 |
| 60-64 | 19310 | 82.52 ± 14.21 | 42221 | 76.87 ± 16.42 |
| 65-69 | 14041 | 82.64 ± 14.44 | 34219 | 74.45 ± 16.47 |
| 70-74 | 10814 | 82.23 ± 14.60 | 29432 | 71.69 ± 16.92 |
| 75-79 | 8294 | 82.71 ± 15.28 | 25628 | 67.86 ± 17.33 |

| | | | | |
|---|---|---|---|---|
| 80-84 | 5923 | 82.61 ± 16.23 | 20490 | 64.35 ± 17.69 |
| 85-89 | 3690 | 83.19 ± 17.27 | 14455 | 60.51 ± 17.75 |
| ≥90 | 2438 | 84.74 ± 18.88 | 9285 | 56.53 ± 17.13 |

The value of eGFR used for the CKD diagnosis in this research dataset is reported in Table 3.9 and this value represents either eGFR value directly reported by the patient's GP or eGFR calculated from the SCr using MDRD formula. From Table 3.9, it is clear that the SCr values are significantly different from the eGFR values used for the diagnosis of CKD for each age group and gender and hence, using just the SCr value alone for the CKD diagnosis is not accurate. From the same table, the negative correlation between the SCr and eGFR values can also be seen.

Table 3.8: Frequencies of patients in each co-morbidity group

| | Co-morbidities N (% of total) | | | | |
|---|---|---|---|---|---|
| | IHD | Hypertension | Diabetes | Anaemia | Obesity |
| | 32036<br>(3.7) | 161534<br>(18.4) | 45355<br>(5.2) | 21294<br>(2.4) | 122187<br>(13.9) |
| Age Group (years) | N (% within the age group) | | | | |
| <20 | 4<br>(0.0) | 129<br>(0.3) | 163<br>(0.4) | 467<br>(1.1) | 1636<br>(3.9) |
| 20-24 | 8<br>(0.0) | 676<br>(0.9) | 355<br>(0.5) | 1428<br>(2.0) | 4822<br>(6.7) |
| 25-29 | 30<br>(0.0) | 1684<br>(1.8) | 553<br>(0.6) | 2072<br>(2.2) | 7700<br>(8.2) |

| | | | | | |
|---|---|---|---|---|---|
| 30-34 | 53<br>(0.1) | 2982<br>(3.2) | 865<br>(0.9) | 2369<br>(2.6) | 9197<br>(9.9) |
| 35-39 | 150<br>(0.2) | 5375<br>(6.0) | 1463<br>(1.6) | 2422<br>(2.7) | 11691<br>(13.1) |
| 40-44 | 350<br>(0.4) | 8455<br>(9.8) | 2221<br>(2.6) | 2000<br>(2.3) | 13166<br>(15.3) |
| 45-49 | 776<br>(1.0) | 12025<br>(15.2) | 3209<br>(4.1) | 1532<br>(1.9) | 13848<br>(17.5) |
| 50-54 | 1441<br>(2.2) | 14718<br>(22.9) | 4056<br>(6.3) | 940<br>(1.5) | 12266<br>(19.1) |
| 55-59 | 2142<br>(3.8) | 16699<br>(29.9) | 4658<br>(13.6) | 612<br>(1.1) | 10941<br>(19.6) |
| 60-64 | 3571<br>(6.6) | 19859<br>(36.5) | 5546<br>(10.2) | 682<br>(1.3) | 11036<br>(20.3) |
| 65-69 | 4009<br>(10.0) | 17812<br>(44.3) | 5480<br>(13.6) | 735<br>(1.8) | 8524<br>(21.2) |
| 70-74 | 4725<br>(14.1) | 17386<br>(51.7) | 5497<br>(16.4) | 992<br>(3.0) | 7115<br>(21.2) |
| 75-79 | 5187<br>(18.4) | 16340<br>(57.9) | 4933<br>(17.5) | 1300<br>(4.6) | 5234<br>(18.6) |
| 80-84 | 4481<br>(20.8) | 13143<br>(61.0) | 3560<br>(16.5) | 1394<br>(6.5) | 3084<br>(14.3) |

| | | | | | |
|---|---|---|---|---|---|
| 85-89 | 3310 (21.6) | 9450 (61.6) | 1985 (12.9) | 1352 (8.8) | 1481 (9.7) |
| ≥90 | 1789 (21.0) | 4801 (56.4) | 811 (9.5) | 997 (11.7) | 446 (5.2) |
| Gender | N (% within the gender group) | | | | |
| Female | 12307 (2.78) | 83251 (18.78) | 20594 (4.65) | 17961 (4.05) | 69809 (15.75) |
| Male | 19729 (4.55) | 78283 (18.05) | 24761 (5.71) | 3333 (0.77) | 52378 (12.08) |
| Ethnicity | N (% within the ethnic group) | | | | |
| White | 15832 (5.22) | 68756 (22.66) | 19343 (6.37) | 6939 (2.29) | 53372 (17.59) |
| Mixed | 174 (1.82) | 1279 (13.38) | 547 (5.72) | 315 (3.29) | 1601 (16.74) |
| Asian/Asian British | 2756 (4.32) | 11053 (17.31) | 7448 (11.66) | 3243 (5.08) | 9048 (14.17) |
| African/Caribbean Black | 50 (1.91) | 670 (25.55) | 231 (8.81) | 111 (4.23) | 328 (12.51) |
| Other Black | 56 (1.20) | 681 (14.62) | 229 (4.92) | 216 (4.64) | 833 (17.88) |
| Ethnicity cannot be determined | 13168 (2.67) | 79095 (16.05) | 17557 (3.56) | 10470 (2.12) | 57005 (11.57) |

From Table 3.10, it can be observed that, in this dataset IHD tends to be most common in the older age group, peaking in those aged 85-89 and more common in males than females. It is also more prevalent in Whites than in other ethnic groups.

Similarly, hypertension is also more common in the older age groups and highest amongst the 85-89 category. However, it is slightly more common in females than males. It is more common in African/Caribbeans and Whites compared with other ethnic groups.

However, diabetes follows a rather different pattern from IHD or hypertension. Diabetes is most prevalent in the early elderly age groups, peaking in the 75-79 category, and slightly more common in males than females. It is most prevalent in the Asian ethnic group, but also quite common amongst African/Caribbean blacks.

Anaemia most frequently occurs in the very oldest age groups, and is much more common in females than in males. It is most prevalent amongst the Asian, followed by the Black ethnic groups.

Obesity is most common in the 65-76 age groups and more prevalent in females than males. It tends to be less common amongst African/Caribbean and Asian ethnic groups than the others.

## 3.5 Summary of chapter 3

The data used in this research was collected from 129 GPs over England and Wales, containing about 1.1 million patients where all available information obtained from this collection was taken into account. This dataset was then cleaned and validated to ensure that the dataset is a good representation of the UK, so that any conclusions drawn from this dataset can be applicable to apply to population.

In cleaning and validation steps, the dataset was used as a whole where later, when the interest was investigating patients with CKD, patients diagnosed with CKD at stages 3-5 (6.7% of the whole dataset) were flagged and used. From the results of the descriptive statistics applied on whole dataset, it was found that older patients and female patients are more likely to have CKD than others. It is also observed that high proportion of patients from Asian ethnic background were diagnosed with CKD at stages 3-5. Average person in the dataset is found to have age 46, high HbA1c values, obese, high glucose level, high cholesterol values, normal blood pressure and good level of haemoglobin, serum creatinine, albumin, HDL, triglyceride and MCV. Additionally, in this research dataset, when all patients are taken into account, IHD, hypertension and anaemia are found to be the comorbidities of CKD that are the most prevalent in older age groups, 85-89, 85-89, and ≥90 categories respectively. On the other hand, diabetes and obesity are found to occur mostly in slightly younger age groups 75-79 and 65-74 respectively. Hypertension, anaemia and obesity are found to exist more in females compared to males, whereas for IHD and diabetes the opposite is observed. The highest proportion of patients who have been diagnosed to have IHD are found to be from the white ethnic background but, for the other comorbidities, the highest proportion of patients with hypertension is from the African/Caribbean ethnic background, for both diabetes and anaemia, those from Asian ethnic background, but for obesity those from other Black ethnic backgrounds. It is concluded that eGFR is used to diagnose CKD and is derived from serum creatinine (SCr). Using SCr alone is not a representation of eGFR and negative association is found between eGFR and SCr. Overall, the population distributions of our GP records on key variables relative to the investigation of CKD suggest that the data is a reasonably representative sample of the UK population as a whole. As such any further analyses should provide reliable indicators of the patterns and associations within the clinical progress of CKD.

# 4 Logistic Regression

## 4.1 Logistic Regression for modelling factors contributing to CKD

## 4.2 Logistic Regression

A first attempt at modelling the cleaned data aims to investigate factors which might influence diagnosis of CKD, i.e. those factors that might be correlated with whether a patient will be diagnosed with CKD or not. The aim is to compare individual patient characteristics and co-morbidities (such as diabetes, anaemia, hypertension, ischaemic heart disease, stroke and obesity) between the two groups, CKD and non-CKD, with a view to identifying any significant differences between them. As the outcome of interest is dichotomous, i.e. 'diagnosis of CKD or not' the investigation employs a Logistic regression model.

In logistic regression, the predicted response measures the logit of the probability (p) of an event occurring given that the values of significant independent predictors are known. This means that the discrete outcome of belonging to either group was transformed to probabilities. This transformation was achieved by taking the mean of the outcome at the value of each predictor variable. Hence, cumulative probability distribution was obtained where the new probability was estimated as a result of the addition of a current probability to the total of the previous probabilities.

The main purposes of the logistic regression analyses were to identify significant explanatory variables and to evaluate the effect size of each such variable (i.e. presence of certain characteristics and co-morbidities) on the diagnosis of CKD (or not). The technique models the effects of the explanatory variables on the response variable simultaneously and hence the effect of predictors on other predictors are automatically controlled.

In this chapter, two applications of the logistic regression were performed on the dataset. In the initial application (model 1 section 4.2.1.1.3) the model is applied to the complete cleaned dataset. This application allows identification of which factors affect whether or not a patient is diagnosed with CKD and which are most influential. Our findings are compared against current knowledge in the CKD field.

A second application, (models 2 and 3 section 4.2.1.2.1 and 4.2.1.2.2 respectively) begins to investigate the time and decline aspects of the project and focuses only on patients with a CKD diagnosis (58675 patients). The CKD patients are classified according to whether progression of their disease can be considered as rapid decline or not. Within the CKD community there are two definitions of rapid decline (definition 1; a change of 5 mL/min/1.73m$^2$ over 1 year or; definition 2; a change of 10 mL/min/1.73m$^2$ over 5 years). Here we run two replica logistic regression models where only the definition of the outcome differs. The two different outcomes align with the two definitions of rapid decline as described above. All logistic models presented in this chapter are fitted using SPSS v.21.

### 4.2.1 Applications of logistic regression

#### 4.2.1.1 Initial Application on Whole Dataset – Model 1

The logistic model applied to the data is

$$logit(p) = a + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} \dots + b_nx_{ni} + \varepsilon_i \qquad \text{eq. (4.1)}$$

(Burns and Burns, 2009).

Where;

p is the probability of the outcome of occurrence,

a is the intercept of the best fit line

$b_1$ to $b_n$ are the coefficients where each predictor has its own beta coefficient. B's represent the slope of the equation and are defined as logits which is the log transformation of odds and

$\varepsilon_i$ is the error term.

**Goodness of fit**

The logistic method uses Maximum Likelihood estimation (MLE) to estimate the difference between the logits of the dependent variable (Field, 2013). The MLE approach maximises the capability of predicting the probability of the outcome using the information from predictor variables. When compared to the linear regression, the logistic model requires a higher sample size since maximum likelihood estimation is a technique applied on larger sample sizes; a minimum of 50 observations per predictor variable is reasonable (Field, 2013). Likelihood is a measure of the overall fit of the model; but since likelihood usually results in small numbers, the natural logarithm of the likelihood is used to examine model fit. Log-likelihood is calculated;

$$Log - likelihood = \sum_{i=1}^{N} [Y_i \ln(P(Y_i)) + (1 - Y_i)\ln(1 - P(Y_i))]$$

eq. (4.2)

(Field, 2013).

Where;

Ln represents the natural logarithm,

$Y_i$ is the outcome,

$P(Y_i)$ is the probability of outcome and

i represents the patients from 1 to N where N is the total number.

Log-likelihood indicates the amount of unexplained variation left after the current model is fitted (Burns and Burns, 2009). Due to the probabilities always being less than one, log likelihoods were always negative. The likelihood ratio test is used to assess model fit (Burns and Burns, 2009).

**Model fitting: Comparing model fits**

To compare two logistic models we look to the difference in the Log likelihoods. For example to compare the baseline model (that containing the constant term only) against a model where additional predictors have been added, the change in model fits is evaluated as the difference between the two log-likelihoods multiplied by 2. The differences is multiplied by 2 in order for the resultant value to follow a chi-square distribution (Burns and Burns, 2009; Field, 2013). A chi-square test is then used to measure the significance of the difference. The equation used for this was;

$$\chi^2 = 2[LL(new) - LL(baseline)]$$

eq. (4.3)

$where\ (df = \hbar_{new} - \hbar_{baseline})$

(Field, 2013).

Where;

$\hbar$ represents the degrees of freedom, degrees of freedom shows the number of predictors and the number of predictors at the baseline model which contains only the constant term is taken as 1 and LL represents the log-likelihood ratio.

An R statistic, similar to the R and the R$^2$ values evaluated in linear regression, can be calculated for a logistic regression as.

$$R = \pm\sqrt{\left(\frac{Wald - (2\ \times df)}{-2LL(original)}\right)}$$

eq. (4.4)

(Field, 2013).

Where;

df represents the degrees of freedom of the model,

-2LL represents the log-likelihood ratio of the model and

Wald represents the Wald statistics obtained from the model.

The R-statistic shows the partial association between the response and each of the independent predictors (Field, 2013). The value of the R-statistics can range from -1 to 1. A positive R is interpreted as showing a positive relationship between the outcome and the predictor, so that an increase in the predictor results in an increase in the likelihood of the event occurring. A negative value of R indicates a negative relationship between the response and the predictor where an increase in the predictor results in a decrease in the likelihood of the event occurring. The strength of the association is determined as strong if the absolute value of R was close to 1 and weak if the absolute value of R was close to zero. The strength of the association is evaluated as strong, when the impact of the variable on the overall model is high and vice versa.

As can be seen from the equation above, R-statistics are calculated using the Wald test (Field, 2013) and hence should be applied with caution because Wald statistics can produce imprecise results in certain cases. For this reason, Hosmer and Lemeshow (1989) developed an alternative statistic calculated as:

$$R_L^2 = \frac{-2LL(model)}{-2LL(original)}$$

eq. (4.5)

(Hosmer and Lemeshow, 1989; Field, 2013).

The formula (eq. 4.5) shows how much development was made on the model fit when a predictor was added. Since it is a squared term, the values of this statistic range from 0 to 1 and hence negative values of $R^2$ are prohibited. This statistic represents the reduction in the absolute value of the log-likelihood of the current model relative to the absolute value of the log-likelihood of the baseline model. In the same year, Cox and Snell reported an alternative $R^2$ measure formula as:

$$R_{CS}^2 = 1 - e^{\left[-\frac{2}{n}(LL(new))-(LL(baseline))\right]}$$

eq. (4.6)

(Cox and Snell, 1989; Field, 2013).

While this statistic is also based on the log-likelihood (eq. 4.6), it differs from the Hosmer and Lemeshow $R^2$ in that it takes into account the sample size, n. However as the natural logarithm can never result in zero, this statistics never reaches to the value of 1 which is the maximum limit of the R$^2$ statistics (Field, 2013). Hence, in 1991, Nagelkerke improved this statistics and indicated a modified version of the Cox and Snell's statistics as;

$$R_N^2 = \frac{R_{CS}^2}{1 - e^{\left[\frac{2(LL(baseline))}{n}\right]}}$$

eq. (4.7)

(Nagelkerke, 1991; Field, 2013).

The advantage of this statistic (eq. 4.7) is that it takes into account both the log-likelihood ratios and the sample size and in addition it can reach the maximum limit of 1.

##### 4.2.1.1.1 Assessing and interpreting individual parameters: Wald Statistics and Odds ratios

Wald statistics are used in logistic regression to test the significance of each predictor variable. The Wald statistic follows a Chi-square distribution and is calculated with the equation;

$$Wald = \frac{b}{SE_b}$$

eq. (4.8)

(Field, 2013).

Where;

b is the coefficient for each predictor and

SE is the corresponding standard error for that predictor.

Wald statistics, though, might produce biased results especially if the standard errors are inflated. Such a case happens when the coefficient of the predictor is large, so that the standard error is larger and hence the Wald statistics results in very small number. Hence, the Wald statistics is underestimated and shows the predictor as very significant, as this significance is caused by the inflation of the standard errors due to the high beta coefficients (Menard, 1995). Therefore the result is imprecise. Consequently, it can be stated that it is better to use log-likelihood ratio statistics to measure the significance of the predictor rather than the Wald statistics, and the researcher minimises the likelihood of making a type II error (i.e. probability of failure to reject the null hypothesis when it is actually false).

#### 4.2.1.1.2 Odds ratio: Exp(B)

The odds ratio illustrates the change the odds of a particular outcome due to a unit change in the predictor (Burns and Burns, 2009). Odds are used to describe the probability of an event occurring relative to the probability of the same event not occurring (Field, 2013). Odds is calculated as:

$$odds = \frac{P(event\ Y)}{P(no\ event\ Y)}$$

eq. (4.9)

Where;

$$P(event\ Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1)}}$$

eq. (4.10)

and

$$P(no\ event\ Y) = 1 - P(event\ Y)$$

eq. (4.11)

(Field, 2013).

The odds ratio can be explained as the change in odds. If the result of the odds ratio is greater than 1, this means that as a unit increase in the predictor variable was observed and hence the odds of the event Y occurring was also increased. On the other hand, if the odds ratio resulted in a value smaller than 1, this means that a unit change in the predictor variable causes the odds of the even Y occurring to decrease (Humphrey, 2009).

### Assumptions of Logistic regression and stepwise methods

The assumptions of the logistic regression model include; the requirement of categories of outcome variable as mutually exclusive and as exhaustive (Burns and Burns, 2009), meaning that every patient in the research dataset should belong to one category of outcome and can only be in one group at a time. Hence, errors are assumed to be independent. The model does not assume a linear relationship between the dependent and the independent variables and the independent variables do not have to be normally distributed, do not need to be interval and do not need to have equal variance within the group (Burns and Burns, 2009).

Similar to the linear regression, the logistic regression also assumes that there is no multicollinearity between the independent predictors.

The results of a logistic regression model can be used to predict probabilities and hence can take any value within the range of 0 to 1 rather than being restricted to take just a value of 0 or 1. The outcome obtained from the model indicates the probability of a subject belonging to either of the two groups. In order to normalise the probability distribution, further transformation was needed and the probabilities were transformed to logits by performing the log transformation on the probabilities. Transforming from probabilities to logits, removed the limitation on the predicted outcome restricted to lie between 0 and 1. Therefore, logits can take any value from negative infinity to positive infinity.

**Stepwise procedures:**

When performing logistic regression in SPSS, there are various options for finding the most parsimonious model. The forced entry method is the default unless an alternative method is selected.

In using stepwise methods, the purpose is to investigate whether or not the addition of each predictor improves the model. While both forward stepwise and backward elimination methods can be used, and in this thesis, same results were obtained by using both methods, we have opted for a backward elimination method in the following analyses. The procedure starts with a model containing all the predictor variables (Field, 2013) and then removes a non-significant predictor each step, starting with the variable with the highest p-value, i.e. the predictor which has the smallest effect on outcome.

The backward elimination method is often chosen over the forward methods for this type of investigation. The reason for this is, in forward stepwise methods, predictors having a suppressor effect are removed and hence this increases the possibility of a type II error. Suppressor effect occurs when a predictor is only significant under the condition of keeping the other predictors constant.

#### 4.2.1.1.3 Results of Logistic Regression (Model 1)

Initially a multiple logistic regression model was applied to the cleaned dataset, comprising 876951 patients from 129 General Practices in England and Wales. The outcome was whether or not a patient has been diagnosed with CKD, where a patient who has been diagnosed with CKD was coded as 1 and those who have no diagnosis of CKD are coded as 0. In the dataset 58675 (6.7%) patients were classified as having been diagnosed with CKD (stages 3-5) and the remaining 818276 (93.3%) patients had not (been diagnosed with CKD). The independent predictors considered were diagnoses of different co-morbidities namely; anaemia, diabetes, ischaemic heart disease, stroke, hypertension and obesity. All of these independent variables are also binary coded; 1 meaning that the patient has been diagnosed with that disease and 0 indicating they have not.

All of the predictors namely; anaemia diagnosis, diabetes diagnosis, hypertension diagnosis, IHD diagnosis, stroke diagnosis and obesity diagnosis were added to the logistic regression analysis and a backwards stepwise procedure was applied using the log-likelihood ratio statistics as the criterion for removing the variable from the model. The stepwise procedure produced a model of best fit that included all but except one of the six predictor variables included initially (Table 4.3). The results indicate that obesity (measured as yes/no) is not statistically significant in predicting whether or not a patient has been diagnosed with CKD. Hence obesity is removed and the model refitted. The results of this analysis (Table 4.3) show all remaining predictor variables to be significant, i.e. they have coefficients significantly different from zero. The results from the initial model are presented in Table 4.3.

Table 4.1: Results of logistic regression model for diagnosis of CKD

| | | | | | | | 95% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| **Predictor** | **Beta** | **SE** | **Wald** | **Df** | **Sig.** | **Exp(B)** | **Lower** | **Upper** |
| Anaemia | 1.037 | 0.020 | 2666.810 | 1 | 0.000 | 2.820 | 2.711 | 2.933 |
| Diabetes | 0.480 | 0.014 | 1176.259 | 1 | 0.000 | 1.1616 | 1.573 | 1.661 |
| Hypertension | 1.672 | 0.010 | 30424.066 | 1 | 0.000 | 5.323 | 5.223 | 5.423 |
| IHD | 1.170 | 0.014 | 6557.963 | 1 | 0.000 | 3.222 | 3.132 | 3.3314 |
| Stroke | 1.025 | 0.018 | 3279.157 | 1 | 0.000 | 2.787 | 2.691 | 2.887 |
| Constant | -3.433 | 0.007 | 267241.034 | 1 | 0.000 | 0.032 | | |

The equation of final logistic regression model was;

$$\begin{aligned} &Logit(CKD\ Diagnosis\ (Y|N)) \\ &\quad = -3.433 + 1.037(Anaemia\ Diagnosis) + 0.480(Diabetes\ Diagnosis) \\ &\quad + 1.672(Hypertension\ Diagnosis) + 1.170(IHD\ Diagnosis) \\ &\quad + 1.025(Stroke\ Diagnosis) \end{aligned} \qquad \text{eq. (4.12)}$$

The estimates of the β's (Table 4.1) are used to define the logistic regression equation from which we can compute the probability that a patient was diagnosed as having CKD (or not) given a particular combination of co-morbidities. The β values correspond to the change in the logit of the outcome variable when a unit change is observed in the predictor variable. The logit of the is the natural logarithm of the odds of outcome Y occurring; in this case Y is a patient having had a diagnosis of CKD

The odds ratio (EXP(B)) is the more commonly used interpretation of the parameter estimates, β.

From Table 4.1, it can be concluded that the odds of a patient with hypertension being diagnosed with CKD are 5.323 times higher than for a patient who does not have hypertension. The results suggest that hypertension is the most influential predictor for diagnosis of CKD. The second highest predictor affecting the diagnosis of CKD is the presence of IHD where the odds of a patient who was diagnosed with IHD as being diagnosed with CKD was 3.222 times greater than those of a patient who does not have IHD. After IHD, factors affecting the outcome are anaemia, stroke and diabetes respectively with odds ratios of 2.820, 2.787 and 1.616.

Table 4.2: Model Summary of the Final Logistic Regression Model

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerle R Square |
|---|---|---|---|
| 1 | 366684.435 | 0.070 | 0.181 |
| 2 | 366685.210 | 0.070 | 0.181 |

The overall fit of the final logistic regression model was assessed using the -2 Log likelihood statistics (Table 4.2). Since step 2 was the final step of the stepwise procedure, we can compare fit between our best model (step 2) against the baseline model (containing constant term only) which had -2LL = 430699.293. The decrease in -2LL between models suggests that the addition of predictors have improved the model and the significance of this improvement is tested by comparing the change in -2LL = 64014.083 against $\chi^2_{0.05,4} = 9.488$. It can be concluded that the model including all 5 predictors is a significantly better fit to the data than the baseline model (no predictors).

Both Cox & Snell R Square and Nagelkerke R Square measures from Table 4.2 were used as an effect size measures of the model, showing how effective the independent predictors are on determining the outcome of the model. Using such information from Table 4.2, concluded that 18.1% of the total variation in the outcome can be explained by knowing the information about these predictors.

The results from the final model of binary logistic regression predict that 4960 patients have CKD. This prediction was found to be correct for 2532 patients, therefore 2428 patients were misclassified. The model also predicted that 871991 patients do not have CKD but this prediction was only correct for 815848 patients, therefore 56143 patients in this category were misclassified. This means that for the diagnosis a patient, who has CKD, 4.3% of the cases were correctly classified and for the diagnosis of a patient who does not have CKD, 99.7% of the cases were correctly classified. In this application the outcome modelled is whether the patient has been diagnosed to have CKD or not. The results show only 18.1% of variation explained by the model, a low sensitivity suggesting that the model as constructed is not useful for predicting whether a patient has a diagnosis of CKD. Hence we cannot determine from this model that the co-morbidities considered are useful predictors of the prevalence of CKD in individuals. As progression of CKD is of more interest than prevalence in this research, a different approach is taken for a second logistic regression application in order to look in more detail on how such potential predictors affect the progression of CKD.

#### 4.2.1.2 Logistic Regression Application Two: Rapid decline in CKD status

As one of the main aims of this research is to examine the changes in CKD status over time, a further application of the logistic regression was used in order to develop some understanding of the patterns of the progression of CKD by considering the progression of CKD using the two definitions of rapid decline. Rapid decline and rate of decline are of great interest to the clinical community and according to the CG73 Chronic Kidney Disease NICE Guidelines, 2008, there are two standard definitions of 'rapid decline' of kidney function when the progression of chronic kidney disease is considered. The first (definition 1) is defined as a decline in eGFR of more than 5 mL/min/1.73m$^2$ within 1 year. The second (definition 2) indicates rapid decline as a decrease of more than 10 mL/min/1.73m$^2$ in eGFR over a 5 year time period.

Here logistic regression is used to investigate progression of the CKD measured as rapid decline, and the impact of co-morbidities on this progression. Two separate but identical models are used to investigate the rapid decline corresponding to the two definitions described above.

The CKD population is also refined according to findings from model 1. As obesity was found not to be an influencing factor in diagnosis of CKD and it is known that the MDRD formula is not valid for obese people (having BMI values above 30), we have removed patients with BMI> 30 from our analysis sample. This was done in order to improve validity of our model estimates by including only those with reliable CKD diagnosis. In doing so our full analysis sample was decreased to 754764 patients. From the remaining sample, in order to identify rapid decline, patients must have at least two repeated eGFR measurements. Hence a further, 507825 patients who do not have any repeated eGFR measurements (including not even a single measurement) are excluded leaving 246939 patients who have repeated eGFR measurements. In this group patients can have between two to fifteen eGFR measurements. The next step was to identify patients who have experienced rapid decline at any time point within the sub-group of patients who have been diagnosed with CKD, only the patients who have been diagnosed with CKD are taken into consideration (23478 patients from the total 110958 patients, 21.2% of 110958). Rapid decline was defined for those patients according to two different definitions and two different logistic regression models were generated.

In order to define a rapid decline, the change in eGFR was calculated from the first available measurement (at first time point). As the time of first eGFR reading was different for each patient, the date for the first measurement (for each individual) is denoted as time zero and subsequent readings are used to calculate the following time points by subtracting the date of the baseline measurement from the date of the next measurement. Hence a time related variable was created where the value of the first measurement of time was zero and the value for the following measurements of time was increasing cumulatively.

On completion of this process, 1514 (6.4% of 23478) patients were found as having a rapid decline based on definition 1 (i.e. more than 5mL/min/1.74m$^2$ decrease of eGFR within 1 year) and 3306 (14.1% of 23478) patients are found as having a rapid decline based on definition 2 (i.e. more than 10mL/min/1.74m$^2$ decrease of eGFR within 5 years). It was notable that the number of patients with rapid decline over 5 years (definition 2) was more than the double of number of patients who showed a rapid decline of kidney function within a year (definition 1). This shows that rapid decline over 5 years is more common than rapid decline over a year. When a patient suffers from a rapid decline of kidney function, presence of CKD for that patient is diagnosed and hence, suitable therapy is started after the awareness of diagnosis of CKD. Slower deterioration of kidney function could be obtained as a result of awareness and therapy after diagnosis of CKD. The results obtained here shows that there are less amount of patients who are suffering from rapid decline over a year than the number of patients suffering a rapid decline over 5 years and this suggest that kidney function continue to have a rapid decline for majority of patients which means that more patients are diagnosed with CKD at later stages such as stages 3-5.

#### 4.2.1.2.1 Results of Logistic Regression Model 2

In this section, the model fitted into the data (model 2) aims to model rapid decline based on definition 1 (i.e. more than 5mL/min/1.74m$^2$ decrease of eGFR within 1 year). The co-morbidities namely, anaemia, diabetes, hypertension, IHD and stroke diagnosis were added to the model initially.

Table 4.3: Results of final logistic regression model for rapid decline based on definition 1

| | | | | | | | 95% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| **Predictor** | **Beta** | **SE** | **Wald** | **Df** | **Sig.** | **Exp(B)** | **Lower** | **Upper** |
| Anaemia Diagnosis | 0.420 | 0.084 | 25.248 | 1 | 0.000 | 1.522 | 1.292 | 1.793 |
| Diabetes Diagnosis | 0.284 | 0.066 | 18.820 | 1 | 0.000 | 1.329 | 1.169 | 1.511 |
| Constant | -2.770 | 0.031 | 7959.683 | 1 | 0.000 | 0.063 | | |

The results in Table 4.3 suggested that only anaemia and diabetes were significant predictors of rapid decline as defined by definition 1 (>5mL/min/1.74m$^2$ decrease within 1 year). The equation of the final logistic regression model for the rapid decline based on definition 1 was found as being;

$$
\begin{aligned}
&Logit(Rapid\ Decline\ (Y|N)) \\
&\qquad = -2.770 + 0.420(Anaemia\ Diagnosis) + 0.284(Diabetes\ Diagnosis)
\end{aligned}
\tag{eq. (4.13)}
$$

The results (Table 4.3) further suggest that a previous diagnosis of anaemia was the most influential factor in rapid decline as a patient with anaemia is 1.522 times more likely to experience a rapid decline of kidney function (definition 1) than a patient who does not have anaemia. Similarly, the odds of a diabetic patient experiencing rapid decline of kidney function (definition 1) is 1.329 times more than a patient who does not have diabetes.

#### 4.2.1.2.2 Results of Logistic Regression Model 3

The same model was set up to examine the impact of co-morbidities on rapid decline, defined as more than 10mL/min/1.74m$^2$ decrease of eGFR within 5 years (definition 2).

Table 4.4: Results of final logistic regression model for rapid decline based on definition 2

| | | | | | | | 95% CI for Exp(B) | |
|---|---|---|---|---|---|---|---|---|
| **Predictor** | **Beta** | **SE** | **Wald** | **Df** | **Sig.** | **Exp(B)** | **Lower** | **Upper** |
| Anaemia Diagnosis | 0.618 | 0.058 | 112.951 | 1 | 0.000 | 1.854 | 1.655 | 2.078 |
| Diabetes Diagnosis | 0.560 | 0.045 | 115.110 | 1 | 0.000 | 1.751 | 1.603 | 1.912 |
| Hypertension Diagnosis | 0.308 | 0.042 | 53.835 | 1 | 0.000 | 1.360 | 1.253 | 1.477 |
| IHD Diagnosis | 0.239 | 0.044 | 29.851 | 1 | 0.000 | 1.271 | 1.166 | 1.384 |
| Stroke Diagnosis | 0.215 | 0.052 | 17.263 | 1 | 0.000 | 1.240 | 1.120 | 1.372 |
| Constant | -2.285 | 0.038 | 3618.972 | 1 | 0.000 | 0.102 | | |

The results of this model (Table 4.4) indicated that all of the co-morbidities considered were significantly associated with the rapid decline of CKD (according to definition 2).

The equation of final logistic regression model for rapid decline based on definition 2 (Model 3) was found to be;

$$\begin{aligned} &Logit(Rapid\ Decline\ (Y|N)) \\ &\quad = -2.285 + 0.618(Anaemia\ Diagnosis) + 0.560(Diabetes\ Diagnosis) \\ &\quad + 0.308(Hypertension\ Diagnosis) + 0.239(IHD\ Diagnosis) \\ &\quad + 0.215(Stroke\ Diagnosis) \end{aligned}$$

eq. (4.14)

Again the presence of anaemia was the most influential factor in rapid decline (definition 2), hence a patient who has anaemia, was 1.854 times more likely to exhibit a rapid decline in kidney function than a patient who does not have anaemia. The values reported in Table 4.4 also suggest that diabetes was a significant factor in rapid decline, with diabetic patients being 1.751 times more likely to show rapid decline than non-diabetic CKD patients.

When the figures of odds ratios were compared, both anaemia diagnosis and diabetes diagnosis influenced the outcome even more when the rapid decline was defined according to definition 2, rather than identifying the patients suffering from rapid decline based on definition 1. According to definition 1, other factors such as hypertension, IHD and stroke have not provided any additional benefit to the diagnosis of rapid decline whereas with definition 2, significant contributions are achieved from each predictor. This means that, an initial faster decline of kidney function occurring for a short period of time was affected from anaemia and diabetes whereas a slower, gradual decline of kidney function over longer period of time was affected from other factors such as hypertension, IHD and stroke as well as anaemia and diabetes.

In summary, when using logistic regression to model diagnosis of CKD (model 1), the predictor having the highest effect on the outcome was the diagnosis of hypertension with an odds ratio of 5.323, followed by diagnosis of IHD with an odds ratio of 3.222, after that diagnosis of stroke with an odds ratio of 2.787, then diagnosis of anaemia with an odds ratio of 2.820 and lastly the diagnosis of diabetes with an odds ratio of 2.787. However the best model did not provided a very good fit.

The logistic model was much more useful in the second application where it was used to model rapid decline in CKD status, or in other words progression of CKD. In this application it was found that a co-diagnosis of anaemia was the most influential factor in rapid decline whether using definition 1 or 2. Further we found that diabetes also had a significant effect on rapid decline according to definition 1 but that the other comorbidities were not significant. However when considering an alternative definition of rapid decline (definition 2) we found that all co-morbidities considered had an influential impact, but again anaemia was the most significant with an odds ratio of 1.854 (followed by diabetes, OR= 1.751; hypertension, OR= 1.360; IHD, OR =1.271 and Stroke, OR= 1.240).

In conclusion while all of the co-morbidity predictors were significantly important for both types of outcome, the primary factors affecting the prevalence of CKD were different to the primary factors affecting the progression of CKD. While hypertension and IHD were found as the predominant factors affecting the diagnosis of CKD, anaemia and diabetes were found as the predominant factors affecting the progression of CKD when a patient was categorised as having rapid decline of kidney function based on definition 1 or definition 2.

# 5 Parametric Modelling

## 5.1 Introduction

The results obtained thus far indicate that the co-morbidities which have the greatest effect on the prevalence of CKD (identified by diagnosis of CKD) are different from the co-morbidities which have the greatest effect on rapid decline of a patient's CKD status. This suggests that the presence and impact of different combinations of co-morbidities may affect progression of CKD in different ways. Therefore, in order to fully understand variation in the progression of CKD, there is a need to investigate which individual factors (primarily co-morbidities) influence progression of CKD, and how.

To investigate changes in CKD status we want to examine how the dependent variable (i.e. a patient's eGFR value) changes over time (measured in years). In this chapter, linear mixed models, generalized linear mixed models and polynomial mixed models are applied to examine variations in the progression of CKD over time. Two different types of methodologies, a data-driven approach and a theory-driven approach can be used in modelling the type of change being investigated here and both are discussed and compared below.

## 5.2 Data-driven approach or Theory-driven approach?

A data-driven approach is used when the pattern of change in the dependent variable is unknown and experimental data is available to identify and describe this pattern. On the other hand, a theory-driven approach is the preferred technique when the expected pattern in the change of response is known and the interest lies in building a model to describe this pattern (Tuma, 2013).

In this research, routinely collected general practice data provides empirical data about CKD and its progression due to having repeated measurements of eGFR (a biomarker for the diagnosis of CKD) for each patient. While kidney function (in both CKD patients and non-CKD patients) is expected to decline over time, faster decline in kidney function is expected in patients diagnosed with CKD. The definite pattern of this decline is not known but is believed to differ between patients, due to other factors such as the presence of different combinations of co-morbidities (de Lusignan *et al.*, 2005). Therefore, a data-driven approach is used here to examine patterns of change in patients with CKD and to associate patterns of change in this phenomenon with other patterns that are known to be related to this change.

Applying a data-driven approach involves using existing methodologies suited to the main features of the empirical data to describe the pattern of change in the dependent variable. The availability of repeated eGFR values for each patient (between 2 to 15 per patient) provide a series of repeated measurements, which are considered as the 'outcome', and change in this outcome is analysed over time. Hence, the change modelled here is the progression of CKD over time. A time variable based on a continuous scale, with unequal time intervals between measurements and unequal numbers of time measurements for different patients, is created from the repeated eGFR readings. The first measurement for each patient is classified as time zero ($t_0$=0) and the subsequent measurements form a cumulative set of measurements.

Existing models of change are usually variable-based (Tuma, 2013). In this research, this means that the "units of analysis" are people and the characteristics of each person are measured using variables such as age, gender, BMI, eGFR. The aim of the research being to describe, explain and predict the change (in CKD measured by eGFR) over time according to the characteristics of the units of analysis, in this case CKD patients.

The models used here aim to explain patterns of change in the responses as a function of individual characteristics (age, gender, and ethnicity) and changes in co-morbidities (i.e. differences in co-morbidities between patients). This means that such co-morbidities can affect changes in CKD but does not imply that they are necessarily the causes of variation occurring in the responses. In this dataset, in using these models as constructed it is assumed that the co-morbidities considered in the model were diagnosed before the first diagnosis of CKD and the results obtained should be interpreted bearing this in mind. This assumption is made because since the recording of CKD is added on P4P reward scheme in 2006, the recording of the diagnosis of other co-morbidities by general practices were older than the recording of the diagnosis of CKD. The models used in this chapter are carried out on sub sample of 472 patients who are diagnosed with CKD from the first eGFR measurement. Since the recording of diagnosis of CKD in general practice records occurred after the recording of the diagnosis of other co-morbidities, according to our assumption, if the patients are selected such that they have been diagnosed to have CKD from the first eGFR measurement, the diagnosis of CKD will then be assumed to be after the diagnosis of other co-morbities for all patients. We believe that in order to observe the best picture for the pattern of change, at least 8 repeated eGFR measurements are needed. Even the models described in this chapter allow the use of missing observations, just to make the modelling procedure faster and to be able to compare with the models described in later chapter such as in chapter 6 where such models does not allow the use of missing observations, sub sample of 472 patients are used in the models described in both chapter 5 and chapter 6 where those patients have full 8 repeated eGFR measurements and have been diagnosed to have CKD from the first observation point.Therefore the patterns suggested using this approach allow the researcher to predict future outcomes for the value of eGFR at the next observation time point (e.g. the value of a patient's eGFR after another 5 years) and hence enable prediction of that patient's stage of CKD in the future.

## 5.3 Models for Analysis of Repeated Measurements

In this section, classical approaches for the analysis of repeated measures are discussed. Repeated measurements can be analysed using one of two broad categories of methodologies, namely univariate and multivariate approaches (see Figure 2.2). Here, such methods and their suitability for the data being investigated are discussed, providing a rationale for the models subsequently used for the analysis of our data.

### 5.3.1 Univariate Approaches

Univariate methods (Diggle *et al.*, 1996; Twisk, 2003) can only be used if the repeated measurements from each patient are reduced to a single measurement by reducing the vector of repeated measurements (Davis, 2002). In other words, these methods are applicable where summary statistics data (e.g. mean, mode, median) can be used as a single measurement of response. Univariate approaches such as summary statistics and ordinary least square (OLS) methods are then employed as analysis techniques (Wishart, 1938; Pocock, 1983; Frison and Pocock, 1992; Dawson, 1994; Davis, 2002).

The benefits of such approaches lie in the simplicity and ease of use of the methods in applications, but they are inappropriate in situations where the number of repeated measurements from each patient is different, resulting in incomplete data. When the number and the temporal patterns of repeated measures differ for each patient, univariate approaches can lead to inaccurate results when making comparisons between patients (Davis, 2002). In such cases, to counteract this problem, weighted analysis of summary statistics can be used as an alternative to summary statistics to account the fact that some data points contribute more than others (such as when calculating the arithmetic mean of a measure of two classes when two classes contains different number of data points) (Matthews, 1993). The main disadvantage of the univariate approach is the loss of information in summarising results and the potential for missing information if the chosen summary statistics do not truly represent the actual data (Davis, 2002). Hence, univariate approaches are not applied in this study.

### 5.3.2 Multivariate Approaches

The most commonly used multivariate approaches (Twisk, 2003) comprise four different techniques namely; unstructured multivariate approaches, multivariate analysis of variance (MANOVA) techniques, repeated measures ANOVA and linear mixed models. The advantage of multivariate approaches over univariate approaches lies in that all of the data points are utilised rather than just a single summary measure. In all of the multivariate techniques described below, the dependent variable is assumed to be normally distributed (Davis, 2002).

#### 5.3.2.1 Unstructured Multivariate Approaches

When interest lies in testing whether or not the repeated measurements are different from each other within individual patients in a single sample, Hotelling's T-statistic, a generalization of one sample t-test, can be used (Hotelling, 1931; Albert, 1999). The dependent variable is assumed to follow a multivariate normal distribution and there should be fewer independent predictors than the number of patients. However, this test is not suitable for use where there is missing data and so is not applicable in this project (Davis, 2002).

### 5.3.2.2 Multivariate Analysis of Variance Approaches

Multivariate analysis of variance methods, namely multivariate general linear model (based on MANOVA) and profile analysis are employed when there are two groups of patients (Twisk, 2003). In MANOVA, since the data contains multiple responses from the measurements at different times from each subject, the within (WSS), between (BSS) and total sum of squares (TSS) from the standard univariate ANOVA are replaced by three sum of squares matrices. The diagonal elements (also called "sums of squares") of the covariance matrix represent the variance of each variable and the off-diagonal elements represent the covariance between pairs of variables (also called cross-products).

Profile analysis is used when the aims are to test if the profiles of different groups of patients within the data are parallel; i.e. testing if there is a difference in time profiles between groups and if there is a difference between measurements at different time points within a group. Profile analysis (using plots of the data to visually compare between groups) can be used in a very broad context so long as there is a multivariate outcome for each patient. However, profile analysis does not consider the natural order of the repeated measurements over time. However, as the data set used in this study is based on a single sample, these methods which require two samples (unstructured, MANOVA and profile analysis) are not appropriate (Davis, 2002).

Growth curve analysis (Potthoff and Roy, 1964) can be used with both single and multiple sample data. This is another multivariate approach and it is used when the numbers of repeated measurements within patients are large (Rao, 1965, 1966, 1967; Khatri, 1966; Davis, 2002). However, growth curve analysis assumes that the time points at which the repeated measurements are made are equally spaced (Davis, 2002) and since that is not the case for our data, again these types of models are not applicable here.

### 5.3.2.3 Repeated Measures ANOVA (ANCOVA) Approaches

Repeated measures ANOVA (or ANCOVA) is another multivariate analysis technique which can be applied to both single and multiple sample datasets (Twisk, 2003). The difference between this method and the unstructured multivariate and MANOVA techniques is that repeated measures ANOVA makes an assumption on the covariance structure of the vector of repeated measurements. The model assumes that repeated measurements are continuous and normally distributed for each patient and that;

- All random components including the random effects for each patient and random error component are independent.
- Repeated measurements for any one patient are correlated.
- Sphericity: the covariance matrix has all diagonal elements equal and all off-diagonal elements equal, which means having a compound symmetry covariance structure (Huynh and Feldt, 1970).
- The correlation between any pair of repeated measurements for any patient is same.

(Hedeker and Gibbons, 2006).

In the dataset used here, the repeated eGFR measurements are not equally spaced in time and correlations between any pair of repeated measurements are not necessarily the same. Hence the assumptions of repeated measures ANOVA are not fully met. Furthermore, in our patient record data, repeated eGFR measurements taken closer together in time tend to be more strongly correlated than those taken further apart in time and so careful consideration is needed before we may assume a compound symmetry covariance matrix. Due to strict assumptions such as the assumption of equally spaced repeated measurements and the assumption on the correlation between the measurements, the repeated measures ANOVA method is not an appropriate technique for use in this research.

In summary, the classical multivariate approaches described above (i.e. unstructured multivariate, MANOVA, ANCOVA) are not applicable here for the reasons described above. The unsuitability of these traditional approaches is mainly due to variable numbers of repeated eGFR measurements amongst patients and so, in effect, there is missing data within the repeated measures, i.e. the data can be considered incomplete. Furthermore, the repeated measurements within patients are not equally spaced in time, leading to the data being considered as unbalanced (Fitzmaurice *et al.*, 2004). Therefore, alternative modelling approaches must be considered.

#### 5.3.2.4 Linear Mixed Models (LMM)

Typically, data from longitudinal studies is unbalanced and incomplete. Linear mixed models have been developed from ANOVA models by including extra random effect terms. Addition of these random effect terms allow certain regression coefficients to vary randomly between individuals, resulting in statistical models that have an ability to handle unbalanced and incomplete measurements in a natural way. These methods are therefore suitable for modelling heterogeneity both within and between individuals in complex, longitudinal outcomes containing multiple sources of variation (Laird and Ware, 1982; Singer and Willlett, 2003). Hence, Linear Mixed Models (LMMs) (Ware, 1985; McLean *et al.*, 1991) provide an alternative to modern methods based on classical multivariate approaches and can deal with the challenges typical of longitudinal data, for example when missing repeated measurements are located at the end of the measurement period for each patient rather than located in the middle, between two actual repeated measurements (Diggle *et al.*, 1994; Longford, 1994).

The "mixed" concept of such models combines the need to consider both *within* patient differences and *between* patient differences simultaneously. Instead of applying additional constraints to meet the assumptions of the model, such models allow the patient level information to be used in generating the model. In simple terms, a linear mixed model analysis can be considered as a univariate or multivariate regression analysis of responses with correlated errors (Bates, 2012).

The general form of a LM can be expressed as;

$$y = X\beta + \varepsilon$$

eq. (5.1)

Where;

$y$ is the vector of responses in of n × 1 dimensions consisting independent observations,
$\beta$ is the vector of unknown parameters to be determined in dimensions of p × 1, namely the coefficients of the predictors.
$X$ is the model matrix in of n × p dimensions where $x_{ij}$ represents predictor j for individual i and

$\varepsilon$ is the vector of independent errors with mean zero and constant variance.

(Davis, 2002).

In linear mixed models, the unknown β parameters are evaluated using the ordinary least squares method. The main purpose of using this method is to model the average of the outcome of interest for a given patient. If the assumption of independent errors (ε) employed in LMMs is changed, so that the errors are assumed to have mean zero but differing variances per patient, this enables modelling of the outcome using a weighted least squares (WLS) approach, that is the approach of weighting observations to estimate variance based on known positive constants to counteract the problem of non-constant variance, rather than ordinary least squares. On the other hand, if the errors are assumed to have zero mean and have a non-trivial covariance matrix, then the generalized least squares (GLS) approach which is the generalization of WLS is used to estimate the β coefficients. GLS is used if in addition to unequal variances where WLS can be used, correlation exists between observed variances. LMM is a "mixed" model containing fixed and random effects (Bickel, 2007). Therefore, the model equation becomes

$$y = X\beta + Z\gamma + \varepsilon$$

eq. (5.2)

Where;

$y$ is the n × 1 vector of observations,

$X$ is the model matrix which is n × p,

$\beta$ is p × 1 vector of unknown parameters for the coefficients of the model matrix for fixed effects,

$Z$ is a given n × q matrix for the coefficients for random effects,

$\gamma$ is an unobservable random vector in q × 1 dimensions and

$\varepsilon$ is n x 1 vector of errors.

(Davis, 2002).

Random effects are unobservable and unmeasurable. Therefore, estimating the random effect component of the model is the biggest challenge. Usually parameters in such models are estimated using iterative procedures, such as maximum likelihood approaches (Longford, 1987; Singer and Willlett, 2003; Beaumont, 2012).

Harville (1977) used maximum likelihood (ML) estimation to estimate both fixed effects and variance components (i.e random effects). It was found that when a ML estimation procedure is used to estimate variance components, it results in biased estimates of parameters. An alternative approach called restricted maximum likelihood (REML) estimation was suggested by Patterson and Thompson (1971). The difference between ML and REML is that the maximum likelihood approach (ML) uses the original observed values whereas the REML method uses the likelihood function calculated from transformed data. In REML approach, original data is replaced by "contrasts" which are the combination of two or more factor level means where the coefficients add up to zero and likelihood function is calculated from the contrasts. This takes account of the degrees of freedom that are lost while estimating fixed effects. This then results in less biased estimates of variance components (Goldstein, 1995).

For longitudinal data, the general LMM is;

$$y_i = X_i\beta_i + Z_i\gamma_i + \varepsilon_i \qquad i = 1,2,3,\dots n \qquad \text{eq. (5.3)}$$

Where for each individual *i*;

$y_i$ is the n × 1 vector of observations,
$X_i$ is the n × p model matrix,
$\beta_i$ is p × 1 vector of unknown parameters,
$Z_i$ is a given n × q matrix,
$\gamma_i$ is an q × 1 unobservable random vector and
$\varepsilon_i$ is n x 1 vector of errors.
(Laird and Ware, 1982; Jenrich and Schlucter, 1986; Laird *et al.*, 1987; Diggle, 1988; Lindstrom and Bates, 1988; Jones and Boadi-Boateng, 1991; Jones, 1993).

Laird and Ware (1982) stated that linear mixed models can be considered as a two-stage random effects model. For instance, in the dataset used in this research project, repeated measurements of eGFR values are taken from each patient and these repeated measurements can be classified as measurements at level 1 which are grouped within patients, whereas level 2 is defined as the between patient level. Therefore, at level 1 we can identify differences between measurements within a patient and at level 2, we can determine differences between the individual patients (Verbeke and Molenberghs, 2000).

The equation for level 1 is expressed as;

$$y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i \quad \text{Where } \varepsilon_i \sim N\ (0, W_i)$$

eq. (5.4)

Where $W_i$ is the non-constant variance for each $i$, $i = 1,2,3, \dots n$.

In this type of LMMs, an expectation-maximization (EM) algorithm (Laird and Ware, 1982; Lindstrom and Bates, 1988) is applied as an iterative procedure to obtain the estimates for ML and REML parameters; where the estimation step in the EM algorithm is defined as the E step and the maximization step is defined as the M step. By using a two-level random effects model, unobservable random parameters can be estimated given that there is no missing data in the repeated measurements. Jennrich and Schluchter (1986) proposed using LMMs to investigate unbalanced and incomplete responses with the main purpose being estimation of the model coefficients, the βs.

Where there are missing responses, then the total set of repeated measurements for each subject are expressed as a submatrix of the model matrix. When the covariance is estimated per patient i.e, it is subject-specific, then Jennrich and Schluchter (1986) suggest using Newton-Raphson (NR) and Fisher scoring (a different form of Newton's method) algorithms instead of an EM algorithm to find the ML and REML estimates (Longford, 1987). Jennrich and Schluchter developed a generalized expectation-maximization (GEM) algorithm where, at each M step of the GEM algorithm, the likelihood function is increased. The GEM approach is only applicable if the data is incomplete, but the main benefit is that it can be used to fit covariance matrices when the number of parameters in the dataset is large (Jennrich and Schluchter, 1986; Laird *et al.*, 1987).

Laird *et al* (1987) suggested using an EM algorithm with incomplete datasets to find the estimates for ML and REML parameters. In such cases, at each iteration step, m steps of repetition have to be carried out before moving to the next iteration step. In previous research, the EM algorithm is used in random effects models where the data is complete with no missing values. When the data is incomplete, i.e. in our case, there are missing eGFR readings, the EM algorithm is inappropriate and other estimation procedures are more favourable, for example the Newton- Raphson or Fisher scoring methods. Comparison between these two approaches (Newton-Raphson and Fisher scoring) concludes that the NR approach is usually preferred over Fisher scoring, as it generates estimates with lower standard errors. Laird *et al* (1987) further suggested that the generalized EM algorithm (GEM) can be used where data is incomplete, and this will be appropriate to a specific case of the random effects model. Therefore, using GEM approach allows the random effects models to be applied to incomplete data. Overall, Jennrich and Schuluchter (1986) concluded that where the number of covariate parameters is small, the NR approach is better due to its speed and the method being not limited to incomplete datasets. However, if the number of covariate parameters is large, measurements from more than 10 time points will be needed to fit the unstructured covariance matrix and in such cases the GEM algorithm is found to be better than other iterative procedures. In addition, using the GEM algorithm to estimate the random effects model does not require the covariance matrix to be specified in advance.

Diggle (1988) proposed a model;

$$y_i = X_i\beta + \varepsilon_i$$

eq. (5.5)

Where $\varepsilon_i \sim N(0,\Sigma_i)$ $\Sigma_i = t^2 I + v^2 J + \sigma^2 R(t_i)$

J is a square matrix every element equal to 1,

$t_i = \left[\, t_1, t_2, t_3, t_{4,\dots} t_{i_{ni}}\right]$ is the vector of measurement times.

R(t) is a symmetric matrix with $R(k,l) = exp( -\sigma|t_k - t_l|^c$ when c=1 or 2. (Diggle, 1988).

In equation (5.5) the within subject covariance structure contains three parameters, $t^2$, $v^2, \sigma^2$ and these parameters are estimated using the ML or REML approaches.

In the model formulation of linear mixed models, a covariance structure should be specified in advance. The most commonly used covariance structures that can be assumed in LMM approaches are; compound symmetry, first-order autoregressive (AR-1), independence and unstructured (Zimmerman, 2000).

Compound symmetry covariance structure (Tiwari and Shukla, 2011) is the basic, well known covariance structure where all variances are assumed to be the same and additionally, the correlation between any two successive repeated measurements are assumed to be same for all individuals irrespective of the time between these repeated measurements. The compound symmetry covariance structure is in the form;

$$\begin{bmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 \end{bmatrix}$$

Compound symmetry covariance structure assumes a constant variance and covariance among the repeated measurements.

The AR-1 covariance structure (Tiwari and Shukla, 2011) takes the form of;

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

Where ρ is the correlation between successive measurements.

The AR-1 structure is known as first order autoregressive covariance structure and it assumes that variances among repeated measurements taken from each patient are the same. It also assumes that the association between any two adjacent repeated measurements from the same patient is equivalent to rho (ρ) which indicates that ρ represents the direct relationship between successive measurements, but only indirect correlation between the measurements further apart. In the data used in this research, repeated measurements are taken from each individuals at various different time points. The first observation is denoted as being at ($t_0$) and the following observations are identified by time (in years) after the first measurement. Since the repeated measurements are not taken at equally spaced time values, covariance structures containing autoregressive models are not appropriate.

The independent covariance structure (Tiwari and Shukla, 2011) is in the form;

$$\begin{bmatrix} \sigma_A^2 & 0 & 0 \\ 0 & \sigma_B^2 & 0 \\ 0 & 0 & \sigma_C^2 \end{bmatrix}$$

The independent covariance structure is also called a diagonal covariance structure. The independent covariance structure assumes that variances between repeated measurements taken from each distinct patient are different and there is no association among repeated measurements from the same patient.

Unstructured covariance structure (Tiwari and Shukla, 2011) takes the form of;

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}$$

The unstructured covariance structure is the fully general covariance structure which does not have any restrictions on variance and covariance of repeated measurements either within or between patients other than that the variances and covariances are constant over time. In this way, different variances and covariance can be observed between repeated measurements.

While analysing the data using LMM techniques, the function of the hypothesis test used is dependent on how close the chosen covariance structure is to the true covariance structure of the dataset (Laird and Ware, 1982).

In general, to choose the best covariance structure for the data, different models are applied and compared. These different models are implemented using the same fixed effects where these fixed effects are estimated by using ML or REML approaches.

Where there is a small number of repeated measurements at equally spaced time points per subject, and when the data is complete and all of the independent predictor variables are categorical variables, then the likelihood ratio test can be used. However, where there are a large number of repeated measurements per subject not equally spaced in time, the likelihood ratio test is not suitable for identifying the best covariance structure.

When the models under consideration are nested, as they are here, models can be compared by looking at Akaike's (1974) Information Criteria (AIC) and Schwarz's (1978) Bayesian Information Criteria (BIC). These information criteria are based on the log-likelihood, both also consider the number of observations or/and number of parameters involved in the model (Vallejo *et al.*, 2011).

$$AIC= -2LL + 2P$$

$$BIC= -2LL + P\log(n)$$

Where;

LL is the log likelihood,

P is the number of parameters and

n is the number of observations (Davis, 2002).

It can be seen from the above equations that AIC penalizes the log-likelihood based on the number of parameters; whereas BIC penalizes log-likelihood according to number of observations and the number of parameters in the model (Jones, 1993; Vallejo *et al.*, 2011).

In the models investigated in this research project (see sections 5.4 & 5.6.1), the nature of the AIC and BIC mean that the lower the AIC (or BIC) values, the better the model fit. The random intercept and random slope models are used as initial linear mixed modelling approaches. These models are attractive in that they compute only four parameters to form the covariance structure but the pitfall is that they do not have a stable covariance matrix of the vector of observations.

The purpose of this research project is to identify within-subject differences as well as between-subject differences. Therefore, the covariance structure of the model should allow the intercept, which is the initial condition (i.e. based on the first of the all repeated measurements), to vary from patient to patient. As well as this, the model should enable the rate of change of the repeated measurements over time (i.e. the slope) to vary between patients. Thus, an appropriate covariance structure should take account of both a random intercept and a random slope. In addition, inclusion of independent within-subject variations in the covariance structure will allow modelling of within subject differences as well. Here, the most appropriate covariance structure of the model is the unstructured covariance structure. The model formulated using an unstructured covariance structure is then compared with the basic model which has the compound symmetry covariance structure.

LMMs are applied to our dataset using both types of covariance structures; compound symmetry and unstructured covariance structure. The two models are then compared, to identify which covariance structure is more applicable to the data.

## 5.4 Application of LMMs on Research Dataset

The linear mixed models described and used in this chapter follow a two level repeated measures structure where the lower level (level 1) observations are eGFR measurements over a period of time. Level 1 observations are nested within patients at the higher level (level 2).

Variation between eGFR measurements for each individual is analysed at level 1 and variation between individuals (patients) is analysed at level 2. In effect, this means that a different regression line will be estimated for each patient, and regression parameters specific to patient attributes, which are called "random effects", are modelled at level 1 (Laird and Ware, 1982).

The Linear mixed model is formulated as;

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i \qquad b_i \sim N_q\ (0,\psi) \qquad \varepsilon_i \sim N_{ni}\ (0,\sigma^2\Lambda_i)$$

eq. (5.6)

Where;

$Y_i$ is the response vector, which is the set of repeated eGFR measurements in $n_i \times 1$ dimensions, where $n_i$ is the total number of observations for individual $i$,

$X_i$ is the model matrix for the fixed effects, which is in $n_i \times p$ dimensions, where $p$ is the total number of fixed effects,

$\beta$ is the vector of fixed-effects coefficients in $p \times 1$ dimensions,

$Z_i$ is the model matrix for the random effects, which has $n_i \times q$ dimensions, where $q$ is the total number of random effects,

$b_i$ is the vector of random-effects coefficients in $q \times 1$ dimensions and, $\varepsilon_i$ is the vector of errors in $n_i \times 1$ dimensions,

$\psi$ is the covariance matrix for random errors in $q \times q$ dimensions,

and $\sigma^2\Lambda_i$ is the covariance matrix for errors in $n_i \times n_i$ dimensions. (Laird and Ware, 1982).

Initially, unconditional models, described below, are fitted to our data. These unconditional models are used to identify whether there is any systematic variation in the repeated eGFR values and, if so, where that variation lies- either *within* or *between* individuals. These are relatively primitive models but they are useful in providing a basic evaluation about within- and between-subject heterogeneity, which can be used to improve the subsequent models which will include substantive predictors (Peugh and Enders, 2005).

### 5.4.1 Model 1 - Unconditional means model

The first model was an unconditional mean model, a two-level model with a random intercept, used to estimate the total variation in the response. At this point, the model does not include any predictors at either level. This is equivalent to a one-way ANOVA model with a random effect (Sullivan *et al.*, 1999). The model estimates the grand mean of repeated eGFR values across all individuals and all measurement times and estimates the amount of variation existing in within- and between-subjects (without considering time). The Intra-Class Correlation Coefficient (ICC) is used to compare the relative magnitude of variance components by estimating the proportion of total variation in the responses that lies between patients (level 2) (Twisk, 2003). This represents the percentage of total variation in eGFR values that is attributable to differences between patients (Shek and Ma, 2011).

Model 1 (Unconditional mean model) has the form;

$Level\ 1: Y_{ij} = \beta_{0j} + r_{ij}$

$Level\ 2\colon \beta_{0j} = \gamma_{00} + u_{oj}$

$$Combined\colon Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

eq. (5.7)

Where;

$Y_{ij}$ is the measured outcome for individual i at measurement point j (i.e. the mean value for measurement j across all patients),

$\beta_{0j}$ is the group mean,

$r_{ij}$ is the residual term for an individual $i$ at measurement $j$,

$\gamma_{00}$ is the grand mean of the response values, across all patients and measurements and

$u_{0j}$ is the individual specific random effect term. (Kwok *et al.*, 2008).

### 5.4.2 Model 2 - Unconditional linear growth model

In an attempt to improve on Model 1, an unconditional linear growth model is applied. This model includes the single predictor, time. This baseline growth model allows investigation of variation in growth rates (in this case, deterioration of CKD status). The linear slope aspect of this model, represent rates of change of eGFR with respect to time for the individual and the slope is allowed to vary randomly between individuals (Shek and Ma, 2011).

Model 2 (Unconditional linear growth model):

$Level\ 1\colon Y_{ij} = \pi_{0i} + \pi_{1i}(Time_{ij}) + r_{ij}$

$Level\ 2\colon \pi_{0i} = \beta_{00} + u_{0i}$

$Level\ 2\colon \pi_{1i} = \beta_{10} + u_{1i}$

$$Combined\colon Y_{ij} = \beta_{00} + \beta_{10}(Time_{ij}) + u_{0i} + u_{1i}Time_{ij} + r_{ij}$$

eq. (5.8)

Where;

$Y_{ij}$ is the measured outcome for individual i at measurement j,

$\pi_{0i}$ is the intercept, which is the initial status of eGFR,

$\pi_{1i}$ is the slope, which is the rate of change of eGFR over time,

$r_{ij}$ is the time specific residual term for each individual,

$\beta_{00}$ is the mean intercept,

$u_{0i}$ is the individual specific random term for the intercept,

$Time_{ij}$ is the time of $j^{th}$ measurement for individual $i$, relative to that individual's first measurement,

$\beta_{10}$ is the mean slope,

and $u_{1i}$ is the individual specific random term in the slope (Kwok *et al.*, 2008).

### 5.4.3 Model 3 - Conditional Linear growth model

A further development of these models can be achieved by inclusion of predictor variables at the between-individual level (i.e. between subjects, level 2). In the case of this study, these are diagnoses of known co-morbidities at baseline (i.e. first eGFR measurement). Initially, a separate linear mixed model was produced for each of the co-morbidities considered (i.e. diabetes, cardiovascular disease, anaemia), followed by a single model containing all patient level predictors. The aim of this is to determine how much of the between-subject variation can be explained by knowledge of co-morbidity diagnoses at baseline for individual patients.

The conditional linear growth model (Peugh and Enders, 2005) take the form:

$$Level\ 1: Y_{ij} = \pi_{0i} + \pi_{1i}(Time_{ij}) + r_{ij}$$

$$Level\ 2: \pi_{0i} = \beta_{00} + \beta_{01}(X_i - \bar{X}) + u_{0i}$$

$$Level\ 2: \pi_{1i} = \beta_{10} + \beta_{11}(X_i - \bar{X}) + u_{1i}$$

$$\begin{aligned} Combined: Y_{ij} \\ &= \beta_{00} + \beta_{01}(X_i - \bar{X}) + \beta_{10}(Time_{ij}) + \beta_{11}[(X_i - \bar{X}) \times (Time_{ij})] + u_{0i} \\ &+ u_{1i}Time_{ij} + r_{ij} \end{aligned}$$

eq. (5.9)

Where;

$Y_{ij}$ is the measured outcome for individual i at measurement j,

$(X_i - \bar{X})$ is the difference between the predictor for individual i and the mean value of that predictor across all individuals,

$\beta_{00}$ is the mean intercept,

$\beta_{10}$ is the mean slope,

$\beta_{01}$ is the coefficient for the intercept in the difference explained above and

$\beta_{11}$ is the coefficient for the slope with respect to the same difference. (Kwok *et al.*, 2008).

### 5.4.4 Results from all three models

The sample of GP data used to formulate Model 1 – Model 3 contains only patients that have been diagnosed to have CKD at stages 3 to 5 from the first observation. Results obtained from all three models are given in Table 5.1 below.

Table 5.1: Results from all three models

| | **Model 1** | **Model 2** | **Model 3** |
|---|---|---|---|
| **Fixed Effects** | | | |
| **Intercept ($\gamma_{00}$)** | 48.91 | 51.19 | 52.17 |
| **(SE)*** | (0.394) | (0.392) | (0.480) |
| **p-value** | p<0.001 | p<0.001 | p<0.001 |
| **Time ($\gamma_{10}$)** | | -0.48 | -0.20 |
| **(SE)*** | | (0.060) | (0.105) |
| **p-value** | | p<0.001 | p<0.005 |
| **HbA1c** | | | 1.23 |
| **(SE)*** | | | (0.368) |
| **p-value** | | | p<0.001 |
| **Anaemia × Time** | | | -0.40 |
| **(SE)*** | | | (0.187) |
| **p-value** | | | 0.017 |
| **Cardiovascular Disease × Time** | | | -0.30 |
| **(SE)*** | | | (0.134) |

| | | | |
|---|---|---|---|
| **p-value** | | | 0.002 |
| **Random Effects** | | | |
| $u_{00}$ | 68.65 | 62.37 | 49.94 |
| $u_{11}$ | | 1.30 | 0.91 |
| $u_{01}$ | | -1.85 | -1.13 |
| $\sigma^2$ | 36.08 | 21.96 | 21.34 |
| **Overall Model Test** | | | |
| **-2LL**[**] | 25569.609 | 24630.347 | 13443.495 |
| **AIC**[**] | 25573.609 | 24638.347 | 13451.495 |
| **BIC**[**] | 25586.082 | 24663.291 | 13474.061 |

[*] Standard error in parentheses.

[**] Each one of these information criteria are such that the lower the value, the better the fit of the model to the data.

The unconditional mean model (Model 1) reveals that about one third of the total variation in the eGFR readings (34.45%) is within-subject variation, i.e. variation between eGFR readings over time within patients. The remaining two thirds (65.55%) of variation in the data is found between subjects, i.e. due to differences between the patients. This suggests that differences in average eGFR measures are due more to differences between patients than due to changes in eGFR measurements within each individual patient. The intra-class correlation coefficient (ICC) is the proportion of variance explained by between subject factors (Kwok *et al.*, 2008) and is 65.55% here.

Since our sample of the GP data contains only patients that have been diagnosed to have CKD at stages 3 to 5 from the first observation, which is the point where the measurement time was set to zero, the grand mean of the eGFR values resulting from the unconditional mean model (48.91 mL/min/1.73m$^2$) was considerably lower than the grand mean of the whole GP data set that contains all patients both with and without CKD (80.18 mL/min/1.73m$^2$).

In Model 2, the unconditional linear growth model, which estimates fixed effects parameters, statistically significant estimates were achieved for both intercept and linear slope, suggesting that initial status and linear growth rate for eGFR were non-zero across patients. The intercept value represents the expected initial eGFR value at time t=0, i.e. eGFR at the first CKD diagnosis for each patient. The intercept estimate is 51.19 mL/min/1.73m$^2$, a higher increase in the intercept compared to Model 1, indicating that the average initial eGFR measure is higher than the average over all observations (as estimated in Model 1) for most patients. The average rate of change of eGFR is estimated at -0.48 units per year, which indicates that on average the eGFR value for most patients was decreasing with time. The estimate of the residual in the covariance was reduced, meaning that a higher proportion of within-subject variability is explained, as a result of the addition of the level 1 covariate (i.e. time). Here, 39.14% of variation at level 1 (between eGFR readings) which is 13.48 % of the total variation was explained by time variable. However, it can be concluded that 60.87 % of within-individual variance (i.e. 20.97 % of the total variation) was still unexplained. The random error terms associated with the intercept and linear effect are also both found to be statistically significant. Therefore, by the addition of time variable into the unconditional means model, small amount of improvement is made on the explanation of variation in the outcome.

Furthermore, a significant amount of variation in the outcome can be explained by further adding between subject (i.e. level 2) factors and covariates into the unconditional linear growth model. In this way, further explanation of the variability in eGFR measurements between patients are achieved by expanding Model 2 to include between-subject (i.e level 2) factors and covariates, i.e. by moving from Model 2 to Model 3 (Kwok *et al.*, 2008).

The conditional linear growth model (Model 3) provides the best fit of all the three models. It includes baseline diagnosis of statistically significant co-morbidities; cardiovascular disease, anaemia, and the baseline value of HbA1c (the biomarker for diabetes). These parameters are included as level 2 factors and as covariates with time. The unstructured covariance structure was used, where all elements in the covariance structure are unique and estimated from the data separately.

The best fit combined model (Model 3) was found to be;

$$eGFR_{ij} = 52.17 + (-0.20)(Time_{ij}) + (1.23)(HbA1c_i) + (-0.40)(Anaemia_i) * (Time_{ij}) \\ + (-0.30)(Cardiovascular\ disease_i) * (Time_{ij})$$

eq. (5.10)

Where $eGFR_{ij}$ is the value of eGFR at time point $j$ for an individual $i$, $i$=1,2,3...472 (since we have 472 patients having 8 repeated eGFR measurements) and HbA1c is a continuous variable on the grand mean centred and $0 \leq Time_{ij} \leq 18$ years. $Time$ is a continuous variable in years indicating the actual time of each repeated measurement and $j$ ($j = 1,2,3,4,\ldots,15$) indexes the time points where the measurements for that individual were made. HbA1c; is a biomarker used in the diagnosis of diabetes which is reported using a continuous scale. (The grand-mean centred HbA1c value is used in the model to allow for meaningful interpretation of the intercept, i.e. HbA1c=0 represents the average value for the study group).

The baseline diagnoses of anaemia and cardiovascular disease were binary coded using the SPSS convention, where absence of disease is coded as 1 and presence of disease is coded as 0. ("CVD present" is defined as a previous/current diagnosis of cardiovascular disease (CVD), peripheral vascular disease (PVD) or ischaemic heart disease (IHD). However, this convention was not the case for the diagnosis of CKD where the diagnosis of CKD is kept coded as 1 and the absence of the disease is kept coded as 0.

In this model, (eq. 5.10), each factor included is found to have some statistically significant effect on eGFR. Since the independent variable called "HbA1c*time" is found to have p-value greater than 0.05, the interaction of HbA1c with time is found to have a non-significant effect on the eGFR value and hence, HbA1c was found to affect the intercept alone, indicating that this covariate has an effect on the baseline diagnosis of eGFR only, whereas presence of anaemia and cardiovascular disease were found to affect the rate of progression of the disease, but not the baseline value. Where a patient with CKD has none of the co-morbidities diagnosed, the average baseline eGFR value was found to be 52.17 mL/min/1.73m$^2$ and, for each one year increase in time, the eGFR value decreases by 0.20 units on average.

In general, diabetes is diagnosed when HbA1c values exceed 6.5 %. In our data set, the average HbA1c value is found to be 7.05 %, indicating that many of the CKD patients (around 50%) had a high HbA1c value and were therefore diabetic. An increase in average HbA1c value indicates a worsening of diabetic condition, whereas a decrease => improving diabetic condition. Our results show that for each one percentage point that a patient's HbA1c is above the average, meaning that their diabetic condition is worse, typically corresponds to an expected increase in the initial eGFR value of 1.23 units. Thus, more serious cases of diabetes would have a better initial eGFR value that is indicating a better initial state of kidneys. However, this initial increase in eGFR value does not mean that the condition of kidneys are getting better. Instead, it means that when a patient has diabetes, it has a greater chance of detection of CKD and having initial treatment for CKD results an initial rise in eGFR value. It is also concluded that over the time, patients having diabetes had a greater decline of eGFR compared to non-diabetic patients.

If a patient has been diagnosed with anaemia only, a decrease in eGFR of 0.60 units is expected for every additional year in time from diagnosis. This compares to a decrease of 0.20 units for a patient without a diagnosis of anaemia. Thus, anaemic patients tend to have a faster decline in kidney function than non-anaemic patients. Similarly, if a patient has been diagnosed with cardiovascular disease, each additional year after diagnosis will typically result in a decrease in eGFR of 0.50 units, compared to 0.20 units in non CVD patients. This means that diagnosis of either (or both) of these co-morbidities result in a faster decline of eGFR over time, increasing the rate of progression of CKD.

The covariance between intercept and slope in the final model was found to be negative and statistically significant. From this, assuming that the decline of eGFR over time is linear for now, we can deduce that when a patient has a higher initial (baseline) eGFR value, a faster decline of eGFR over time is expected, hence increasing the speed of progression of CKD towards the end stage renal disease (ESRD). We can further infer that patients who also have diabetes, and hence a higher initial eGFR due to the contribution of the HbA1c term in eq. (5.10), would show a more rapid decline in eGFR, despite this not being explicitly shown in eq. (5.10). Model 3 (i.e. the conditional linear growth model) is found to be the best fitting model of the three since it gave the lowest log-likelihood (-2LL), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values, indicating a better model fit. (Table 5.1) (Leyland, 2004).

Pseudo-$R^2$ statistics from the variation between all the intercepts for all individuals in the dataset (Model 3, conditional linear growth model) and average intercepts (Model 2, unconditional model) shows that 19.92 % of variation in the intercept (i.e. the variation of eGFR value between all patients at baseline) can be explained by Model 3. Pseudo-$R^2$ statistics from the variation between all annual changes (Model 3, conditional linear growth model) and the average annual change (Model 2, unconditional model) shows that 30.77 % of variance of the annual change in eGFR value across all patients is explained by Model 3. From these results, it can be concluded that the co-morbidities investigated here have medium to large effects on the eGFR value at the initial time point and large effects on the linear rate of change in eGFR value over time (Kwok *et al.*, 2008).

Results obtained from applying LMMs to this data suggest that the research presented here confirms existing reports that CKD is related to other clinical conditions, known as co-morbidities. While it is recognised that diabetes, anaemia and cardiovascular diseases have significant influence on eGFR and hence diagnoses of CKD, our findings show and evaluate the impact of these conditions on the rate of decline of CKD status. Interpretation of the results obtained from our models reveal a clear trend for decline in eGFR over time and indicate that progression of CKD was faster for patients with the co-morbidities anaemia, diabetes and cardiovascular diseases both singly and in various combinations. Here, patients with these pre-existing conditions who are subsequently diagnosed with CKD are found to be at higher risk of accelerated deterioration of kidney function over time.

## 5.5 Extensions of Linear Models

GLM models can be combined together with quasilikelihood methods to extend the GLM methodology for repeated measurements. There are four different types of such models (Zeger and Liand, 1992) namely; random-effects models (Laird and Ware, 1982), response condition models (Neuhaus, 1992), transition models (Diggle *et al.*, 2002) and marginal models (Liand and Zeger, 1986). All of these models can be used to analyse repeated measurements taken from the same subject.

### 5.5.1 Random Effects Model

Random effects models are also called linear mixed models, as described in a previous section (see section 5.3.2.4). In random effects models (Twisk, 2003), as well as considering the whole model as formed of two separate part, namely random and fixed components, subject-specific random effects are also added to the model equation. Random effects are different from the random component of the model. The random component of the model includes the individual repeated measurements whereas random effects take account of heterogeneity between subjects caused by unmeasured variables. In this way, the fact that the every individual is different is considered in the model formulation. Therefore, this methodology is assumed to describe all of the within-subject correlation included in the dataset. Responses for a subject are dependent on the random effect for that subject. However, since random effects for each subject are different, the set of responses from each subject are assumed to be different to and independent of those from each other subject. This means that responses within any subject are assumed to be dependent on the random effects and the responses between subjects are assumed to be independent. Besides, subject-specific random effects are assumed to be identically, independently distributed (i.i.d.), assuming that the random effects follow a normal distribution (Fieuws *et al.*, 2007; Verbeke *et al.*, 2012).

A random effects model, which is also known as "subject-specific" or "cluster-specific" approach, is used to model between-subject differences by using subject-specific effects (Zeger *et al.*, 1988; Neuhaus *et al.*, 1991). This means that within-subject effects are used to formulate between-subject differences. In this way, the between-subject variation is identified by investigating within-subject variation. When looking from the perspective of this research project, random effects models are considered to be the optimal linear models to analyse the repeated measurements in this dataset in order to examine the effect of factors and covariates on both an individual's response and on the average response of the population.

However, in this research, even random effects models are considered as initial modelling techniques, in the following parts of this section, alternative extensions of linear models are also explained and reasons for not choosing these alternatives are discussed.

### 5.5.2 Response-conditional models

Response-conditional models were studied by Neuhaus (1992). According to Neuhaus and Jewell (1990), response-conditional models are only applicable if the main purpose of the research is to determine the inter-dependence of repeated measurements within the individual rather than considering the effects of predictors on the response (Neuhaus and Jewell, 1990). Since the main aim of this research project includes looking at both within-individual and between-individual differences, and the effects of predictors on the response are important, these methods are not suitable for this research work.

### 5.5.3 Transition Models

In transition models (Twisk, 2003), it is taken into account that the current repeated measurement from a same subject is dependent on the previous value observed for that same particular subject. Therefore, the latest response is dependent on the previous responses for the same individual and hence the previous values are used as predictors in the model formulation. The function that describes the pattern of the behaviour of the outcome is not known and hence the vector containing the parameters that describes the functional behaviour of past observations are unknown in the model formulation. Furthermore, the variance of the response is conditional on the past responses and it is described by the product of an unknown scale parameter and the known variance function that is represented in terms of the mean of the responses being conditional on the past responses (Kon and Whitternore, 1979; Kaufmann, 1987; Ware *et al.*, 1988).

Transition models are appropriate for use in analysing the repeated measurements if the interest is focused on looking at the effects of time-dependent covariates on the within-subject responses (Neuhaus, 1992). Since the assumption of transitional models require the assumption that the response is based on the past observations, this type of model should be used if the response follows a stochastic process (i.e. assuming that the evolution of the response variable over time is a random process). In transitional models, only within-subject variation is considered and since in this research the main aim is to investigate the effect of factors and covariates between-subjects as well as looking at within subject differences, transitional models are not the optimal approaches to analyse the dataset of interest in this present work.

### 5.5.4 Marginal Models

A marginal model allows modelling of the mean response (μ) against independent predictors. Hence a marginal model would model the average eGFR value for a class of patients all having the same types of diseases (e.g. with both diabetes and anaemia). Marginal models include Generalized Linear Models (GLMs) and the types of models fitted using Generalized Estimated Equations (GEE) (Liang and Zeger, 1986).

In the GEE approach, a model is specified using a known link function, g(.), that can connect the predictors to the mean response ($\mu_{ij}$). In addition to the link function, another assumption is made about the distribution of the response and, finally, a working correlation matrix is assumed (Zeger and Liang, 1986). This working correlation matrix is used to take account of dependencies between the repeated measurements within each individual. Models created using GEE approaches are very similar to GLM models, where a common effect of the exposure (i.e. the effect of the diagnosis of a co-morbidity) is assumed over the entire population and the mean of the outcome is conditioned on the exposure (Zeger, 1988). The main differences between the GEE and GLM approaches lie in the estimation of beta coefficients in the model (represented as βs in the previous sections of this chapter) and the evaluation of standard errors in these beta coefficients. In GLMs, beta coefficients are estimated using maximum likelihood (ML) approaches, whereas in GEE methods, they are estimated using estimated equations (EE) (Liang *et al.*, 1992). Furthermore, in GLMs standard errors do not take any account of any correlation between repeated measurements for each individual, whereas in GEE methods, correlation between repeated measurements within an individual is modelled using the working correlation matrix structure. This gives the GEE approach greater efficiency over GLMs. Using robust estimators which rule out the influence of outlying observations (i.e. when trimmed mean is used instead of arithmetic mean in descriptive statistics) can result in more precise standard error estimates in GLMs (Twisk, 2003). However, modelling the correlation structure as close as possible to the true correlation structure adds even more accuracy in models using GEE approaches.

To rule out the use of the GEE methodology, a model using the GEE approach was fitted to a sample of the data. The sample included all patients who had been diagnosed with CKD at stages 3-5 from the first observation (i.e. with initial eGFR less than 60 mL/min/1.743m$^2$) and having exactly 8 repeated eGFR measurements (a total of 3776 measurements from 472 patients). In this data, 52.5% of the total 472 patients had been diagnosed with CVD, 47.5% of the total patients had been diagnosed with diabetes and 14.6% of the total patients had been diagnosed with anaemia. The mean eGFR value, with corresponding standard deviation was found to be 48.91 ± 10.23 mL/min/1.73m$^2$. The average value of the time measurement across all measurements for all patients was found to be 4.93 ± 3.52 years. The resultant model equation was given as;

$$\begin{aligned}\ln(eGFR) = 3.887 &+ 0.045(Diabetes\ diagnosis)_i + 0.023(CVD\ diagnosis)_i \\ &- 0.005(time)_{ij} - 0.016\big(Anaemia\ diagnosis)_{ij} * (time_{ij}\big) \\ &- 0.004(CVD\ diagnosis)_i * (time_{ij})\end{aligned}$$

eq. (5.11)

All of the predictors retained in eq. (5.11) were found be statistically significant at the 99% confidence level, with associated p-values less than 0.01. In the model presented in eq. (5.11), since the co-morbidities were binary categorized, therefore only taking values of zero or one (i.e. 1 when disease was diagnosed and 0 when disease has not been diagnosed) and hence, the interaction terms of these co-morbidities with time are not time-dependent covariates. For this reason, in eq. (5.11), only time-dependent covariate was time and since, marginal models as discussed below, evaluated the mean model, time was kept at the mean level which was 4.93 years.

Marginal models are also known as "population-averaged" models and such models are used to investigate the effect of covariates on the population-averaged response only. Even smaller standard errors were estimated in the marginal models, such as when the GEE approach was employed in the model presented in eq. (5.11), than those obtained using linear mixed models, parameter estimates were found to be very similar. A marginal model assumes that every individual has the same response to the co-morbidities (i.e. having the same co-morbidity status) and therefore does not allow individuals to vary in their initial states (i.e. their initial eGFR values), nor does it allow individuals to vary in their progression over time (i.e. their slopes). In marginal models even within-individual dependencies are taken account in the same way as in random-effect models, marginal models do not allow the researcher to investigate the division of the total variation into within-individual and between-individual components (Twisk, 2003). Therefore, since the interest in this research project is to look at the variation in the outcome in terms of between-individual and within-individual components, and to understand the progression of CKD over time by allowing individuals to differ in their both initial states and their slopes, just like transitional models, marginal models are also not ideal for the analysis of the type of repeated measurements in this dataset, particularly due to the nature of the research questions discussed in Chapter 1. Marginal models are only useful if the main interest of the study is to focus on population averages. In cases where the covariates are time-independent (i.e. constant over time), marginal models can be the optimal type of model to use in the analysis in order to predict expected averages (Zeger *et al.*, 1988; Neuhaus *et al.*, 1991; Graubard and Korn, 1994).

In random effects models and transitional models, a single model equation is used to formulate both within-individual and covariate effects. However, in marginal models, iterative procedures are used and a sequence of several equations are formulated to analyse the population-averaged response.

## 5.6 Generalized Linear Mixed Models (GLMMs) as Extensions of Linear Mixed Models

When the repeated responses are measured on a continuous scale and can be assumed to follow a normal distribution, then normal parametric methods such as linear mixed models can be applied. However, if the repeated responses are found to have a non-Gaussian distribution, then extensions of linear mixed models (i.e. generalisations such as GLMMs) can be used, provided that the response variable follows a known distribution such as Poisson, Bionomial, Gamma, etc. (Molenberghs and Verbeke, 2005) (see Figure 2.2).

The extensions of linear mixed models which are described below are based on the univariate general linear model (GLM). Such developments of GLMs are take account of random variables by the addition of random effect term in the model formulation (Nelder and Wedderburn, 1974) and also enables the analysis of other types of outcomes, such as binary outcomes. In general, the random variables in the random effect component of the model are assumed to be independently, identically, normally distributed (i.i.d.) with constant variance per random effect.

Wedderburn (1974) suggested an additional improvement to GLM models by establishing that the use of quasilikelihood can make GLMs suitable for a wider range of applications. However, some challenges (e.g. assuming that observations, both within and between subjects, are uncorrelated) that need to be considered still exist in GLM approaches. Therefore, mixed modelling approaches, such as Generalized Linear Mixed Models (GLMMs) are usually preferred.

A general linear model for Gaussian outcomes is;

$$y_i = \beta_0 + \beta_1 X_i + \sigma\varepsilon_i \qquad i = 1,2,3, \dots n$$

eq. (5.12)

Where;

$y_i$ is the vector of the outcome,

$\beta_0$ is the vector of intercepts,

$\beta_1$ is the vector of fixed-effects coefficients,

$X_i$ is a vector which represents the independent predictors,

$i$ is the individual,

$n$ is the total number of individuals and

$\varepsilon_i$ is a random error variable, where $\varepsilon_1, \varepsilon_2, \dots \varepsilon_n$ are each $\sim N\,(0,1)$ (McCullagh and Nelder (1989), Atkin *et al.* (1989), Dobson (1990)).

In the GLM model (Cnaan *et al.*, 1977), the outcome variable $y_i$ is assumed to be normally distributed with mean, $\mu_i$ and constant variance, $\sigma^2$, so $y_i \sim N\,(\mu_i, \sigma^2)$.

$$\mu_i = \beta_0 + \beta_1 X_i$$

eq. (5.13)

(Davis, 2002).

The main purpose of the GLM model is to examine the differences between the patient-specific means by using independent predictors, and to analyse associations between these independent variables and the outcome variable. In this way, the impact of each predictor on the response variable can be estimated. This is achieved by modelling the patient-specific means indirectly by first transforming the patient-specific mean via a link function (Albert, 1999). The link function is denoted by g(.) and the model then becomes;

$$g(\mu_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$$

eq. (5.14)

(Davis, 2002).

Where the function $g(\mu_i)$ is modelled as a linear function of the predictors.

The error term in the above equation, eq. (5.14), must be generalized, based on the transformation applied using the function, g(.).

The GLM model consists of a random component, a systematic component and a link between the random and systematic components. The random component is used to categorize the distribution of the response variable. The systematic component is included in the model to identify the independent predictor variables that will be used in the model.

The random and systematic components are connected using the link function g(.), so that the model can examine the influence of independent predictors on the mean of the repeated measurements for each patient (i.e. the patient-specific means) (Davis, 2002).

The appropriate link function, such as the identity link, log link or logit link, is chosen based on the distribution of the repeated measurements. Commonly used link functions include;

Identity link, where; $g(\mu) = \mu$

Log link, where; $g(\mu) = \log(\mu) = \ln(\mu)$

Logit link, where; $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ (Davis, 2002).

In many of the models that follow in this study, the log link function is used to allow the association between the mean of the responses and the independent predictors to be nonlinear.

When observations are taken at unequally-spaced time values, the problem of data being unbalanced caused by the measurements being unequally spaced is solved by considering generalized linear mixed models (GLMMs) (Jones and Boadi-Boateng (1991), Jones (1993)). GLMMs require either one of two assumptions to be valid. One assumption is to assume that the observed responses are uncorrelated; the alternative is to assume that responses are correlated and that the covariance matrix structure of the repeated measurements is continuous first-order autoregressive (AR-1). In GLMMs employed in this study, Maximum Likelihood estimations are carried out by using nonlinear optimization techniques. The greatest benefit of using such nonlinear optimization methods in ML estimation for GLMMs is the ability to estimate likelihood repetitively by breaking down the intervals between time measurements into small increments.

Generalization of linear mixed models enables GLMM methodologies to be used for both normal and non-normal outcomes. For instance, if the response follows a normal distribution, LMMs can be used or if the response follows a non-normal distribution such as the gamma distribution, Generalized Linear Mixed Model (GLMM) methods can be used via a link function. GLMM methodologies also allow the use of both continuous and discrete responses. In terms of covariates, both time-dependent and time-independent predictors can be used, and missing data can be naturally handled by the method given that the missing data mechanism is either missing completely at random (MCAR; where missing observation does not depend on observed or unobserved measurement) or missing at random (MAR; where given the observed data, the missing observation does not depend on unobserved data) (Keselman *et al.*, 2001).

Generalized Linear Mixed Models (GLMMs) are an extension of general linear models (GLM) which take both random and fixed effects into account and are used when the assumption of independence between observations is violated (e.g. in longitudinal studies, where repeated measurements are taken from the same individual) (Zeger and Karim, 1991; Berslow and Clayton, 1993; Brown and Prescott, 1999). GLMM models are the extension of LMMs to account for the response following various non-normal but standard distributions, such as the gamma distribution or binomial distribution. GLMMs allow the linear predictor to have, in addition to fixed effects, one or more random components, each with an assumed normal distribution of mean zero and constant variance for each one individual. In this way, the correlation between observations from the same individual is taken into account in GLMM models (Diggle *et al.*, 2002).

The general form of the GLMM is the same as in eq. (5.6). However, in a GLMM, instead of modelling the response itself, a link function is used to transform the response into a quantity which can be modelled using linear predictors. The link function is represented by a generic link that is denoted by g(.), and the linear predictor is formed from the combination of fixed and random effects, excluding the residuals.

A linear predictor has the form;

$$\eta = X\beta + Z\gamma$$

eq. (5.15)

Where;

**X** is the matrix of predictors (e.g. independent variables), with corresponding $\beta$s which are the fixed effect parameter coefficient estimates for these regressors,

**Z** is the matrix of variables having random effects, with corresponding random effects denoted by $\gamma$ where both $\beta$ and $\gamma$ are vectors of the corresponding coefficients (Hedeker, 2005).

An inverse link function, denoted by $h(.) = g^{-1}(.)$, is used to convert the transformed response back to the original dimensions. For non-Gaussian data, the assumption of correlation between individual measurements is different from that for Gaussian data and hence results in different interpretations of the regression coefficients in the model (Diggle *et al.*, 2002).

In GLMMs, the effect of a covariate on the mean response for that individual is estimated conditionally based on the random effect for that individual (McCulloch *et al.*, 2008). The type of the link function and corresponding family of the distribution for the outcome is chosen based on whether the outcome is binary, discrete or continuous (Molenberghs and Verbeke, 2005; Zhang *et al.*, 2008).

In this study, the eGFR responses are measured on a continuous scale and since the distribution of eGFR responses are skewed, underline distribution is assumed to be gamma distribution. The most common link functions used with gamma distribution are inverse link function, log link function and identity link function as discussed before. Since the aim is to transform the responses, so that the responses will follow a normal distribution, identity link function is not appropriate. When both link functions (inverse link and log link) are employed on the same data by using gamma distribution with same fixed and random effects, the models are compared and log link function with a gamma distribution is shown to be the most appropriate choice in this study, proving a better fit to the model.

The log link function (Azeuro *et al.*, 2010) is defined by;

$$g(u) = \log(u)$$

eq. (5.16)

Where here log, means the natural logarithm (log to the base $e$).

The inverse link function is then defined as;

$$u = h(s) = e^{\log(u)} = e^{s}$$

eq. (5.17)

The probability density function for the general gamma distribution (Hedeker, 2005) is defined as;

$$f(x,a,b) = \frac{1}{b^{a}\Gamma(a)}x^{a-1}e^{-\frac{x}{b}} \qquad a > 0,\ b > 0,\ 0 < x < \infty$$

eq. (5.18)

Where "$a$" is the shape parameter and "$b$" is the scale parameter of the gamma distribution and $\Gamma(a)$ is the gamma function, and mean and variance of the distribution is defined by;

$$E(X) = ba$$

$$Var(X) = b^{2}a$$

eq. (5.19)

The gamma distribution used in this research represents a general family of distributions where the exponential distribution and chi-square distribution are special cases with $a$ = 1 and with $b$ = 2 respectively. The gamma distribution is employed in this research, so that the original responses (i.e. eGFR values) can be transformed by using gamma distribution to meet the normality assumption required by the GLMMs.

### 5.6.1 Application of GLMMs to this Research Dataset

A main objective of this study is to determine if there is any association between the changes in eGFR over time and the co-morbidities of interest (i.e. diagnoses of anaemia, diabetes and cardiovascular diseases), and so the next step is to build the best possible model for our data.

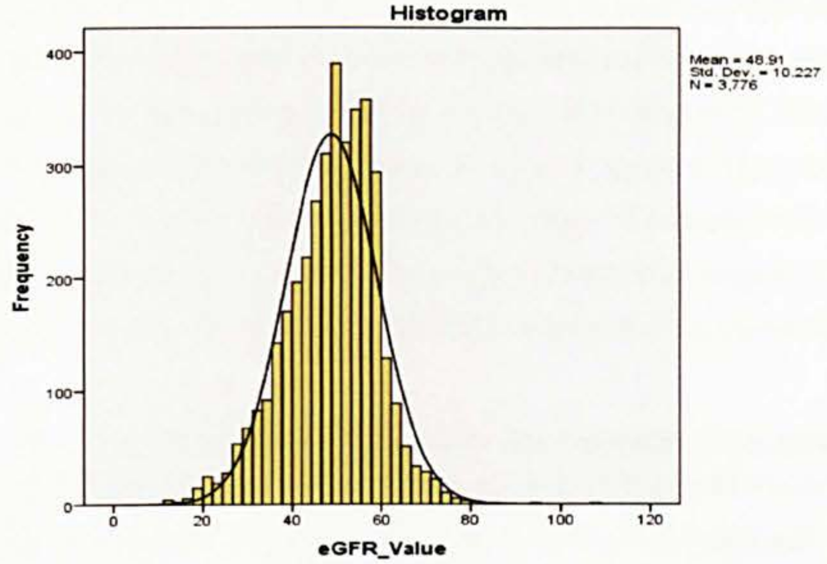

Figure 5.1: Distribution of eGFR values across all measurements for all patients within our dataset who have been diagnosed with CKD between stages 3 to 5. The solid line is the best-fitting normal distribution.

A histogram of the distribution of eGFR values across all the measurements for all patients (Figure 5.1) shows that the data is negatively skewed (i.e. the peak value of the data is to the right of that for the best-fitting normal distribution).

Since the normality assumption assumed under the LMM method is violated, we look instead to using the GLMM methodology, assuming a gamma distribution for the response and using a log link function. The random effect here is a correction appropriate for a particular individual patient. For both the linear mixed models and generalized linear mixed models, the coefficients were computed using restricted maximum likelihood estimation in SPSS.

In total, five models (1 LMM and 4 GLMMs) were created and analysed for our data set. The purpose of forming out five different models in total is to find the "best" model with fewest parameters, best fit to the data, and with the least complex covariance structure. As each of these models are computed, the "goodness of fit" of each one to our data is found using the Akaike Information Criterion (AIC), -2 LogLikelihood (-2LL) and Bayesian Information Criterion (BIC), in order to compare the models and ensure that a better model is obtained at each step (see Table 5.8). Each of these criteria is such that the lower (less positive or more negative) the value of the statistic, the better the model fits the data (Lv and Liu, 2013). In order to keep consistency between models and to make fair comparisons between them, the same co-morbidities are initially included in each of the models studied, and in all five models, the same sample of the research dataset is used as for the LMM and the GLMM (i.e. using eq. (5.6), namely 472 patients with CKD at stages 3-5 with 8 repeated observations for each patient, resulting in 3776 observations of eGFR values in total). The results of the models are presented in Tables 5.2-5.3 and Tables 5.5-5.7. The results in these tables show the best fitting models (i.e those containing only the statistically significant coefficients for co-morbidities of each type).

### 5.6.1.1 Model 1 – Modelling eGFR directly as a function of comorbidities and time

Initially, Model 1 is produced assuming a normal distribution with identity link function. This model is essentially described in eq. (5.6) using the LMM approach. In this model, the co-morbidities and other terms found to be significant and hence taken into account are diagnoses of diabetes and cardiovascular diseases at baseline and time and the interaction between time with the diagnoses of anaemia and of cardiovascular disease. The coefficient values for these terms, their standard errors and significance levels are given in Table 5.2.

Table 5.2: Results of Model 1 – eGFR as a linear function of co-morbidities and time

| **Model 1**<br>**Model Term** | **Coefficient** | **Standard Error**<br>**in coefficient** | **P-Value** |
|---|---|---|---|
| Time (in years) | -0.213 | 0.085 | $0.013^{*}$ |
| Diagnosis of CVD | 2.066 | 0.777 | $<0.001^{***}$ |
| Diagnosis of Diabetes | 3.016 | 0.724 | $0.008^{**}$ |
| (Diagnosis of Anaemia)*Time | -0.567 | -0.567 | $<0.001^{***}$ |

$^{*}$ p<0.05, $^{**}$ p<0.01, $^{***}$ p<0.001

The coefficients which best fit our CKD data are calculated and reported in the eq. (5.20) for Model 1 which is;

$$y = 48.592 + 3.016(diabetes\ diagnosis)_i + 2.066\ (CVD\ diagnosis)_i - 0.213(time)_{ij} - 0.567\ (Anaemia\ diagnosis)_i * (time_{ij}) - 0.328\ (CVD\ diagnosis)_i * (time)_{ij}$$

eq. (5.20)

Where $y$ represents the eGFR value, and each diagnosis is 1 if the disease is present or 0 otherwise. All other potential coefficients proved not to be statistically significant.

### 5.6.1.2 Model 2 – Modelling log(eGFR) as a linear function of the co-morbidities and time

In order to investigate whether a multiplicative rather than an additive model would be more appropriate for this data, the eGFR values are transformed to the natural logarithmic domain that forms log(eGFR) which we assume to be normally distributed. Model 2 is computed using a normal distribution but with log link function. The coefficients found, together with their standard errors and significance levels, are given in Table 5.3. The equation for Model 2, eq. (5.22), with only significant coefficients retained was found to be;

$$\ln(eGFR) = 3.877 + 0.060(Diabetes\ diagnosis)_i + 0.043(CVD\ diagnosis)_i - 0.007(time)_{ij} - 0.013(Anaemia\ diagnosis)_i * (time_{ij}) - 0.006(CVD\ diagnosis)_i * (time_{ij})$$

eq. (5.21)

Since Model 2 is evaluated in the log domain, very different coefficients are observed from before. When the AIC, BIC and -2LL information criteria for Models 1 and 2 are compared, it can be observed that transforming the eGFR values using the natural logarithm improved the model fit by a large amount, even though normality assumption was still retained (see Table 5.4).

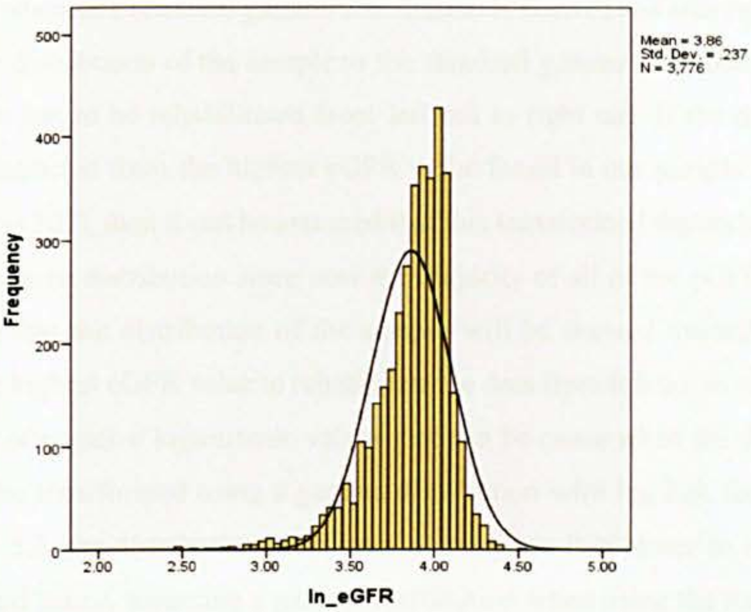

Figure 5.2: Distribution of ln(eGFR) values across all measurements for all patients within our dataset who have been diagnosed with CKD between stages 3 to 5. The solid line is the best-fitting normal distribution.

Table 5.3: Results of Model 2 – log(eGFR) as a linear function of co-morbidities and time

| **Model 2**<br>**Model Term** | **Coefficient** | **Standard Error** | **P-Value** |
|---|---|---|---|
| Intercept | 3.877 | 0.016 | $<0.001^{***}$ |
| Time (years) | -0.007 | 0.002 | $<0.001^{***}$ |
| Diagnosis of CVD | 0.043 | 0.016 | $0.007^{**}$ |
| Diagnosis of Diabetes | 0.060 | 0.015 | $<0.001^{***}$ |
| Diagnosis of Anaemia*Time | -0.013 | 0.004 | $0.003^{**}$ |
| Diagnosis of CVD*Time | -0.006 | 0.003 | $0.021^{*}$ |

$^{*}$ p<0.05, $^{**}$ p<0.01, $^{***}$ p<0.001

### A step to adjust the non-normality of the research dataset in using GLMMs

Since the sample from the research dataset contains only the patients with CKD at stages 3-5, the majority of all of the eGFR values across all patients in this sample are lower than the average eGFR value of the whole dataset. This makes the distribution of the sample to be skewed towards left whereas a standard gamma distribution is skewed towards right. Therefore, in order to ament the distribution of the sample to the standard gamma distribution, the majority of the eGFR values has to be rehabilitated from left tail to right tail. If the dependent variable (i.e. eGFR) is subtracted from the highest eGFR value found in our sample dataset (i.e. a constant value which is 107), then it can be assumed that this transformed dependent variable will follow a standard gamma distribution since now the majority of all of the eGFR values will be on the right tail, so that the distribution of the sample will be skewed towards right. The reason of choosing the highest eGFR value to rehabilitate the data from left tail to right tail is the eliminate the problem of negative logarithmic values that can be cause when the dependent variable (i.e. eGFR) will be transformed using a gamma distribution with log link function. As can be seen from Figure 5.3, the distribution is still not normal, but it is closer to normality compared to Figure 5.2 and hence, assuming a gamma distribution when using the transformed eGFR value rather than eGFR value itself is considered superior.

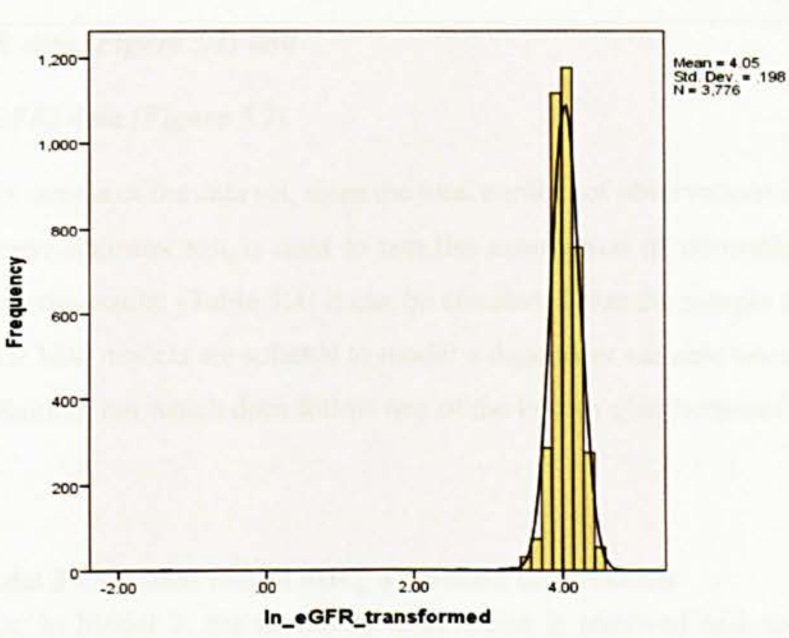

Figure 5.3: Distribution of ln(Transformed eGFR) values across all measurements for all patients within our dataset who have been diagnosed with CKD between stages 3 to 5. The solid line is the best-fitting normal distribution.

**Test of Normality:**

$H_0$: *The sample data are statistically not different from a normal population*

$H_1$: *The sample data are statistically different from a normal population*

Table 5.4: Normality Test Results for eGFR and ln(eGFR) datasets

| | **Kolmogorov-Smirnov** | | | **Shapiro-Wilk** | | |
|---|---|---|---|---|---|---|
| | **Statistics** | **df** | **Sig.** | **Statistics** | **df** | **Sig.** |
| **(a)** eGFR Data (Figure 5.1) | | | | | | |
| **eGFR Value** | 0.054 | 3776 | <0.001 | 0.985 | 3776 | <0.001 |
| **(b)** ln(eGFR) Data (Figure 5.2) | | | | | | |
| **ln(eGFR) Value** | 0.101 | 3776 | <0.001 | 0.917 | 3776 | <0.001 |

***(a) for eGFR data (Figure 5.1) and***

***(b) for ln(eGFR) data (Figure 5.2).***

In this sample of the data set, since the total number of observations is greater than 2000, the Kolmogorov-Smirnov test is used to test the assumption of normality of the dependent variable. From the results (Table 5.4) it can be concluded that the sample data is not normally distributed. GLMM models are suitable to model a dependent variable which does not follow a normal distribution, but which does follow one of the known distributions from the exponential family.

### 5.6.1.3 Model 3 – GLMM model using a Gamma distribution

Hence, in Model 3, the normality assumption is removed and, the eGFR values are modelled using a gamma distribution with log link function. In this way, the natural logarithm of the mean of the eGFR values over all measurements across all patients is modelled, the new model coefficients were calculated and these, their standard errors and significance values are given in Table 5.5. The equation with best coefficient values for this model (Model 3) with only statistically significant terms retained is found to be;

$$\ln(eGFR) = 3.869 + 0.064(Diabetes\ diagnosis)_i + 0.050(CVD\ diagnosis)_i - 0.006(time)_{ij} - 0.015(Anaemia\ diagnosis)_i * (time_{ij}) - 0.007(CVD\ diagnosis)_i * (time_{ij})$$

eq. (5.22)

Table 5.5: Results of Model 3 – GLMM model using a Gamma distribution

| **Model 3**<br>**Model Term** | **Coefficient** | **Standard Error** | **P-Value** |
|---|---|---|---|
| Intercept | 3.869 | 0.017 | $<0.001^{***}$ |
| Time (years) | -0.006 | 0.002 | $0.003^{**}$ |
| Diagnosis of CVD | 0.050 | 0.050 | $0.004^{**}$ |
| Diagnosis of Diabetes | 0.064 | 0.064 | $<0.001^{***}$ |
| Diagnosis of Anaemia*Time | -0.015 | 0.005 | $0.004^{**}$ |
| Diagnosis of CVD*Time | -0.007 | 0.003 | $0.014^{*}$ |

$^{*}$ p<0.05, $^{**}$ p<0.01, $^{***}$ p<0.001

A gamma distribution is usually employed when the data is positively skewed. However, in this study the data is negatively skewed (see Figure 5.1). Therefore, when the difference between the information criteria for Model 3 and Model 2 are compared with the corresponding difference between Model 2 and Model 1 (see Table 5.8), only small improvements are observed in the former case. In order to get a further improved model, the eGFR values are first manipulated to reverse the shape of the distribution from negatively-skewed to positively-skewed (i.e. so that the peak value of the data will be to the left side of that for the corresponding normal distribution). This transformation of eGFR values is carried out by subtracting each eGFR value from the whole number just greater than the maximum eGFR value found amongst all our CKD patients (i.e. max eGFR=107). This ensures any potential problems due to having to find the logarithm of a negative-valued quantity are removed. In this way, the distribution is changed to positively-skewed and hence will be more appropriate for being modelled using a gamma distribution in the analysis.

### 5.6.1.4 Model 4 – GLMM using log link function and Gamma distribution to model transformed data

Therefore, Model 4 is performed on the manipulated eGFR values as described above as response variable, using a gamma distribution with log link function. The equation with optimal coefficients for this model (Model 4), with only statistically significant terms retained is found to be;

$$\begin{aligned}\ln(107 - eGFR) &\\ &= 4.060 - 0.052(Diabetes\ diagnosis)_i - 0.033(CVD\ diagnosis)_i \\ &+ 0.003(time)_{ij} + 0.009\big(Anaemia\ diagnosis)_i * (time_{ij}\big) \\ &+ 0.005\big(CVD\ diagnosis)_i * (time_{ij}\big)\end{aligned}$$

eq. (5.23)

The full set of coefficients and the significant values are given in Table 5.6.

Table 5.6: Results of Model 4 – GLMM using log link function and Gamma distribution to model transformed data

| Model 4<br>Model Term | Coefficient | Standard Error | P-Value |
|---|---|---|---|
| Intercept | 4.060 | 0.012 | $<0.001^{***}$ |
| Time (years) | 0.003 | 0.001 | $0.05^{*}$ |
| Diagnosis of CVD | -0.033 | 0.013 | $0.014^{*}$ |
| Diagnosis of Diabetes | -0.052 | 0.012 | $<0.001^{***}$ |
| Diagnosis of Anaemia*Time | 0.009 | 0.003 | $0.001^{***}$ |
| Diagnosis of CVD*Time | 0.005 | 0.002 | $0.002^{**}$ |

$^{*}$ p<0.05, $^{**}$ p<0.01, $^{***}$ p<0.001

When the information criteria for Model 4 and Model 3 are compared, a major improvement is observed in all the three measured goodness of fit criteria, indicating that Model 4 is a much better model for this data than Model 3 (see Table 5.8).

The association between the initial eGFR status and the progression of eGFR over time is estimated by calculating the covariance matrix. The "unknown structure" of the covariance matrix is estimated by the SPSS software. In each of the models above (Models 1 to 4), the covariance matrix was evaluated with the "unstructured" covariance option selected and the programme then estimated the covariance. However, for each of these models, the covariance between the intercept and the slope is estimated to be zero. Therefore, a simpler covariance matrix structure, such as a variance component (diagonal) matrix, could possibly be used to achieve a better model with lower computational requirements.

#### 5.6.1.5 Model 5 – GLMM assuming a simpler covariance matrix

In Model 5, the process used to obtain Model 4 is repeated, but with the "variance component" option rather than "unstructured" selected for the form of the covariance matrix in the calculations. In this approach, a better fit to the data (in terms of information criteria) is achieved using this simpler covariance matrix (see Table 5.8). The coefficients, their standard errors and significance levels for this simpler model (Model 5) are given in Table 5.7. Model 5 with only statistically significant terms retained is;

$$\begin{aligned}\ln(107 - eGFR)&\\ &= 4.060 - 0.050(Diabetes\ diagnosis)_i - 0.033(CVD\ diagnosis)_i\\ &+ 0.003(time)_{ij} + 0.008\big(Anaemia\ diagnosis)_i * (time_{ij}\big)\\ &+ 0.005(CVD\ diagnosis)_i * (time_{ij})\end{aligned}$$

eq. (5.24)

Table 5.7: Results of Model 5 – GLMM assuming a simpler covariance matrix

| Model 5<br>Model Term | Coefficient | Standard Error | P-Value |
|---|---|---|---|
| Intercept | 4.060 | 0.012 | <0.001*** |
| Time (years) | 0.003 | 0.001 | 0.05* |
| Diagnosis of CVD | -0.033 | 0.013 | 0.013* |
| Diagnosis of Diabetes | -0.050 | 0.012 | <0.001*** |
| (Diagnosis of Anaemia)*(Time) | 0.003 | 0.002 | 0.003** |
| (Diagnosis of CVD)*(Time) | 0.002 | 0.003 | 0.005** |

* p<0.05, ** p<0.01, *** p<0.001

Table 5.8: Model Comparison

| Model | Type | AIC | BIC | -2LL |
|---|---|---|---|---|
| 1 | LMM | 24553.059 | 24621.572 | 24530.989 |
| 2 | GLMM | -4489.937 | -4421.424 | -4512.007 |
| 3 | GLMM | -4450.508 | -4381.995 | -4472.578 |
| 4 | GLMM | -5724.751 | -5656.239 | -5746.822 |
| 5 | GLMM | -5733.991 | -5671.701 | -5754.049 |

When comparing all five models (Model 1-5) in Table 5.8, the lowest AIC, BIC and -2LL values are found for Model 5 and, hence, it can be concluded that of these, Model 5 is the best-fitting model for our data. The results from all five models indicated that statistically significant parameters in all cases (to account for the changes in initial value of eGFR (i.e. the intercept) across patients) are the diagnoses of CVD and of diabetes, whereas the parameters included to describe the progression of CKD (i.e. the slope) and the effect of co-morbidities on this are in all cases the interaction of the diagnoses of anaemia and of CVD with time.

For evaluation and interpretation of the coefficients in terms of eGFR values, values obtained using Model 5 and appropriate values for the diagnoses and time are transformed back by exponentiation (i.e. the inverse of taking the logarithm), and then subtracting the result from

107, to give meaningful model-predicted eGFR values. The mean eGFR value at time zero is found from Model 5 to be 49.0257 mL/min/1.743m$^2$, given that the patient has not been diagnosed to have CVD or diabetes. If the patient has been diagnosed to have CVD only, this eGFR value at time zero rises to 50.9076 mL/min/1.743m$^2$, whereas if the patient has been diagnosed to have diabetes only, the resulting eGFR value is 51.8531 mL/min/1.743m$^2$at time zero. This means that both diagnoses of CVD and of diabetes tend to increase the initial eGFR value, with the effect of having diabetes being more influential than of having CVD at baseline. However, each yearly increase in time results in a decrease in this predicted eGFR value by a factor of 0.9964 if the patient has none of the co-morbidities. Thus, when a patient has not been diagnosed with CVD or diabetes, then the initial eGFR value (i.e. 49.0257 mL/min/1.743m$^2$) at time zero would be expected to decrease to 48.8515 mL/min/1.743m$^2$ after one year. This decrease of 0.9964 units in one year will be greater if the patient has either anaemia, CVD or both (see eq. (5.24)).

Each regression coefficient was estimated by using a robust method, hence resulting in the corresponding standard errors being low. The regression coefficients for the parameters affecting the progression of CKD show lower standard errors (i.e. are each less than 0.005) than those standard errors for the regression coefficients for the parameters affecting the initial eGFR value (i.e. the standard errors for these are between 0.010 and 0.015).

The higher eGFR values observed for patients with CVD indicates that, for some initial period, they will have better (i.e. higher) eGFR values than patients without that disease. However, the faster rate of decline in eGFR for patients with CVD results in lower eGFR values than those for non-CVD patients after some time, typically around 4.125 years.

Overall, the most appropriate ways of analysing longitudinal data with repeated measurements when the data is incomplete and unbalanced (as it is on this study) are found to be use of methodologies known as Linear Mixed Models (LMMs) and Generalized Linear Mixed Models (GLMMs). LMMs are used when the outcome measure can be assumed to follow a normal distribution, whereas GLMMs are applied otherwise, i.e. when this normality assumption is violated. However, some standard distribution should be assumed for the data in order to be able to perform the GLMM approach, and here a gamma distribution is used since the distribution of the outcome is skewed. Furthermore, with this choice of distribution a natural logarithm link function is used to transform the response. Here, the results obtained from the GLMM approach used in Model 5 indicates that when a patient has a diagnosis of CVD or diabetes, that patient will have a higher initial eGFR value compared with a patient without those diseases. However, having either CVD or anaemia will increase the rate of decline of eGFR, and hence the progression of CKD. The models could be improved if a distribution that better fits the data were used instead of assuming a gamma distribution.

The results of this study are consistent with those of previous research on the progression of CKD (de Lusignan, *et al.*, 2009). However, our work is based on a large sample of routinely-collected General Practice patient records, in contrast to the more usual controlled cross-sectional studies or clinical trials. Our results provide evidence that the methodological approach presented here applied to this routinely collected data is a useful and appropriate mechanism for investigating dynamic relationships within health-related data.

## 5.7 Centring in Regression Analysis and Polynomial Mixed Models

In research papers employing regression analysis, centring of data is not commonly reported (Paccagnella, 2006). However, data centring can be very important in interpretation of research results, as it selects an appropriate reference value for an outcome and modelling of the outcome is relative to that reference value (Kraemer and Blasey, 2004). Hence, in the context of time-dependent models, use of appropriate time centring will ensure that the level-1 (i.e. within individual) growth parameters are meaningful, which will allow the interpretation of the level-1 intercepts as the different statuses of individuals at the reference value of time, and the level-1 gradients as the average rate of change of eGFR for those individuals. The main purpose of centring in regression analysis is to deal with multicollinearity (Glanz and Slinker, 2001).

In this section, initially four linear models are computed to show the effects of different techniques of centring of the time variable. Comparison between these four linear models indicated that mean centring worked best for this data. Subsequently, two more models are analysed to examine non-linear changes in eGFR values over time, using polynomial functions of time, after centring the time variable using the mean centring technique which had been found to be the best for the linear models applied to our data. In all four initial (linear) models, both the diagnoses of three major co-morbidities of CKD - namely cardiovascular diseases (CVD), diabetes and anaemia - and their interactions with time, were taken into account. All statistically non-significant parameters are then removed from each model, using a backward elimination procedure.

Table 5.9: Results of the initial four linear models, with different centrings of time

| Model Number | Model Equation |
|---|---|
| Model 1 | $y_{ij} = 49.73 + 3.07(Diagnosis\ of\ Diabetes)_i - 0.35(Time_{ij})$<br>$- 0.24(Diagnosis\ of\ Cardiovascular\ Diseases)_i * (Time_{ij})$ |
| Model 2 | $y_{ij} = 45.69 + 2.94(Diagnosis\ of\ Diabetes)_i - 0.33(Time_{ij})$<br>$- 0.29(Diagnosis\ of\ Cardiovascular\ Diseases)_i * (Time_{ij})$ |
| Model 3 | $y_{ij} = 48.10 + 3.03(Diagnosis\ of\ Diabetes)_i$<br>$- 4.28(Diagnosis\ of\ Anaemia)_i - 0.22(Time_{ij})$<br>$- 0.46(Diagnosis\ of\ Anaemia) * (Time_{ij})$<br>$- 0.37(Diagnosis\ of\ Cardiovascular\ Diseases)_i * (Time_{ij})$ |
| Model 4 | $y_{ij} = 47.94 + 3.10(Diagnosis\ of\ Diabetes)_i$<br>$- 4.51(Diagnosis\ of\ Anaemia)_i - 0.22(Time_{ij})$<br>$- 0.46(Diagnosis\ of\ Anaemia)_i * (Time_{ij})$<br>$- 0.37(Diagnosis\ of\ Cardiovascular\ Diseases)_i * (Time_{ij})$ |

Equations for Models 1 to 4, show the best fit models, i.e. those including statistically significant parameters (Table 5.9). All the included coefficients are significant at the 2.5% ($p < 0.025$) level, but none of the removed coefficients were significant at the 5% ($p < 0.05$) level. In each model the time variable is centred in a different way.

For Model 1, time is measured relative to the first observation of the data sequence for each subject, which represents the initial eGFR status for that subject. In Model 2, the time is measured relative to the last observation in the data for each patient, which is the final status for that particular individual. For Model 3, the time variable for each individual is centred on the mean time of all his/her eGFR measurements. Model 4 is calculated using a time variable which is centred on the median time of that individual's eGFR measurements.

From the results of all four models (see Table 5.9), for a typical individual, the average eGFR value at initial status (i.e. first measurement) is expected to be 49.73 mL/min/1.743m$^2$, whereas it is found to be 45.69 mL/min/1.743m$^2$ at the final status (i.e. last measurement), 48.10 mL/min/1.743m$^2$ at the mean measurement time and 47.94 mL/min/1.743m$^2$ at the median time, all of these eGFR values assume that the individual patient has not been diagnosed with CVD, diabetes or anaemia at the corresponding time point.

The coefficient of time is defined to be the rate of change of the eGFR value per year. It can be concluded that, for all four models, (Table 5.9) the eGFR value for a typical CKD patient decreases by around 0.2-0.3 units per year increase in time and that diagnosis of diabetes is a statistically significant factor in all four models, increasing the corresponding eGFR intercept value by about 3.0 units. The interaction term of CVD with time is another parameter that is also statistically significant in all four models, accelerating the rate of decline of eGFR value by about 0.3 units per year. Additionally, in Models 3 and 4, diagnosis of anaemia and its interaction with time are observed to be statistically significant, reducing the eGFR intercept by about 4.5 units and increasing the rate of decline of eGFR value by 0.46 units per year. When random effects in each of the four models are compared, there is not much difference in the within-person residual variance between the four models. The different centrings also have negligible effect on the between- person residual variance in the rate of change. However, it is found that centring does have an effect on the covariance between the eGFR intercept and the coefficient of time.

Table 5.10: Model-fit results for the linear models with various time centrings

| Model Number | Description | AIC and BIC criteria |
|---|---|---|
| Model 1 | Time measurement from first | AIC = 24618.161, BIC = 24643.103 |
| Model 2 | Time measurement from last | AIC = 24589.763, BIC = 24614.705 |
| Model 3 | Time measurement centred on mean | AIC = 24578.116, BIC = 24603.055 |
| Model 4 | Time measurement centred on median | AIC = 24584.859, BIC = 24609.799 |

In Table 5.10, the Akaike and Bayesian Information Criteria goodness of fit statistics are shown for all four models computed in this section of this chapter. From these results, although the goodness of fit differences is small, it can be concluded that Model 3 is the best-fitting model for our data, since it shows the lowest AIC and BIC values. Since Model 3 used a time variable centred on the mean measurement time for each individual, this approach is also adopted in computing two further models, using quadratic and cubic polynomials in time, in order to account for non-linear changes of eGFR over time. Model 5 allows quadratic dependence on time, whereas Model 6 permits cubic dependence on time, time being mean centred (i.e. centred on the mean time measurement for each individual) in each case. The polynomial terms are added to both levels (i.e. both level 1, within individual and level 2, between individual) for both models.

Table 5.11: Equations resulting from fitting quadratic and cubic polynomial models.

| Model Number | Model Equation |
|---|---|
| Model 5 | $y_{ij} = 48.10 + 2.69(Diagnosis\ of\ Diabetes)_i$<br>$- 4.81(Diagnosis\ of\ Anaemia)_i + \zeta_{0i} - 0.22(Time_{ij})$<br>$- 0.42(Diagnosis\ of\ Anaemia)_i * (Time_{ij})$<br>$- 0.24(Diagnosis\ of\ Cardiovascular\ Diseases)_i * (Time_{ij})$<br>$+ \zeta_{1i}Time_{ij} + \zeta_{2i}Time_{(ij)^2}$ |
| Model 6 | $y_{ij} = 48.10 + 2.63(Diagnosis\ of\ Diabetes)_i$<br>$- 4.57(Diagnosis\ of\ Anaemia)_i + \zeta_{0i} - 0.22(Time_{ij})$<br>$- 0.35(Diagnosis\ of\ Anaemia)_i * (Time_{ij})$<br>$- 0.28(Diagnosis\ of\ Cardiovascular\ Diseases)_i * (Time_{ij})$<br>$+ \zeta_{1i}Time_{ij} + \zeta_{2i}(Time_{ij})^2 + \zeta_{3i}(Time_{ij})^3$ |

Where;

$y_{ij}$ is the $j^{\text{th}}$ eGFR measurement for patient $i$,

$Time_{ij}$ is relative to the mean measurement time for the $j^{\text{th}}$ patient and

$\zeta_{0i}, \zeta_{1i}, \zeta_{2i}, \zeta_{3i}$ are the random coefficients for each patient for the intercept, time, quadratic time and cubic time terms respectively.

It can be observed that the quadratic and cubic time terms have no significant effect on level 2 (i.e. the between-patient differences). However, they do have an effect at level 1, which is the within-individual level (i.e. different measurements made for a single individual). The residual variance is reduced from 22.01 to 18.25 by moving from Model 1 to Model 5. This means that 17.08% more of the variation within the measurements for an individual is explained when the time squared term is added in Model 5, compared to just having a linear time component in the random effects. On the other hand, the residual variance in Model 6 is reduced to 16.64, which means that a further 8.8% of the within person variation has been explained by including the cubic time term, compared to the quadratic time term in Model 5. From Model 6, it can further be concluded that, by including both squared and cubic time terms to the within- individual level, Model 6 explains 22% more of the within-individual variation compared to the best linear model (i.e. Model 3). In terms of goodness of fit, by comparing the AIC and BIC values reported in Table 5.12 with those in Table 5.10, it can be seen that Model 6 provides the best of these models for our data, explaining more of both between individual and within individual variation than did the earlier models.

Table 5.12: Goodness of fit statistics for the quadratic and cubic time models

| Model Number | Description | AIC and BIC criteria |
|---|---|---|
| Model 5 | Quadratic in time | AIC = 24320.510, BIC = 24364.154 |
| Model 6 | Cubic in time | AIC = 24250.388, BIC = 24318.971 |

Various linear and polynomial mixed models for eGFR measurements on our CKD patients over time have been computed. Hence, it is observed that, for this data, use of centring the time on the mean time of the observation points for each individual leads to more meaningful and reliable values of the level 1 eGFR intercept (i.e. at the reference mean time value) and helps to take into account multicollinearity. Furthermore, when quadratic and cubic time dependent terms were introduced into the models, a higher proportion of both within- and between- individual variation was explained and a better fitting model was obtained, as observed from the AIC and BIC values. Hence, the cubic model is the best-fitting one of these models for this data.

Semi-parametric and non-parametric modelling approaches are explained in the following chapter (chapter 6) (Davidian and Giltinan, 1993, 1995; Vonesh and Chichilli, 1997; Davidian and Giltinan, 003; Molenberghs and Verbeke, 2005).

# 6 Semi-parametric and Non-parametric Modelling

## 6.1 Introduction

When parametric models such as LMMs and GLMMs are applied on the dataset used in this project, results obtained from such models in chapter 5 showed that these models are parsimonious and efficient when correctly specified. However, if one of the assumptions for these models are violated, then LMMs and GLMMs are restrictive and less powerful against these violations. Results obtained in chapter 5 from the application of LMMs and GLMMs on this dataset suggested that the relationship between eGFR and time is non-linear. Hence, the assumption of linearity between response (i.e. eGFR) and independent variable (i.e. time) in LMMs and GLMMs is violated. Therefore, in this chapter, semi-parametric and non-parametric models are investigated on this dataset in order to take account the non-linear relationship between eGFR and time.

Additive models (AMs) are generalisation of linear models (LMs) where smooth functions are used to model the non-linear relationship between response and independent predictor instead of parametric component (i.e. linear term) for continuous covariates. Generalized additive models (GAMs) are generalisation of additive models by allowing the response to follow some distribution from the exponential family of distributions, relaxing the assumption of normality on the outcome. Additive mixed models (AMM) are another generalisation of additive models by the inclusion of random component to the additive model. In this way, AMMs model both random component and fixed linear component, allow modelling of smooth variation about the linear trend and hence enable the modelling of non-linear relationship between the response and covariate effectively. Generalized additive mixed models (GAMMs) are generalisation of AMMs as like the generalisation of GAMs from AMs.

## 6.2 Generalized Additive Models (GAM)

Generalized additive models (GAM) are developed from additive models (Hastie and Tibshirani, 1986, 1990) in much the same way as Generalized Linear Models (GLMs) are a generalisation of linear models. The main difference between the two (i.e GAMs and GLMs) lies in that the linear predictor component (in a GLM) is replaced by using some smooth function(s) in additive models (i.e GAMs). Therefore, the assumption of a linear relationship between the outcome and the covariates (the independent predictors) no longer holds and the relationship between them is represented by some smooth function(s) instead. These known, smooth, monotonic (i.e. meaning either always increasing or always decreasing) functions are evaluated using observed data and specified using a smoothing spline as a base function which is a basis defining the space of functions of which $f$ or close approximation of $f$ is an element. In GAM models, the outcome can follow any known distribution from the exponential family in a similar way to in GLMs. A GAM model has the general form;

$$g(\mu_i) = X_i^*\theta + f_1(x_{1i}) + f_2(x_{2i}) + f(x_{3i}, x_{4i}) + \cdots \qquad \text{(eq. 6.1)}$$

$\mu_i \equiv E(Y_i)$ and $Y_i \sim \mathit{some\ member\ of\ exponential\ family\ of\ distributions}$

Where;

$Y_i$ is the response (outcome),

$X_i^*$ is a row of the model matrix for any strictly parametric component,

$\theta$ is the corresponding parameter vector,

$f_j$ is the $j^{\text{th}}$ smooth function of the $k^{\text{th}}$ covariate, $x_k$ (Wood, 2006).

A typical additive model contains two parts; a parametric component and a non-parametric component. If the relationship between the response and some covariate is linear, then the dependence on this covariate is modelled in the parametric component of the model. If the relationship between the response and a covariate is not linear, then the dependence on that covariate is modelled in the non-parametric component using some smooth function. Another difference between basic additive models and GAM models is the procedure for estimating smoothing parameters. Whereas parameter estimation for additive models uses penalized least squares techniques (PLS), in GAMs the penalized likelihood maximization method is used instead. In GAMs, the reason of using penalized likelihood maximization over likelihood maximization is because if parameters are estimated by maximum likelihood, estimation of splines will be overfitted due to not taking the wiggliness of the function into account (Wood, 2006).

In additive models, PLS technique is used in parameter estimations, so that the additive model can take the response, the penalty matrix and smoothing parameters to calculate the model matrix. However, there is no simple trick to produce penalized likelihood of GAM. Therefore in GAM, the inclusion of smooth functions creates a challenge in estimation of the smoothing parameter associated with each covariate. Once the base function of the smooth function has been selected, the penalized likelihood maximization method is used. However this poses two problems; estimating the smooth term for the parameter where the smooth term is the constant that controls degrees of smoothness (i.e. controls the weight to be given to make the function smooth) and also estimation of the 'wiggliness' of the function. If the smooth term is equal to zero, this results in un-penalized regression estimate. Whereas, if the smooth term tends to infinity, the regression estimate of spline tends to straight line. In order to overcome these two problems mentioned above, penalized iteratively re-weighted least squares (P-IRLS) in two steps method is used which proceeds as follows:

When the current $\beta$ estimate and the corresponding estimated mean response, $\mu$ is given, at the $k^{\text{th}}$ iteration,

$Var(Y_i) = V(\mu_i^{[k]})\phi$

Where

$Y_i$ is the vector of responses and

$\phi$ is a constant where $\phi = \sigma^2$ if the distribution of the response is Gaussian or $\phi = \frac{1}{v}$ if the response follows a Gamma distribution where $v$ is the degrees of freedom.

Then at the first step, $z_i$ and $w_i$ are calculated where $w_i$ is the weight which will be multiplied by the least error that is $(z - X\beta)$ in the second step.

$$z_i = g'\left(\mu_i^{[k]}\right)\left(y_i - \mu_i^{[k]}\right) + X_i\beta^{[k]}$$

$$w_i = \frac{1}{V(\mu_i^k)(g'(\mu_i^{[k]}))^2}$$

Where

$g$ is the link function such as log link function when the distribution of the response is taken as Gamma distribution,

Mean response, $\mu_i$ is taken as outcome that is used in transforming the response via link function and

$X_i$ is the $i^{th}$ row of the model matrix, $X$.

Then, at the second step, $w_i$ is minimized with respect to $\beta$ in order to estimate the next parameter, $\beta^{[k+1]}$. In this way, the least absolute error is minimized rather than least square error. In order to perform this minimization, the statement below is minimized with respect to $\beta$:

$$\left\|\sqrt{W}\,(z - X\beta)\right\|^2 + \lambda\beta^T S\beta$$

Where

$W$ is the diagonal matrix such that $W_{ii} = w_i$,

$X$ is the model matrix,

$\lambda$ is a constant smooth term controlling the degrees of smoothness and

$S$ is a vector of known coefficients.

For instance if the mean response assumed to follow a Gamma distribution and modelled by using a log link function, then $g'(\mu_i) = \mu_i^{-1}$ and $V(\mu_i) = \mu_i^2$. Hence $w_i$ and $z_i$ are calculated as being $w_i = 1$ and $z_i = \left(\frac{y_i - \mu_i^{[k]}}{\mu_i^{[k]}}\right)^2 + X_i\beta^{[k]}$.

(Wood, 2006).

## Use of Spline Functions

A spline is a function generated from polynomials. A spline function is obtained when, for a specific covariate, the area under the polynomial curve is divided into non-overlapping sections and each section is estimated separately by using various polynomials defined as bases or base functions (de Boor, 1978; Wahba, 1990). At the point where the two adjacent sections meet, a vector with elements can be formed giving the position of knots. Regression modelling using smoothing functions uses penalized regression smoothers, based on splines, to model long term variation as a smooth function and to separate the long term overall systematic between subject variations from random between subject variations.

For each non-linear but smooth term corresponding to the contribution to eq. (6.1) from a single independent predictor $x_i$, an appropriate smooth function $f_i(x_i)$ should be found. This function is found by using a suitable set of basis functions. One possible approach is the use of splines. There are several different types of splines, but they all have the same common feature in that each type can approximate any given smooth function by a sequence of functions, each of which is smoothed; over non-overlapping but connected sub-intervals of the domain of the variable $x_i$. The domain for $x_i$ is sub-divided into intervals in such a way that over any one interval, the function $f_i(x_i)$ is monotonic. The bounding points of these intervals are called knots (or nodes). The same type of base spline must be used throughout for the function associated with any given predictor. However, different types of splines can be used for the functions associated with other predictors. Using the appropriate spline optimisation method, a smooth function for each predictor is obtained leading to an overall smooth function which is as close an approximation as possible from which our data comes. Hence, spline bases provide an attractive option for modelling smooth functions of covariates and to estimate the model parameters. Several different kinds of splines are available, for example the most commonly used splines are cubic regression splines, cyclic cubic regression splines, P-splines and thin-plate regression splines.

### 6.2.1 P-splines for GAMs

Initially, all of the four commonly used spline techniques namely; cubic regression splines, cyclic cubic regression splines, P-splines and thin-plate regression splines are studied to investigate which spline technique is more suitable than others to use on the data for this project. Investigations suggested that cubic regression splines and cyclic cubic regression splines are not appropriate for this data due to the strict constrains applied when estimating the smooth function. These constrains include estimating the parameter by fixing the value of the smooth term to the value obtained at the knot for cubic regression spline whereas requiring to have the same value at both the upper and lower boundaries of the knot in cyclic cubic regression splines. P-splines are based on B-splines. However, the reasons of not considering B-splines over P-splines are the lack of availability of the numerical methodologies to achieve the stability of the B-splines and the restriction of B-splines on the smooth function which allows the basis function to be non-zero only for the specific interval (i.e. $m$ + 3 knots where $m$ is the 1 degree less than the basis function). Hence, p-splines and thin-plate regression splines are further investigated.

Further investigations have found that both P-splines and thin-plate regression splines are favourable than other spline techniques because both of these splines use low-rank base function and therefore difference penalty is employed on the estimation of parameter $\beta$ to control the "wiggleness" of the function. However, in the GAM models that follow, P splines have been chosen as the most appropriate for modelling the smooth functions to represent the variation of the outcome within our data. Thin-plate regression splines overcome many problems that occur in other spline techniques such as selection of knot locations, and as a result, such splines produce knot-free bases and create smoothing functions for multiple predictors by finding the best match between the data and smoothing object. Therefore, even thin-plate regression splines are useful specifically when the data is noisy, since P-splines are based on cubic splines, such splines are easier to create and additionally, P-splines can use any order of penalty as well as any B-spline basis to control the "wiggleness" of the function. Therefore, P-splines are more flexible and stable than thin-plate regression splines when knots are equally spaced and in addition to this, P-splines have lower computational cost than thin-plate regression splines when used for large datasets and hence P-splines are chosen over thin-plate regression splines for this project.

P splines are an improved version of B –splines, as proposed by Eilers and Marx in 1996. (B-splines (deBoor, 1978) were developed for spline interpolation purposes. Both B-splines and P-splines are developed from cubic splines.

In cubic splines, it is required to have a continuous function $f_i(x_i)$ that has a continuous first and second derivatives at each knot in order to estimate the $\beta$ parameters. In this way, spline function can be evaluated as;

$$f(x) = \sum_{i=1}^{k} b_i(x_i)\beta_j(x_i) \qquad \text{(eq. 6.2)}$$

Where;

$b_i(x_i)$ is a standard base spline function,

$k$ is the number of knots

$j$ is the number of parameters (Wood, 2006).

A major advantage of cubic spline is that it has easily understandable parameters, and the base function does not demand any re-scaling condition on the covariates. However, the requirements of selecting the number of and locations of the knots are disadvantage.

B-splines are a development of polynomial (i.e. cubic) splines and are considered to be stable and large scale splines. However to achieve stability, B-splines require advanced numerical methodologies and those currently available provide poor performance on stability. P-splines are an improved development of B-splines which overcome this problem (Eilers & Marx 1996). P-splines also have the added advantages of being low rank smoothers (i.e. using the low-rank base functions) meaning that when the knots are evenly spaced, the difference penalty is employed on parameter $\beta$ to control "wiggliness" of the function. P-splines allow use of any order of penalty plus any B-spline as a basis function to estimate the parameter. Hence P-splines are considered to be flexible when the knots are evenly spaced but are problematic for non-evenly spaced knots (Wood, 2006).

Cyclic cubic regression splines, which are a further development of cubic splines, and thin-plate regression splines, that generate the best function that matches the data and the smoothing object (i.e. the function that we are trying to estimate) are the other two alternative spline choices that can be selected by the researcher. After the theoretical investigations on all of the commonly used spline techniques, P-splines are chosen due to the reasons stated earlier. However, all of the commonly used various spline techniques that are investigated theoretically have also been tested on our data for the practical check and it has been concluded that P-splines offer the best fit for modelling the data points in this research and these are used in the models that follow.

## 6.3 Generalized Additive Mixed Models (GAMM)

In a similar manner to progression from GLM to GLMM models (see section 5.6), if random effects are added to a GAM model, then the model becomes a GAMM model. GAMM models are mixed models characterizing both fixed effects with parametric components and random effects describing by smooth functions of the covariates. Non-normal responses are permitted, since these models are generalized versions of additive mixed models. The assumptions of such models are very similar to those of GLMMs except that the assumption of linearity between the response and the independent predictors is removed. Linearity which is the dependence of outcome variable on the covariates (the independent predictors) is replaced by smooth functions describing the influence of the covariates. Therefore, GAMM models are considered to have great flexibility and allow complex relationships to be modelled. The general form of a GAMM model is;

$$g(y) = X^{*}\theta + \textstyle\sum_{j=1}^{k} f_j(x_j) + Zb + \varepsilon \qquad \text{(eq. 6.3)}$$

Where;

$g(y)$ is a monotonic, differentiable link function of the outcome such as log link function,

$\theta$ is the vector of fixed parameters,

$X^{*}$ is the fixed effects model matrix,

$f_j$ is the smooth function estimated for covariate $x_j$, where this $x_j$ is centred on mean, Z is the random effects model matrix,

$k$ is the number of continuous covariates,

$b$ is the vector of random effects coefficients with unknown positive definite covariance matrix $G_\theta$, so that $b \in N(0, G_\theta)$ and

$\varepsilon$ is the vector of residual error with positive definite covariance matrix R, so that $\varepsilon \in N(0, R)$,

The conditional mean of the response given random effects $b$ is defined to be $\mu^{[b]}$ and is linked to the linear predictor, $\eta$ (i.e. the link function such as log link). Then,

$$g(\mu^{[b]}) = \eta = X^{*}\theta + \sum_{j=1}^{k} f_j(x_j) + Zb \qquad \text{(eq. 6.4)}$$

(Lin and Zhang, 1999; Fahrmeir and Lang, 2001; Wood, 2006).

Since flexible covariance structure can be used for the random effects, $b$, by using this linear predictor, $\eta$ (i.e. linear link function), GAMM models can be applied in different study designs, including hierarchical designs. In GAMM models, if the linear link function is used, the link function transform the conditional mean of outcome into a linear outcome, so that the smooth function is reduce to linear which then creates a linear or polynomial model that can be defined as GLMM model and is considered as a special case of GAMM models.

Instead of taking the value of the covariate as an independent predictor as in GLMMs, a linear predictor in GLMMs is improved in GAMM models, so that a single smooth function is formed from the combination of piecewise functions corresponding to that covariate. For each smooth term, two components are evaluated. One is the fixed effect, calculated by an un-penalized component, and the other is the random effect, computed using a penalized component. The random effect of the smooth component is also assumed to be normally distributed. In estimating each smooth parameter in GAMM models, each is first considered as a variance component of the covariate, then each parameter is estimated by using either restricted estimation of maximum likelihood (REML) or penalized quasi-likelihood (PQL) methods (Wood, 2006).

### 6.3.1 Application of GAMM Models

The GLMM models described previously (chapter 5) are based on the assumption of linearity between the dependent and independent variables (i.e. between eGFR and time respectively). This assumption is now examined to investigate the possibility of non-linearity of the association between the outcome and covariates (e.g. between eGFR and time). Up to this point, we have used eGFR which is calculated using the MDRD formula (see section 2.1.3), which includes age, gender, ethnicity and SCr in its calculation as the dependent variable. Hence, the effects of these covariates (i.e. age, gender, ethnicity, SCr) cannot be fully explored using eGFR as the outcome.

In order to examine the effects of age, gender, ethnicity and SCr in more detail, the MDRD formula for eGFR is inverted to extract the corresponding SCr values for each patient. Extraction of SCr values from eGFR using the MDRD formula allows the use of SCr values (obtained directly from a patient's blood sample) as the dependent variable and allows the effects of age, gender and ethnicity to be studied. In addition to this, eGFR calculated by MDRD formula is not commonly reported in the routine practice for every patient. However, SCr is a measurement obtained directly from the blood of the patient and hence more commonly reported in the routine practice. Therefore, every patient might not have a calculated eGFR value but can have SCr measurement. Since GAMM models require a large amount of data points and desire the data to be balance, using SCr as an outcome instead of eGFR in such models can provide better models and this is investigated in this section of the thesis.

The dependent variable now becomes the SCr value, which is transformed by taking the natural logarithm to achieve a normally distributed dependent variable. GAMM models are then applied to the log (SCr) values using the P-splines approach described in section 6.3.1. In order to overcome the disadvantages of P-splines due to needing to decide on the number of knots, the number of knots is not specified when defining the model, so that the number of knots can be chosen automatically by the software package based on the data available for that covariate.

Throughout this section, the GAMM models presented involve the application of smoothing splines to model SCr readings against continuous covariates including time, age, age at diagnosis of CKD, systolic blood pressure value and the interactions of the co-morbidities with time.

Diagnostic testing of the models is achieved using graphical analysis, Quantile Quantile plots (QQ-plots) and histograms to assess normality, 'residuals versus predictor plot' and 'residuals versus fitted value plots' to examine homogeneity, and 'fitted values versus observed values plots' to check the quality of the model fit (e.g. Figure 6.12).

Before describing the models applied on log (SCr) values, a GAMM model is first fitted using eGFR values as an outcome to compare with the GLMM model described in chapter 5.

#### 6.3.1.1 Comparing a GAMM Model using eGFR values as an outcome with GLMM Model

LMM models described in chapter 5 are parametric models which are fully identifiable in finitie dimensitional parameter space. Parameters are estimated and interpreted easily. However, estimation of parameters are only accurate if the assumptions of the parametric models are met by the data such as normality assumption. In such cases, parametric models are more powerful than semi-parametric or nonparametric models.

Even, restrictive assumptions of the parametric models are violated in nonparametric models, parameters in nonparameteric models are in infinite dimensional space and therefore interpretation of the parameters are difficult and estimation can be inaccurate if the model contains large number of independent variables.

Semiparametric models combine components of parametric and nonparametric models where the parameters of interest are in finite dimensional space and nuisance parameters (e.g. mean and variance) are in infinite dimentional space. In semi-parametric models, parameters of interest are interpreted easily due to parametric component and models are flexible due to nonparametric component.

Both GLMM and GAMM models are semi-parametric models. However, GLMM models are more towards the parametric models and GAMM models are more towards the nonparametric models. This is because, even the normality assumption is violated in GLMM models, the known distribution from the exponential family is assumed for the outcome and hence such models are identifiable. In addition to this, linearity between the outcome and independent variables is still assumed in GLMM models. However, in GAMM models, the outcome is not assumed from a known distribution. Hence the models are not identifiable and linearity assumption between outcome and independent variables is also violated. The distribution of the outcome is identified by the data point using the smoothing spline functions. However, GAMM models require large amount of data points and desire the data to be balance.

Firstly, when the GAMM model is fitted using the eGFR values as outcome, it is observed that same data points used to fit the GLMM models in chapter 5 are not enough to fit the interaction effects of eGFR with time and other covariates in GAMM model. Therefore, only the single effects are considered in this model. By using the same single effect parameters employed in the best GLMM model in chapter 5, a GAMM model is carried out and only the diagnosis of anaemia and diabetes are found to be the statistically significant (both with $p < 0.001$) single effect parameters affecting the outcome. Time is taken as a smooth term to model the progression of CKD by looking at how eGFR changes over time and since the smoothing parameter is estimated as 6.459 with $p < 0.001$, it can be concluded that the decline of eGFR over time is nonlinear. When the diagnostic plots are investigated for this model, it is concluded that there is no evidence of serious model fitting problems. However, since this model only explains 4.85% of the total variation in the outcome, it is found to be less powerful than GLMM model when the outcome is used eGFR values. In the next sections, the outcome variable is changed and assessed as SCr values to obtain better fitting models and investigating alternative variables affecting the outcome to explain the variation in the SCr and hence in the eGFR.

#### 6.3.1.2 GAMM Model 1 – Modelling SCr as a smooth function of time

The research question now lies in investigating changes in SCr readings over time and what function affect such changes. The first GAMM model (Model 1, eq. 6.5) is a basic model which models changes in SCr as a smooth function of time, where the subject (i.e. the patient) is the random component. The additive mixed model framework is defined as;

$$y_t = \alpha + f_i(time_t, \lambda_1) + N_t \qquad \text{(eq. 6.5)}$$

Where;

$y_t$ is the $\log(SCr)$ of a given patient at time measurement $t$, where log is the natural logarithm,

$\alpha$ is the intercept,

$time_t$ is the time of the t<sup>th</sup> measurement after the first diagnosis of CKD, in years,

$N_t$ is the identically, independent, normally distributed noise (error) term.

The function $f_i(time_t, \lambda_1)$ is the smoothing curve for time and can have any smooth shape,

$\lambda_1$ is the smoothing parameter for the time, if $\lambda_1 = 1$, then this means that there is a linear relationship between time and SCr, which is the assumption made in GLMM models.

In the results in all models below in this section, *s* represents an additive spline contribution to log (SCr).

**Formulation**

**of Model 1 in R Software**

Model1 <- *gamm4(y~s(Time, bs="ps"), random=~(1|ID), data=CKDdata, method="REML")*

(Wood and Scheipl, 2015) (Wood, 2015).

Table 6.1: Output of Model 1

```
Family: gaussian
Link function: identity

Formula:
y ~ s(Timeyears, bs = "ps")

Parametric coefficients:
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) 2.068248   0.004263   485.2   <2e-16 ***

Signif. codes:
'***': 0 < p-value ≤ 0.001,
'**': 0.001 < p-value ≤ 0.01,
'*':0.01< p-value ≤ 0.05,
'.': 0.05 < p-value ≤ 0.1,
' ': 0.1< p-value ≤ 1

Approximate significance of smooth terms:
                   edf     Ref.df    F       p-value
s(Timeyears)   6.164   6.164    24.33  <2e-16 ***

Signif. codes:
'***': 0 < p-value ≤ 0.001,
'**': 0.001 < p-value ≤ 0.01,
'*':0.01< p-value ≤ 0.05,
'.': 0.05 < p-value ≤ 0.1,
' ': 0.1< p-value ≤ 1

R-sq.(adj) = -0.00176  lmer.REML score = -10440  Scale est. = 0.0024038  n = 3776
```

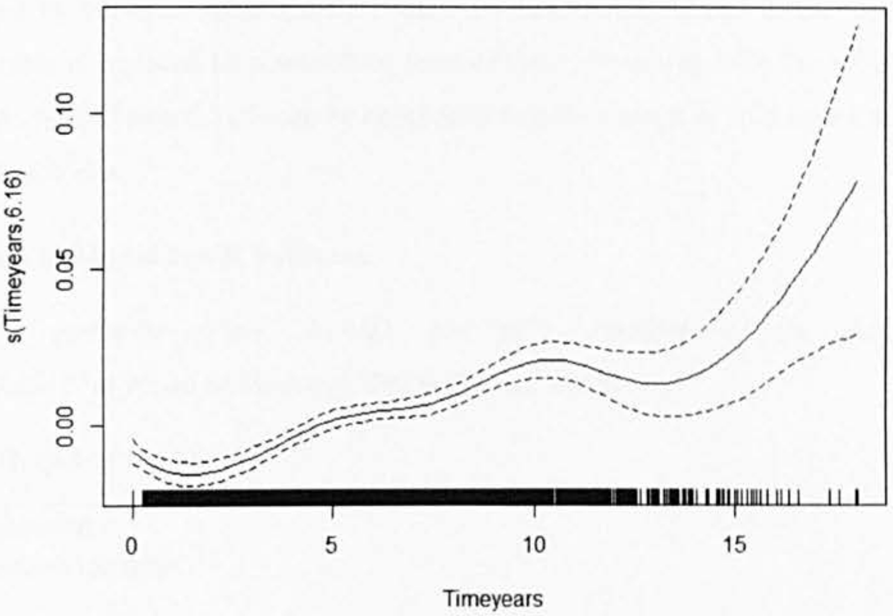

Figure 6.1: The smoothed term of time from Model 1

The "tick marks" in Figure 6.1 on the horizontal axis represent the actual measurement times and the dotted curves represents the limits on 95% Confidence Interval. Vertical line shows the spline function of time.

The further $\lambda$ is away from 1, the greater the non-linearity with respect to the covariate presented. In GAMM models, $\lambda$ is computed by finding the value which gives the lowest AIC. The resulting $\lambda$ value is called the expected degrees of freedom (edf). The results of Model 1 (Table 6.1) show the edf for the time variable to be 6.164, and both Figure 6.1 and output from Model 1 (Table 6.1) show that the patient's SCr has a highly significant nonlinear upward trend with associated p-value of 0.000 (3dp), and hence it can be concluded that eGFR decreases over time (note: as known from the MDRD formula that there is an inverse relationship between eGFR and SCr – as one increases, the other decreases). Figure 6.1 also indicates that SCr changes non-linearly over time which in turn leads to the conclusion that change in eGFR, and hence the progression of CKD, over time is non-linear in nature.

### 6.3.1.3 GAMM Model 2 – Treating measurements from different subjects separately

The next stage of GAMM analysis asks; 'can the non-linear changes in SCr over time be explained by between subject differences'. In this model, Model 2, the smoothed time-dependent term is replaced by a smoothed term of time interacting with the subject. From the output of Model 2 (Table 6.2), it can be concluded that the change of SCr over time is different for different subjects.

**Formulation of Model 2 in R Software**

Model2 <- *gamm4(y~s(Time, by=ID, bs="ps"), random=~(1|ID), data=CKDdata, method="REML")* (Wood and Scheipl, 2015) (Wood, 2015).

Table 6.2: Output of Model 2

```
Family: gaussian
Link function: identity

Formula:
y ~ s(Timeyears, by = ID, bs = "ps")

Parametric coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  2.05874   0.00724     284.4    <2e-16 ***

Signif. codes:
'***': 0 < p-value ≤ 0.001,
'**': 0.001 < p-value ≤ 0.01,
'*':0.01< p-value ≤ 0.05,
'.': 0.05 < p-value ≤ 0.1,
' ': 0.1< p-value ≤ 1

Approximate significance of smooth terms:
                   Edf     Ref.df   F      p-value
s(Timeyears):ID    5.277   5.277    20.15  <2e-16 ***

Signif. codes:

'***': 0 < p-value ≤ 0.001,
```

'**': 0.001 < p-value ≤ 0.01,

'*':0.01< p-value ≤ 0.05,

'.': 0.05 < p-value ≤ 0.1,

' ': 0.1< p-value ≤ 1

**R-sq.(adj)** = 0.00468 **lmer.REML score** = -10349 **Scale est.** = 0.002438 n = 3776

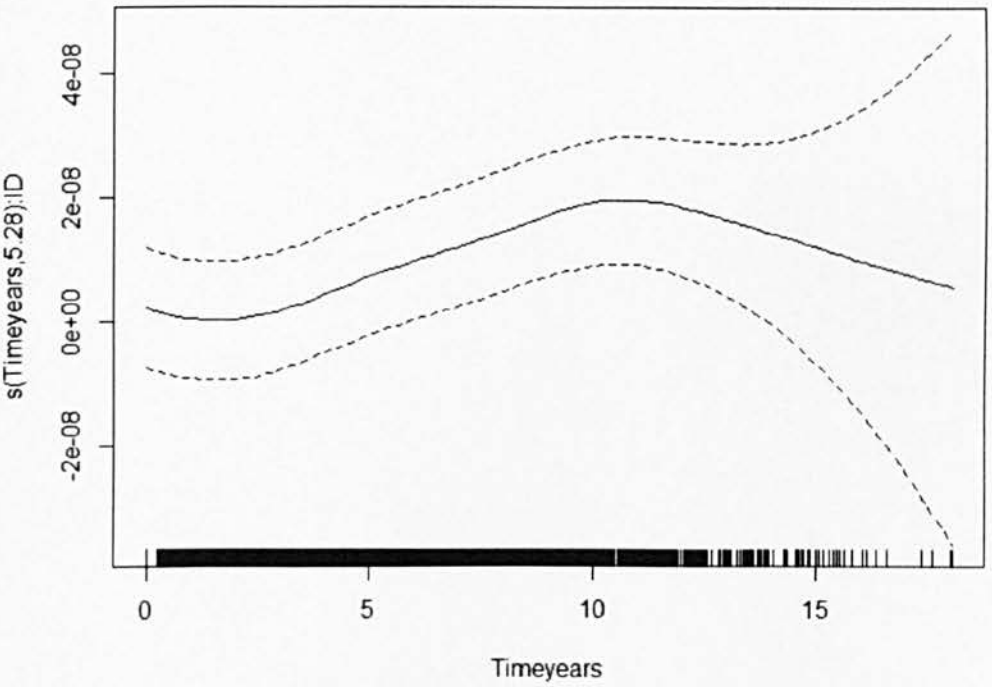

Figure 6.2: Smoothed term of time interaction with the patient from Model 2

Figure 6.2 shows that a very slight decrease is even observed in SCr values for around a year after diagnosis, suggesting a slight improvement in CKD status. This may be explained as a consequence of the effects of medication prescribed to the patient at time of diagnosis providing a minor improvement initially, resulting in a very small temporary increase in eGFR. However, since CKD status is known to generally worsen over time, eGFR will progressively decrease, and SCr increase over time, accordingly. The trend over the subsequent 10 years (from diagnosis) is described as a gradual increase in SCr (i.e decrease in eGFR). Beyond 10 years after diagnosis of CKD, the confidence intervals widen markedly and so the trend indicated on the graph in Figure 6.2 is presumed to be non-significant after 10 years.

#### 6.3.1.4 GAMM Model 3 – Including patient specific factors

In Model 3, additional covariates such as age, age at diagnosis of CKD, gender and stage of CKD at diagnosis are introduced into the model. When the influence of each co-morbidity on rapid decline in CKD status was investigated (see section 4.2.1.2), it was found that anaemia was the most influential single factor. Hence, in Model 3, the smooth term of time is replaced with the smoothed function of the interaction of time with the diagnosis of anaemia.

**Formulation of Model 3 in R Software**

Model3 <- *gamm4(y~s(Time, by=AnaemiaDiagnosis, bs="ps")+ s(Age, bs="ps")+ s(Diagnosticage, bs="ps")+ Gender + Diagnosticstage, random=~(1|ID), data=CKDdata, method="REML")* (Wood and Scheipl, 2015) (Wood, 2015).

Table 6.3: Output of Model 3

```
Family: gaussian
Link function: identity

Formula:
y ~ s(Timeyears, by = AnaemiaDiagnosis, bs = "ps") + s(Ageyears,
   bs = "ps") + s(Diagnosticage, bs = "ps") + Gender + Diagnosticstage

Parametric coefficients:
                            Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                 1.981976  0.003890   509.49   <2e-16 ***
GenderMale                  0.101625  0.004834   21.02    <2e-16 ***
DiagnosticstageStage 3b     0.066880  0.005028   13.30    <2e-16 ***
DiagnosticstageStage 4      0.155996  0.008669   18.00    <2e-16 ***
DiagnosticstageStage 5      0.241947  0.018369   13.17    <2e-16 ***
```

Signif. codes:
‘***’: $0 < \text{p-value} \leq 0.001$,
‘**’: $0.001 < \text{p-value} \leq 0.01$,
‘*’:$0.01 < \text{p-value} \leq 0.05$,
‘.’: $0.05 < \text{p-value} \leq 0.1$,
‘ ’: $0.1 < \text{p-value} \leq 1$

It can be seen from the parametric coefficients that males tend to have higher SCr than females, and more serious the stage of CKD, the higher the SCr (and hence lower eGFR).

**Approximate significance of smooth terms:**

| | edf | Ref.df | F | p-value | |
|---|---|---|---|---|---|
| s(Timeyears):AnaemiaDiagnosisDo not have the disease | 5.964 | 5.964 | 3.035 | 0.005954 | ** |
| s(Timeyears):AnaemiaDiagnosisHaving the disease | 3.635 | 3.635 | 5.595 | 0.000357 | *** |
| s(Ageyears) | 6.285 | 6.285 | 7.028 | 1.25e-07 | *** |
| s(Diagnosticage) | 6.357 | 6.357 | 4.002 | 0.000422 | *** |

Signif. codes:
‘***’: $0 < \text{p-value} \leq 0.001$,
‘**’: $0.001 < \text{p-value} \leq 0.01$,
‘*’:$0.01 < \text{p-value} \leq 0.05$,
‘.’: $0.05 < \text{p-value} \leq 0.1$,
‘ ’: $0.1 < \text{p-value} \leq 1$

**R-sq.(adj)** = 0.572 **lmer.REML score** = -11020 **Scale est.** = 0.0023384 n = 3776

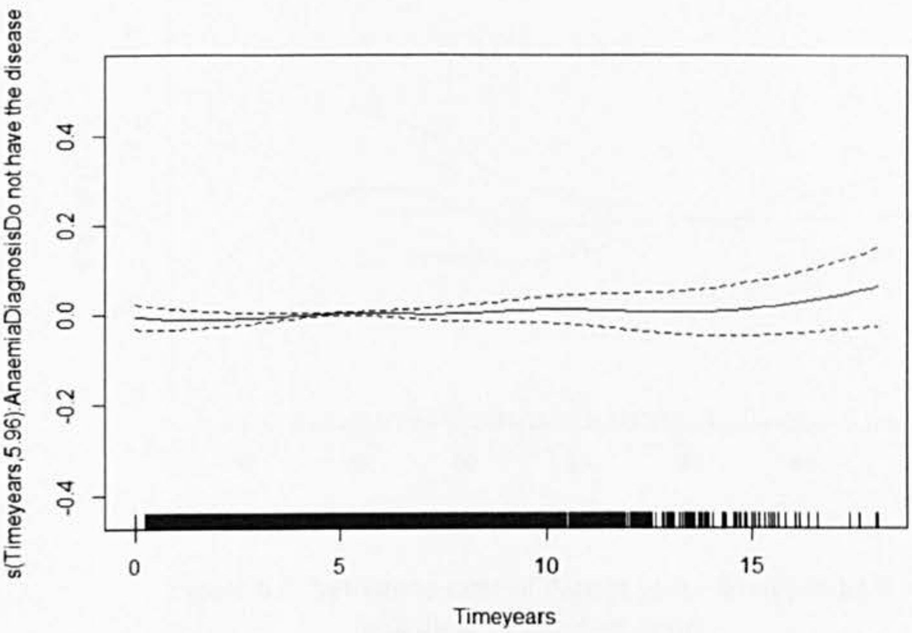

Figure 6.3: Smoothed term of interaction of time with Anaemia Diagnosis, representing Non-Anaemic Patients, from Model 3

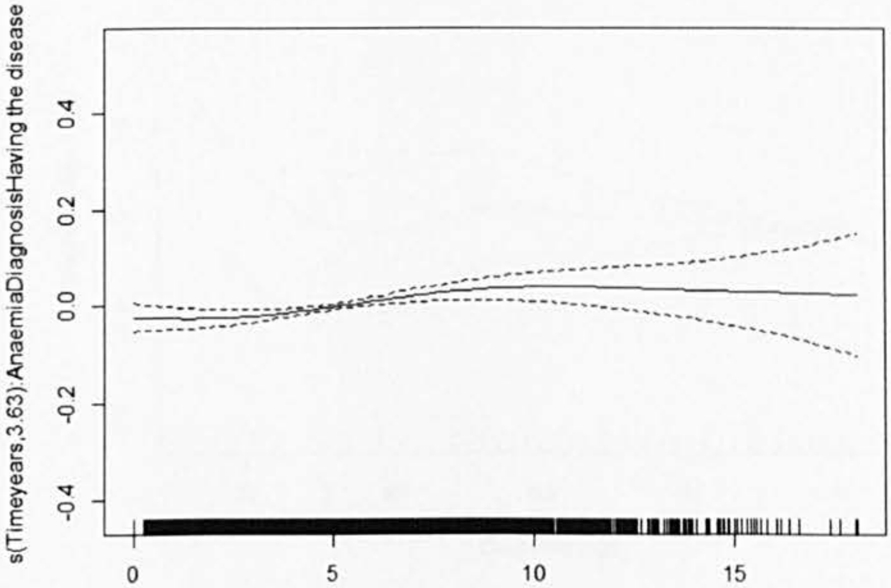

Figure 6.4: Smoothed term of interaction of time with Anaemia Diagnosis, representing Anaemic Patients, from Model 3

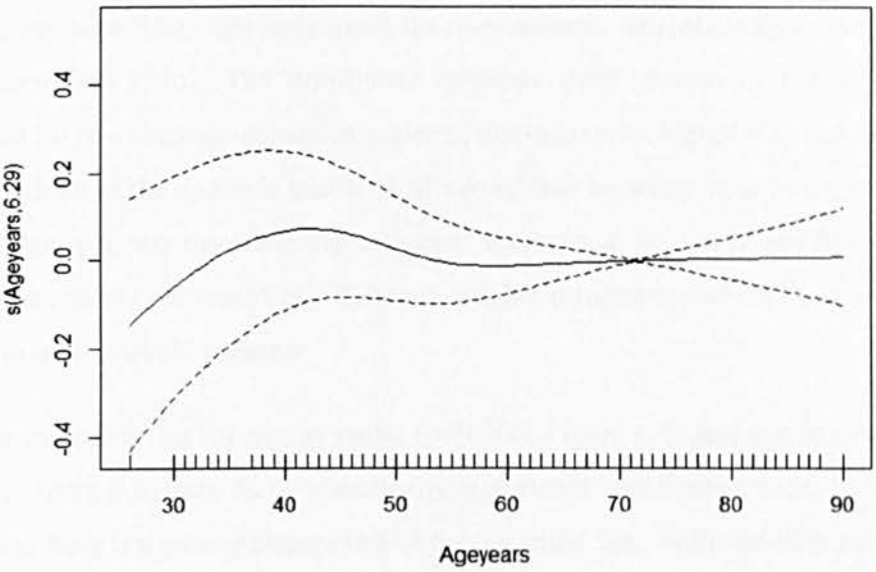

Figure 6.5: Smoothed term of Age in years, from Model 3 regardless of anaemia status

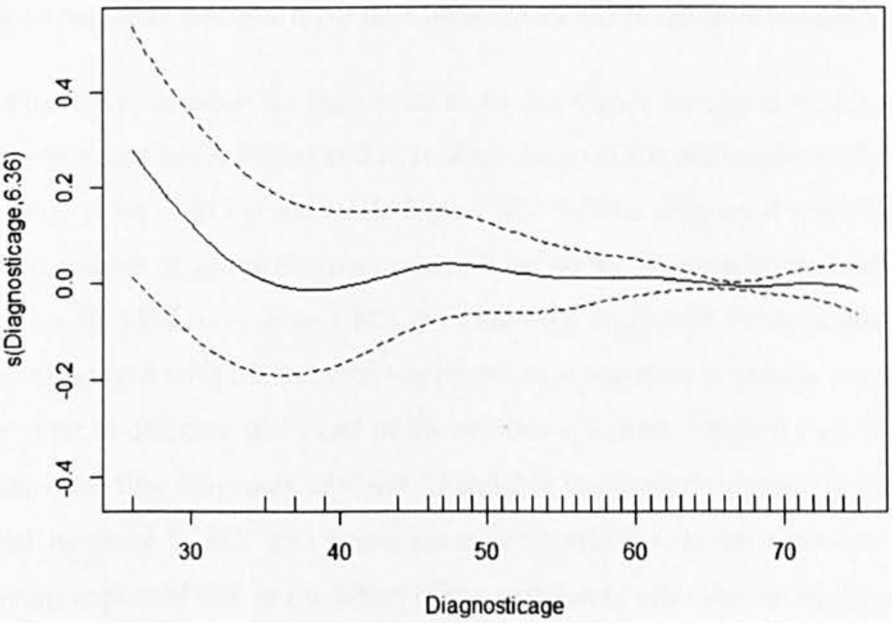

Figure 6.6: Smoothed term of age at the diagnosis of CKD, from Model 3 regardless of anaemia status

In Table 6.3, we see a highly statistically significant upward trend in the change in SCr over time for anaemic patients, with associated edf=3.635 and p-value<0.001 and from Figure 6.3 a small but significant upwards trend for non-anaemic patients with associated edf=5.964 and p-value=0.006 (3dp). The non-linear upwards trend shown in Figure 6.1 is further decomposed for two separate groups of patients, non-anaemic (Figure 6.3) and anaemic (Figure 6.4). Comparison of the upwards trends of SCr over time between these two groups of patients, shows anaemic patients experiencing a sharper increase in SCr over the first 10 years. This translates to a sharper decline of eGFR, hence quicker progression of CKD, in anaemic patients compared to non-anaemic patients.

The smooth terms for age in years, (p=0.000, Figure 6.5) and age at diagnosis of CKD ($p < 0.001$, Figure 6.6) both show statistically significant non-linear trends. In Figure 6.5, it is observed that there is a greater change in SCr for the under 50s, while the effect of age on change in SCr appears to lessen in patients over the age of 50. For patients aged 50 or over (and similarly for patients aged over 50 at the diagnosis of CKD), SCr changes little with increased age (or diagnostic age). However, it should be noted that the MDRD formula used to calculate eGFR from SCr does take the patient's age into account, and hence this effect of little change in SCr with age does not mean that eGFR (or the kidney function) is stable as the patient ages.

In Figure 6.6, between the ages of 20 to 40, the higher the age at the diagnosis of CKD, the lower the SCr, and hence higher eGFR, is observed. eGFR is not routinely tested for younger people, so only those with exceptionally higher SCr will be diagnosed with CKD at younger age. This low number of young diagnoses contributed to the large confidence intervals between the ages 20 to 40. However, since CKD is commonly diagnosed between ages 40-50, when patients are diagnosed with CKD, certain appropriate medication is usually prescribed to those patients in order to decrease the speed of the decline of kidney function over time. In the first instance, after the first diagnosis of CKD, if suitable medication is given to those patients, a slight initial decrease in SCr and hence increase in eGFR may be observed due to kidney function being improved due to the effect of the treatments after the initial diagnosis of CKD. The effect of this slight initial increase in eGFR may explain the slight decrease in SCr in Figure 6.6 between the ages at diagnosis 40-50. Since generally the diagnosis of CKD occurs between ages 35-50, the higher the age at the diagnosis within this range, the higher the SCr and hence the lower the eGFR observed.

#### 6.3.1.5 GAMM Model 4 – including further patient-specific factors

The model is expanded further to include; age at each repeated measure, gender, age at diagnosis of CKD, diagnosis of diabetes, diagnosis of cardiovascular diseases, stage of CKD at diagnosis and mean systolic blood pressure, along with the interaction between time and anaemia factored by two categories of diagnosis of anaemia (i.e. one for non-anaemic patients and one for anaemic patients).

**Formulation of Model 4 in R Software**

```
Model4 <- gamm4(y~s(Time, by=AnaemiaDiagnosis, bs="ps") + s(Age, bs="ps")+ Gender+
Diagnosticstage + DiabetesDiagnosis+ CardiovascularDisease + s(SystolicBP, bs="ps")+
s(Diagnosticage, bs="ps"), random=~(1|ID), data=CKDData)
```

(Wood and Scheipl, 2015) (Wood, 2015).

Table 6.4: Output for Model 4

**Family:** gaussian
**Link function:** identity

**Formula:**

y ~ s(Timeyears, by = AnaemiaDiagnosis, bs = "ps") + s(Ageyears, bs = "ps") + Gender + Diagnosticstage + DiabetesDiagnosis + CardiovascularDisease + s(SystolicBP, bs = "ps") + s(Diagnosticage, bs = "ps")

**Parametric coefficients:**

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.996169 | 0.004840 | 412.463 | < 2e-16 *** |
| GenderMale | 0.103190 | 0.004778 | 21.598 | < 2e-16 *** |
| DiagnosticstageStage 3b | 0.065077 | 0.004887 | 13.317 | < 2e-16 *** |
| DiagnosticstageStage 4 | 0.154962 | 0.008434 | 18.374 | < 2e-16 *** |
| DiagnosticstageStage 5 | 0.235989 | 0.017836 | 13.231 | < 2e-16 *** |
| DiabetesDiagnosisDiabetes diagnosed | -0.016752 | 0.004590 | -3.650 | 0.000266 *** |
| CardiovascularDiseaseDisease present | -0.011831 | 0.004710 | -2.512 | 0.012054 * |

Signif. codes:
'***': 0 < p-value ≤ 0.001,
'**': 0.001 < p-value ≤ 0.01,
'*':0.01< p-value ≤ 0.05,
'.': 0.05 < p-value ≤ 0.1,
' ': 0.1< p-value ≤ 1

**Approximate significance of smooth terms:**

| | edf | Ref.df | F | p-value |
|---|---|---|---|---|
| s(Timeyears):AnaemiaDiagnosisDo not have the disease | 1.000 | 1.000 | 0.416 | 0.518817 |
| s(Timeyears):AnaemiaDiagnosisHaving the disease | 3.612 | 3.612 | 5.577 | 0.000381 *** |
| s(Ageyears) | 6.274 | 6.274 | 6.841 | 2.13e-07 *** |
| s(SystolicBP) | 3.091 | 3.091 | 4.752 | 0.002404 ** |
| s(Diagnosticage) | 6.296 | 6.296 | 4.105 | 0.000336 *** |

Signif. codes:
'***': 0 < p-value ≤ 0.001,
'**': 0.001 < p-value ≤ 0.01,

'*':0.01< p-value ≤ 0.05,
'.': 0.05 < p-value ≤ 0.1,
' ': 0.1< p-value ≤ 1

**R-sq.(adj)** = 0.589 **lmer.REML score** = -10976 **Scale est.** = 0.0023556 n = 3760

The results from Model 4 (see Table 6.4) shows that all of the predictors in the parametric component included in the model are statistically significant. The smoothed term representing the interaction between time and diagnosis of anaemia indicates that, for non-anaemic patients, SCr does not change significantly over time (i.e. change in SCr over time is linear since edf=1.00). However, for anaemic patients, the smoothed term of interaction of time with anaemia is statistically significant (edf= 3.612, p < 0.001), and SCr displays a significant non-linear upwards trend over time. In real terms, this indicates that while there is no significant change in eGFR over time in non-anaemic patients; eGFR decreases significantly and non-linearly over time in patients who have been diagnosed with anaemia.

The most significant contributor to non-linear change in SCr over time (i.e. progression of CKD) is age with associated p-value of $2\times10^{-7}$. This is followed by the nonlinear effect for the smoothed term due to the patient's age at diagnosis of CKD (p=0.00033). A further, albeit the least, statistically significant effect is that of the mean systolic blood pressure reading (p=0.0024).

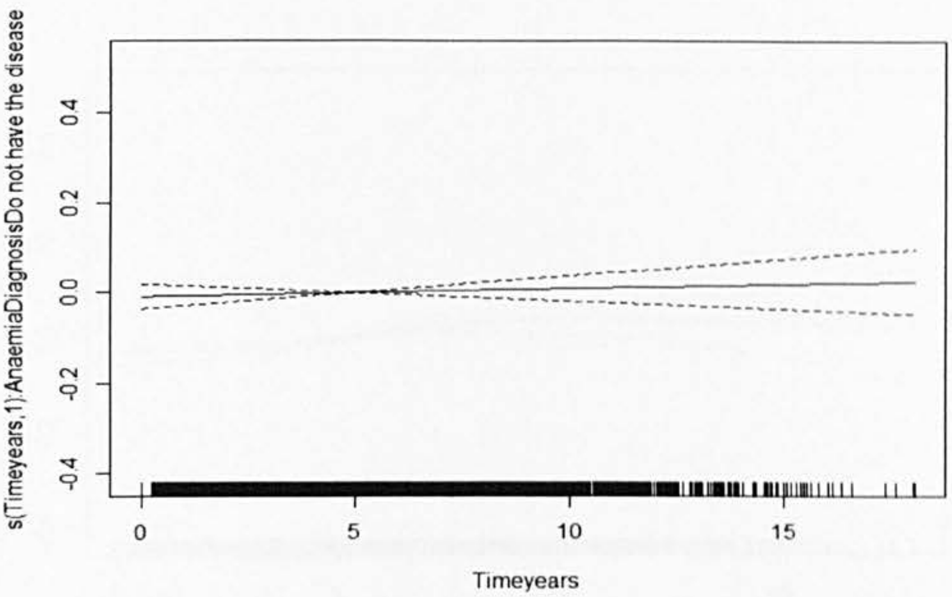

Figure 6.7: Smoothed term of interaction of time with Anaemia Diagnosis for Non-Anaemic Patients, from Model 4

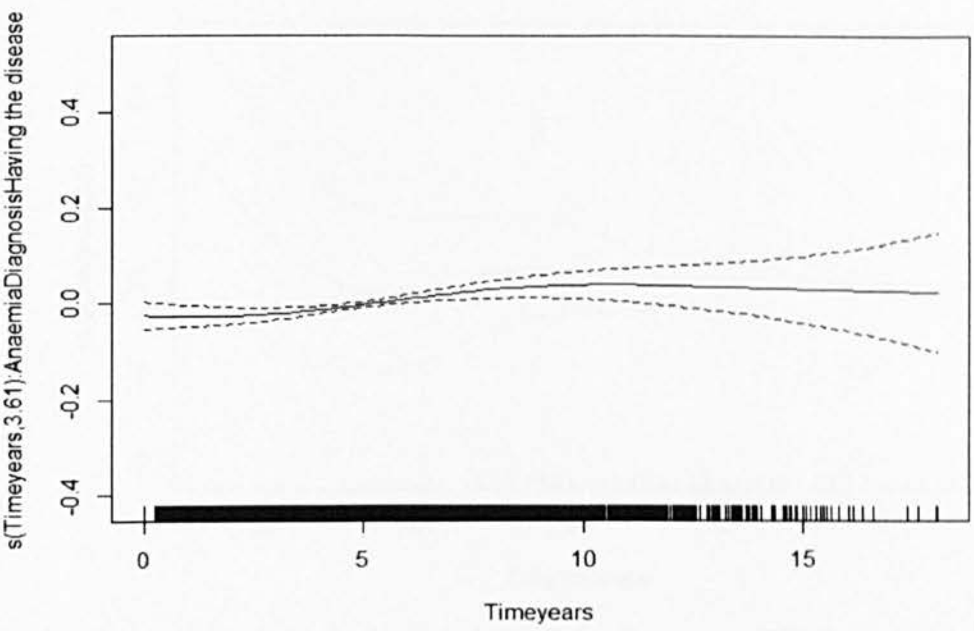

Figure 6.8: Smoothed term of interaction of time with Anaemia Diagnosis, for Anaemic Patients, from Model 4

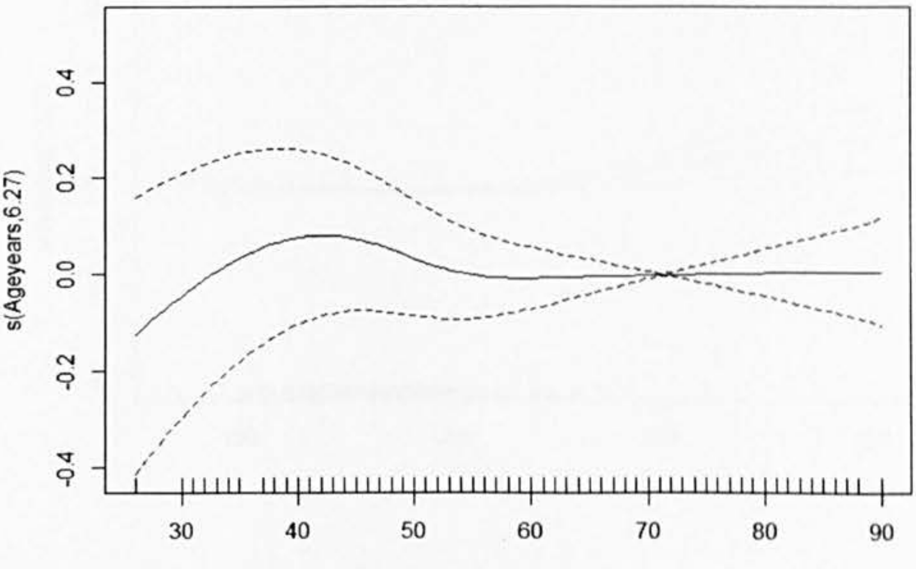

Figure 6.9: Smoothed term of age, from Model 4

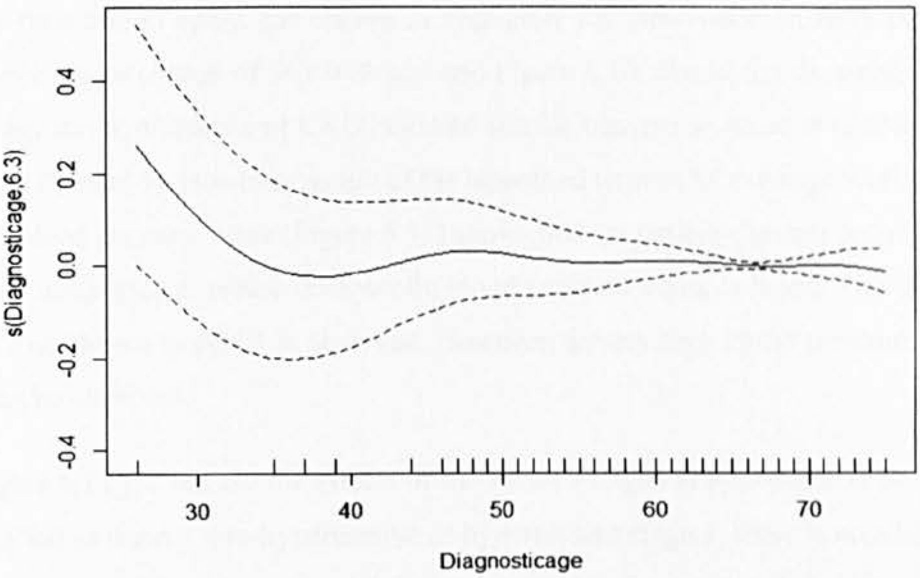

Figure 6.10: Smoothed term of age at the diagnosis of CKD, from Model 4

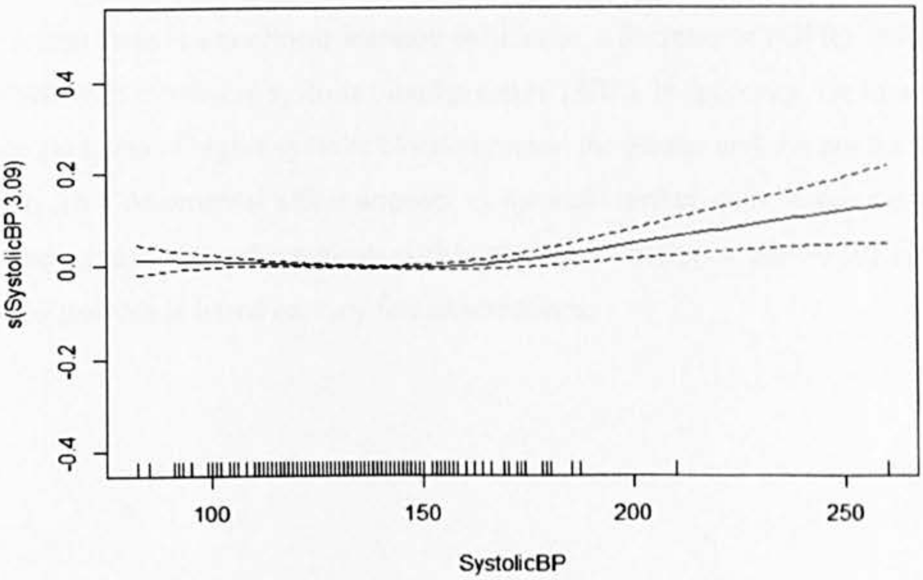

Figure 6.11: Smoothed term of mean systolic blood pressure value, from Model 4

Figure 6.7 shows that for non-anaemic patients, the general trend of SCr over time is linear and upwards, but it is not significant for the first 10 years, suggesting that while eGFR declines over time due to aging, the change is negligible for these patients. Both Figure 6.9, representing non-linear change of SCr with age and Figure 6.10, displaying the dependence on the patient's age at the diagnosis of CKD, showed similar patterns to those obtained from the previous model (Model 3). However, graph of the smoothed term of SCr change with respect to mean systolic blood pressure value (Figure 6.11) shows that for patients having normal systolic blood pressure readings, i.e. where the systolic blood pressure value is below 120, almost no change in SCr and hence in eGFR is observed. However, at very high blood pressure values, a rise in SCr can be observed.

In Figure 6.11 we can see the effects of the various stages of hypertension on SCr. For patients classified as normal, pre-hypertensive or hypertensive stage 1, there is no added effect on SCr due to hypertensive status. A patient is classified to be normal if the systolic blood pressure is between 110 and 120, as being pre-hypertensive by having a systolic blood pressure value between 120 and 139, and classified as being hypertensive at stage 1 by having a systolic blood pressure reading between 140 and 159. For patients with more severe hypertensive status (i.e. stage 2 or crisis) there is a nonlinear increase in SCr (i.e. a decrease in eGFR), indicating a worsening of CKD with increasing systolic blood pressure (SBP). In summary, the more severe the hypertension (in terms of higher systolic blood pressure) the greater and sharper the increase in observed SCr. This detrimental effect appears to increase further with increasing systolic blood pressure, being most severe for patients with hypertensive crisis (i.e. SBP>180). However, it should be noted that this is based on very few observations.

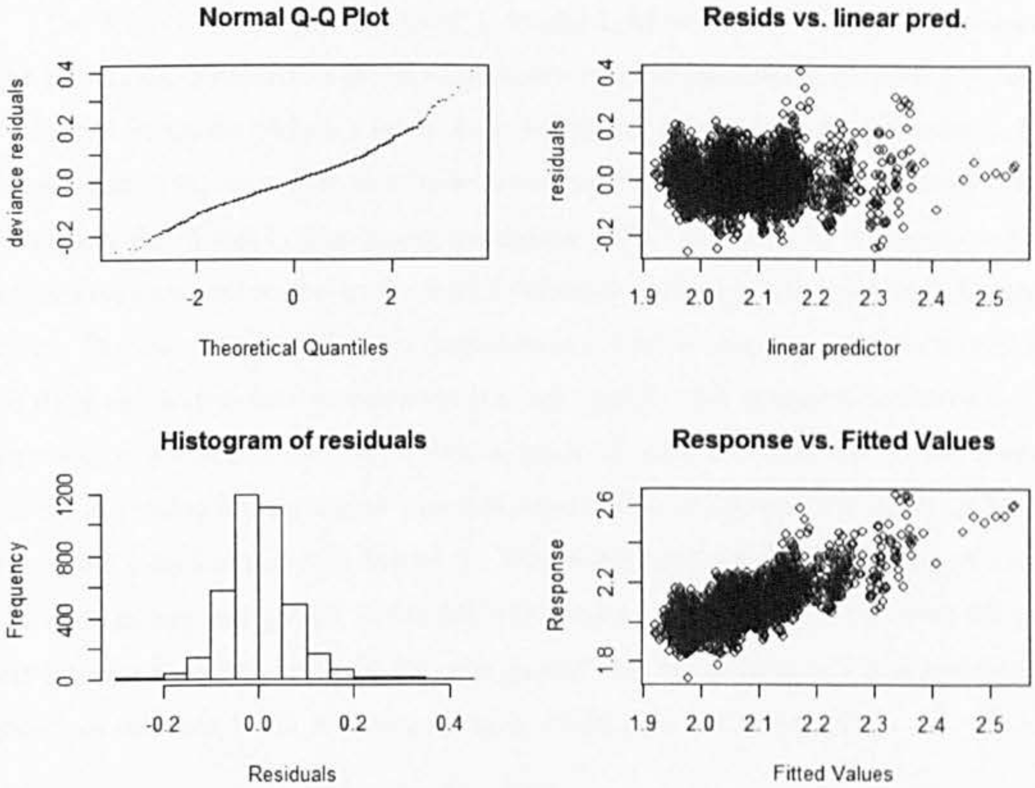

Figure 6.12: Model validation graphs for Model 4

Figure 6.12 shows the model validation graphs for the GAMM model (Model 4) that includes the predictors; age, interaction of time with diagnosis of anaemia, age at the diagnosis of CKD, stage of CKD at the diagnosis of the disease, diagnosis of diabetes, diagnosis of CVD, mean systolic blood pressure and gender. The histogram of the residuals shows that residuals are approximately normally distributed. Furthermore, the QQ-plot shows just a little deviation from being a straight line between the theoretical quartiles. Therefore, it can be concluded from all four graphs in Figure 6.12 there is no evidence of serious model fitting problems.

### 6.3.1.6 Evaluation of GAMM models

The four GAMM models (Model 1, Model 2, Model 3 and Model 4) presented in this section (6.3.1) are compared in terms of goodness of fit to the data by observing REML scores and adjusted R-square values (Table 6.5). From the adjusted R-square values, it can be concluded that, if the smooth term of time is used as the only predictor in the model, the model provides poor fit (Model 1). Explaining the change of SCr over time by taking account that this differs between patients improves the model (Model 2), although a low adjusted-R-value is still observed. The inclusion of additional predictors is needed to explain more of the total variation of the response, and so such information (i.e. age, gender, age at diagnosis, and stage of CKD at diagnosis) is added into the model. The inclusion of these contributions greatly improves the model fit, increasing the explained variation from 0.47% of the total variation (in Model 2) to 57.2% of the total variation (in Model 3). This would also confirm the value of considering factors such as age and gender in the MDRD formula to calculate eGFR from SCr. A small further increase in model fit (to 58.9%) was gained with the addition of covariates representing diagnoses of diabetes, CVD, and mean systolic blood pressure (Model 4).

Table 6.5: Model fit statistics for GAMM models

| Model Number | Model Formula | Adjusted R-square | REML Score |
|---|---|---|---|
| Model 1 | y ~ s(Timeyears, bs = "ps") | -0.00176 | -10440 |
| Model 2 | y ~ s(Timeyears, by = ID, bs = "ps") | 0.00468 | -10349 |
| Model 3 | y ~ s(Timeyears, by = AnaemiaDiagnosis, bs = "ps") + s(Ageyears, bs = "ps") + Gender + Diagnosticstage | 0.572 | -11020 |

| | | | |
|---|---|---|---|
| Model 4 | y ~ s(Timeyears, by = AnaemiaDiagnosis, bs = "ps") + s(Ageyears, bs = "ps") + Gender + Diagnosticstage + DiabetesDiagnosis + CardiovascularDisease + s(SystolicBP, bs = "ps") + s(Diagnosticage, bs = "ps") | 0.589 | -10976 |

## 6.4 A Bayesian Alternative Approach to GAMM Models

In this section, a Bayesian perspective of generalized regression models is considered. Here, inference is achieved through a mixed model representation using penalised likelihood to estimate regression parameters of the model. These models provide an alternative Bayesian perspective of GAMM models, in contrast to the frequentist perspective detailed in sections 6.2-6.3.

Generally, it is known that in the BayesX methodology (Belitz *et al.*, 2012), there are three main approaches for the estimation of regression parameters. These are: full Bayesian inference by using Markov Chain Monte Carlo (MCMC) techniques (Belitz *et al.*, 2012), which is used to formulate generalized regression models from a purely Bayesian perspective; mixed model representation inference (Belitz *et al.*, 2012) which is used to formulate generalized regression models from a Bayesian perspective in relation to mixed models (i.e. an empirical Bayes estimation) and penalized likelihood inference (Belitz *et al.*, 2012) which is used to formulate generalized regression models from a frequentist perspective (i.e. a Non Bayesian estimation, such as stepwise regression). Here, the interest is in finding a Bayesian alternative to the GAMM model. As this is a mixed model, we want to compare the differences between the empirical Bayes model and the full Bayes model. The package 'R2BayesX' is an R interface that is used to estimate Structured Additive Regression (STAR) models (Belitz *et al.*, 2012), which include generalized additive mixed models (Lin & Zhang, 1999), using the BayesX methodology.

When a mixed model methodology is applied, it enables the estimation process for STAR models to be carried out and benefits from the link penalisation of the likelihood and the corresponding random effects distributions. The resulting smoothed terms of the estimated smooth parameters can then be assumed to represent the variance components of the mixed model. For a Gaussian response, variance components are (most commonly) estimated using REML. For non-Gaussian response, the marginal likelihood technique is generally used. In an empirical Bayes approach, regression parameters are estimated using penalised likelihood techniques, where an empirical Bayes posterior distribution is generated and the parameter estimates are classified as penalized likelihood estimates (Shen, 2011).

In a fully Bayesian MCMC method, firstly a prior distribution is defined for all unknown parameters. Then, estimation of parameters is carried out using MCMC techniques (Shen, 2011). This approach in BayesX provides numerically effective methods for STAR models.

The parameters in the model presented here are estimated using penalized likelihood. Model comparisons and parameter selection are achieved using goodness of fit criteria such as AIC, BIC and Generalized Cross-Validation (GCV).

### 6.4.1 Bayesian perspective on continuous covariates using P-splines

An advantage of using a mixed model representation is that it allows the investigation of the challenges inherent in non-parametric regression from an alternative perspective. In order to evaluate the effect of continuous covariates, a P-spline approach is used (Eilers & Marx 1996), where the smooth function $f$ is estimated by a polynomial spline with equally spaced knots. The biggest problem in using this approach is to determine the correct number of knots (too many or too few will cause over/under fitting, which would fail to estimate the smooth function properly). Usually a moderate number of knots, for example between 20 and 40, will be enough to adequately estimate the smoothness but prevent over fitting. The equally spaced knots are used to impose penalties on B-spline coefficients based on first and second differences of the smooth function. As a result, the approach creates penalized likelihood estimation, including penalty terms.

### 6.4.2 Empirical Bayes Estimation Based on Mixed Model Methodology

In the mixed model representation within BayesX, a re-parameterization is applied, so that STAR models can be expressed as GLMM models (Green, 1987). When STAR models are represented as GLMM models, regression parameters and variance components are evaluated by using iteratively weighted least squares together with REML or marginal techniques. In this approach, the vector containing regression coefficients is divided into two separate parts, the un-penalized and the penalized parts. The GLMM consists of un-penalized parameters as fixed effects and penalized parameters as random effects. Smooth functions and variance parameters can be estimated at the same time (Shen, 2011).

In order to carry out empirical Bayes estimates, estimation procedures are performed iteratively in two stages.

At the first stage, penalized and un-penalized regression parameters are estimated by solving a system of equations involving known variance parameters. Then, at the second stage, variance parameters are revised using REML or marginal log-likelihood. These two steps are repeated until convergence is achieved. When marginal likelihood is used, Fisher's Scoring algorithm is used in the convergence criteria and, if the variance component is small, maximization of the likelihood by this algorithm can create problems and fail. For this reason, if the iterated criterion is found to be smaller than the lower limit for the variance component that is specified by the user, then the estimation of the variance component is stopped. The iterated criterion is defined to be;

$$c\left(T_j^2\right) = \frac{||\widetilde{x}_j \widehat{\beta_j^{pen}}||}{||\hat{\eta}||}$$

Where;

$c\left(T_j^2\right)$ is the iteration criterion for variance component $\left(T_j^2\right)$ at time $j$, that is given by solving systems of equations and estimated by maximising the restricted log likelihood,

$\widetilde{x}_j$ is the model matrix estimated at time $j$,

$\widehat{\beta_j^{pen}}$ is the estimated corresponding penalized β parameters from the random component of the model and

$\hat{\eta}$ is the estimated predictor (Belitz *et al.*, 2012).

### 6.4.3 Application of BayesX Models

As an exploration of how these Bayesian approaches might be suitable for modelling and analysis of data of the type used in this project, three models of the types described above are developed. These are computed in order to study which of these three Bayesian methods gives the best match in terms of AIC and BIC to the data in comparison to the GAMM models of section 6.3. However, the Bayesian modelling represented here is an initial "pilot" study just for overall comparison – the results of the individual models are not interpreted in detail and should be primarily regarded as providing stimulus for future research.

The final model obtained in the previous section (6.3) using the GAMM methodology is used as a source in forming the two alternative, Bayesian perspective, full Bayes model using the MCMC technique and the empirical Bayes model using a mixed model representation. Both models are then compared in terms of assessing goodness of fit to the data.

The first model is an empirical Bayes model using a mixed model representation. It includes the predictors 'smoothed term of interaction of time with diagnosis of anaemia in years', gender, a smoothed term of age at (CKD) diagnosis, stage (of CKD at diagnosis), diabetes and CVD. (N.B. smooth terms for mean systolic blood pressure and for age in years are not included in these Bayesian models due to incomplete data on these features. BayesX models do not allow for missing data in observations). Similarly, model formulations including a smoothed term for age (in years) and time-varying covariates are problematic and are not included in the subsequent model.

The interpretation of the parameters obtained for the models below is analogous to that for the GAMM models of section 6.3, since the smooth functions of the covariates are approximated using the same type of splines both here and in the GAMMs.

Hence the Bayesian model fitted initially is;

#### 6.4.3.1 Model 1 – Empirical Bayesian Mixed Model using MGCV

Model1 <- *bayesx(y~sx(Time, by=AnaemiaDiagnosis, bs="ps") + Gender+ Diagnosticstage + DiabetesDiagnosis+ CardiovascularDisease + sx(Diagnosticage, bs="ps"), random=~(1|ID), method="REML", data=CKDData)* (Belitz *et al.*, 2012) (Wood, 2015) (Umlauf *et al.*, 2015).

Table 6.6: Model Output for Model 1

```
bayesx(formula = y ~ sx(Timeyears, by = AnaemiaDiagnosis, bs = "ps") +
   Gender + Diagnosticstage + DiabetesDiagnosis + CardiovascularDisease +
   sx(Diagnosticage, bs = "ps"), data = RData, random = ~(1 |
   ID), method = "REML")

Fixed effects estimation results:

Parametric coefficients:
                                         Estimate   Std. Error   t-value    Pr(>|t|)
(Intercept)                               2.0110     0.0049     408.8323  < 2.2e-16 ***
GenderMale                                0.1024     0.0023      44.1582  < 2.2e-16 ***
DiagnosticstageStage 3b                   0.0661     0.0024      27.9673  < 2.2e-16 ***
DiagnosticstageStage 4                    0.1530     0.0042      36.8442  < 2.2e-16 ***
DiagnosticstageStage 5                    0.2323     0.0087      26.7712  < 2.2e-16 ***
DiabetesDiagnosisDiabetes diagnosed      -0.0179     0.0022      -8.0409  < 2.2e-16 ***
CardiovascularDiseaseDisease present     -0.0105     0.0023      -4.6178  < 2.2e-16 ***

Signif. codes:  0 < '***' ≤ 0.001 '**' ≤ 0.01 '*' ≤ 0.05 '.' ≤ 0.1 ' ' ≤ 1

Smooth terms:
                                                        Variance Smooth Par.  df     Stopped
sx(Timeyears):AnaemiaDiagnosisDonot_have_the_disease 0.0000   7910.5600  2.1501     0
sx(Timeyears):AnaemiaDiagnosisHaving_the_disease      0.0000   256.5230   3.5008     0
sx(Diagnosticage)                                     0.0017   2.6353     11.1929    0

Scale estimate: 0.0044

N = 3776  df = 23.8438  AIC = -16656.4  BIC = -16507.7
GCV = 0.00446674  logLik = 8352.05  method = REML  family = gaussian
```

### 6.4.3.2 Model 2 - Bayesian model using MCMC techniques

Now, a Bayesian model containing the same predictors is formulated, but this time using the MCMC technique to estimate the model parameters instead of the REML approach. This results in the formulation of a full Bayes model. The R syntax formulation for this second model is;

```
Model2 <- bayesx(y~sx(Time, by=AnaemiaDiagnosis, bs="ps") +  Gender+ Diagnosticstage
+ DiabetesDiagnosis+ CardiovascularDisease + sx(Diagnosticage, bs="ps"), random=~(1|ID
), method="MCMC", data=CKDData)
```

(Belitz *et al*., 2012) (Wood, 2015) (Umlauf *et al*., 2015).

Table 6.7: Model Output for Model 2

```
bayesx(formula = y ~ sx(Timeyears, by = AnaemiaDiagnosis, bs = "ps") +
   Gender + Diagnosticstage + DiabetesDiagnosis + CardiovascularDisease +
   sx(Diagnosticage, bs = "ps"), data = RData, random = ~(1 |
   ID), method = "MCMC")

Fixed effects estimation results:

Parametric coefficients:
                                          Mean     Sd     2.5%     50%    97.5%
(Intercept)                              1.6345  0.1449  1.4508  1.5997  1.8741
GenderMale                               0.1020  0.0024  0.0976  0.1020  0.1066
DiagnosticstageStage 3b                  0.0659  0.0023  0.0613  0.0659  0.0704
DiagnosticstageStage 4                   0.1516  0.0043  0.1431  0.1517  0.1599
DiagnosticstageStage 5                   0.2317  0.0086  0.2153  0.2317  0.2487
DiabetesDiagnosisDiabetes diagnosed     -0.0181  0.0023 -0.0224 -0.0181 -0.0136
CardiovascularDiseaseDisease present    -0.0104  0.0022 -0.0146 -0.0104 -0.0061

Smooth terms variances:
                                Mean     Sd     2.5%    50%     97.5%    Min     Max
sx(Timeyears):Non-anaemic      0.0000  0.0000  0.0000  0.0000  0.0001  0.0000  0.0008
sx(Timeyears):Anaemic          0.0001  0.0001  0.0000  0.0000  0.0002  0.0000  0.0014
sx(Diagnosticage)              0.0054  0.0065  0.0005  0.0032  0.0230  0.0003  0.0679

Scale estimate:
                  Mean      Sd      2.5%     50%      97.5%
Sigma2 (σ²)     0.0044   0.0001   0.0042   0.0044   0.0046

N = 3776  burnin = 2000  DIC = 3811.825  pd = 34.14964
method = MCMC  family = gaussian  iterations = 12000  step = 10
```

### 6.4.3.3 Model 3 - Empirical Bayesian Mixed Model using both MGCV and BayesX

The final Bayesian model presented here is the same penalised likelihood (empirical Bayes) model but using mixed model representation and REML with the aid of both BayesX and Mixed GAM Computation Vehicle (MGCV) packages in R. The formulation in R syntax for model 3 is;

Model3 <- *bayesx(y~s(Time, by=AnaemiaDiagnosis, bs="ps") + Gender+ Diagnosticstage + DiabetesDiagnosis+ CardiovascularDisease + s(Diagnosticage, bs="ps"), random=~(1|ID), method="REML", data=CKDData)* (Belitz *et al.*, 2012) (Wood, 2015) (Umlauf *et al.*, 2015).

Table 6.8: Model Output for Model 3

```
bayesx(formula = y ~ s(Timeyears, by = AnaemiaDiagnosis, bs = "ps") +
   Gender + Diagnosticstage + DiabetesDiagnosis + CardiovascularDisease +
   s(Diagnosticage, bs = "ps"), data = RData, random = ~(1 |
   ID), method = "REML")

Fixed effects estimation results:

Parametric coefficients:
                                           Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)                                 2.0117    0.0124    162.8066  < 2.2e-16 ***
GenderMale                                  0.1026    0.0023    44.3708   < 2.2e-16 ***
DiagnosticstageStage 3b                     0.0664    0.0024    28.1464   < 2.2e-16 ***
DiagnosticstageStage 4                      0.1534    0.0041    37.0691   < 2.2e-16 ***
DiagnosticstageStage 5                      0.2326    0.0087    26.8317   < 2.2e-16 ***
DiabetesDiagnosisDiabetes diagnosed        -0.0180    0.0022    -8.1156   < 2.2e-16 ***
CardiovascularDiseaseDisease present       -0.0102    0.0023    -4.5083   < 2.2e-16 ***
---
Signif. codes:  0 < ‘***’ ≤ 0.001 ‘**’ ≤ 0.01 ‘*’ ≤ 0.05 ‘.’ ≤ 0.1 ‘ ’ ≤ 1

Smooth terms:
                                                        Variance  Smooth Par.  df      Stopped
sx(Timeyears):AnaemiaDiagnosisDonot_have_the_disease    0.0000    438.9780     1.9956  0
sx(Timeyears):AnaemiaDiagnosisHaving_the_disease        0.0004    10.1618      3.3130  0
s(Diagnosticage)                                        0.0506    0.0880       7.3670  0

Scale estimate: 0.0045

N = 3776  df = 19.6757  AIC = -16647.9  BIC = -16525.2
GCV = 0.00447675  logLik = 8343.65  method = REML  family = gaussian
```

### 6.4.3.4 Summary and Discussion of Results of Bayesian modelling

Models 1 and 3 are effectively the same model, but use different techniques for the estimation of the smoothed terms within the model (i.e. Model 3 uses MGCV package instead of BayesX). When considering the goodness of fit statistics for Models 1 and 3 (Table 6.9), it can be observed that Model 3 has the lower AIC and BIC scores but higher GCV score. Model 3 therefore will subsequently be compared against Model 2.

Table 6.9: Model Fit Comparison of Model 1 and Model 3

| Model Number | AIC | BIC | GCV |
|---|---|---|---|
| Model 1 | -16656.4 | -16507.7 | 0.00446674 |
| Model 3 | -16647.9 | -16525.2 | 0.00447675 |

Model 1 and 3 are empirical Bayes models, which means that these models are semi-parametric models, allowing mixed model inference from a Bayesian perspective. On the other hand, Model 2 is a fully Bayesian model, which allows the formulation of generalized regression models from a Bayesian perspective without considering a mixed model representation.

The output for models 2 and 3 indicate that all of the predictors included in both of the models are statistically significantly affecting the change in SCr. The significant smooth terms indicate a non-linear relationship between the smooth predictors and SCr.

Table 6.10: 99% Confidence limits for Model 2

| | 0.5% | 99.5% |
|---|---|---|
| (Intercept) | 1.44023315 | 1.887537600 |
| GenderMale | 0.09635132 | 0.108796685 |
| DiagnosticstageStage 3b | 0.06006566 | 0.072057798 |
| DiagnosticstageStage 4 | 0.14044537 | 0.162441255 |
| DiagnosticstageStage 5 | 0.21093842 | 0.251973445 |
| DiabetesDiagnosisDiabetes diagnosed | -0.02405719 | -0.012076600 |
| CardiovascularDiseaseDisease present | -0.01584129 | -0.005213282 |

Table 6.11: 99% Confidence limits for Model 3

| | **0.5%** | **99.5%** |
|---|---|---|
| (Intercept) | 1.97650141 | 2.046938585 |
| GenderMale | 0.09599815 | 0.109177850 |
| DiagnosticstageStage 3b | 0.05966755 | 0.073113455 |
| DiagnosticstageStage 4 | 0.14158953 | 0.165176473 |
| DiagnosticstageStage 5 | 0.20789375 | 0.257310247 |
| DiabetesDiagnosisDiabetes diagnosed | -0.02433946 | -0.011686937 |
| CardiovascularDiseaseDisease present | -0.01668650 | -0.003759899 |

Note that the methodology used for computing Model 2 does not yield "p-values" for the significance of parameters on terms due to its use of the MCMC approach. However, we can infer whether or not each contribution is or is not significant by whether the confidence interval for that term does include (not significant) or does not include (significant) zero. Table 6.10 and 6.11 show the confidence limits for the parametric terms of models 2 and 3 respectively. It can be observed that all the confidence intervals for those parameter do not include zero, so all the parameters are significant.

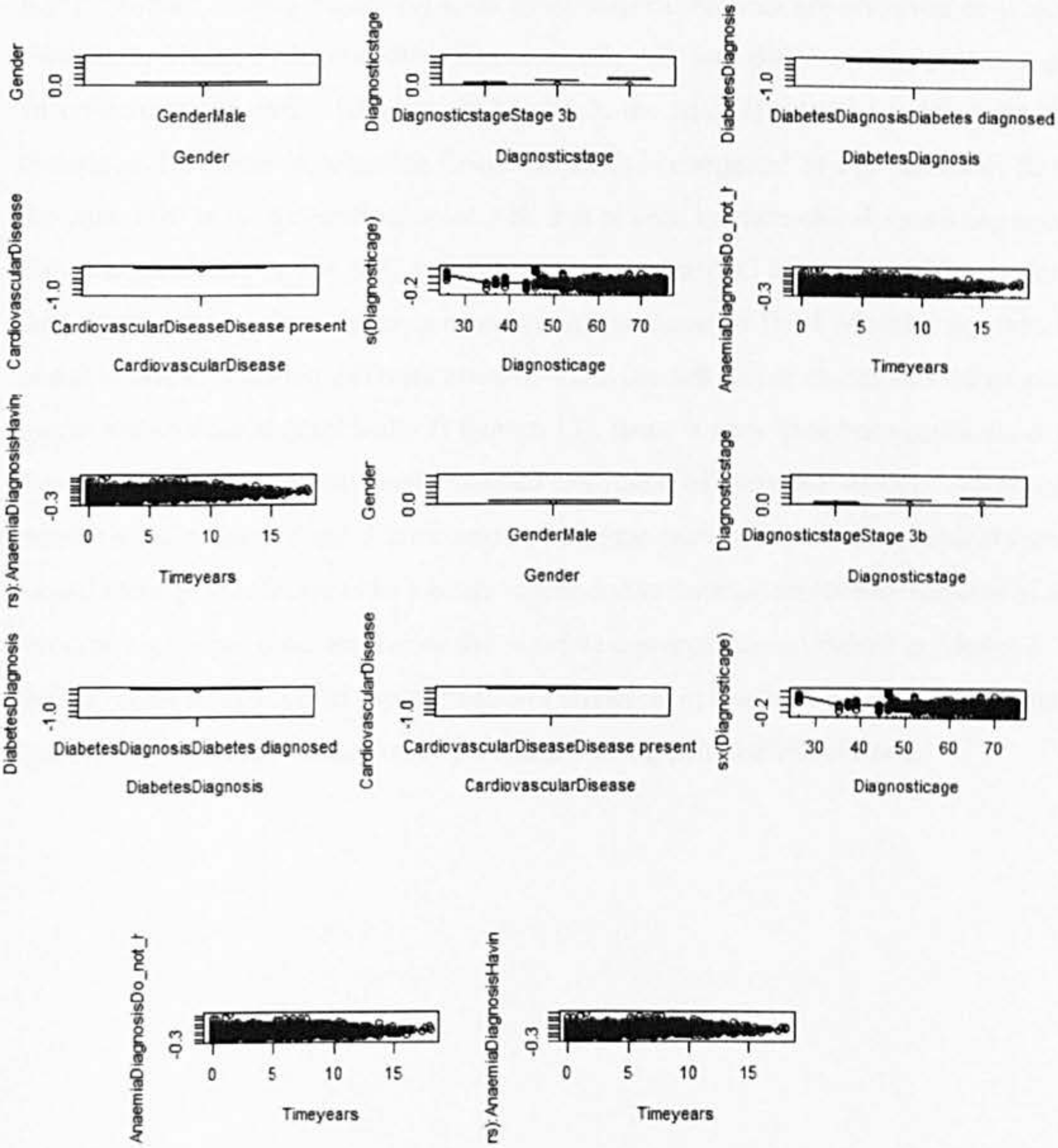

Figure 6.13: Plots of partial effects of the covariates and boxplots for each fixed effects using Model 2 and Model 3 together

It can be concluded that when the partial effects of Model 2 are compared graphically against Model 3 using Figure 6.13, no substantial differences are observed between these two models. In Model 3, the empirical Bayes model, AIC and BIC are computed as a goodness of fit criteria to the data. However, in Model 2, the fully Bayesian Model using the MCMC technique, Deviance Information Criterion (DIC) is computed as a goodness of fit criterion to the data. DIC is the generalisation of AIC that is used in hierarchical modelling specifically in Bayesian models. As like AIC and BIC, the lower the DIC is, the better the model fits to the data. Furthermore, since direct comparison of goodness of fit of Model 2 and Model 3 is not possible due to resulting different criteria, when the full Bayes model and the empirical Bayes model are compared graphically (Figure 6.13), there is very little/not significant difference in the observed parameter estimates obtained from each of these two models. Additionally, when properties of models 2 and 3 are compared against each other, the Empirical Bayesian Mixed Model (Model 3) is found to be a better model due to the high computational cost of the MCMC iteration technique used and hence the possible convergence problems in Model 2. Therefore, the Bayesian mixed model representation (Model 3) is found to be a useful alternative to the frequentist model for estimation of parameters using penalised likelihood.

# 7 Conclusion, Challenges and Future Research

## 7.1 Conclusion

The research presented in this thesis illustrates applications of advanced and developing statistical methodologies to analyse routinely collected longitudinal data, specifically repeated measurements from a representative sample of 129 general practices in England and Wales. The work details many of the new and emerging statistical methodologies that are being used in this area and demonstrates the significant potential for the use of such methods in the field of health research, while highlighting both advantages and disadvantages encountered therein. The application here is based on the study of the progression of CKD, and the results obtained show that the techniques applied can help to further understanding of the natural progression of CKD; and to explain patterns of change and the influence of key factors on this progression. From a wider perspective the procedures followed and mechanisms applied would be applicable to the study of any chronic disease using GP data, and to many other areas of health research and monitoring; as well as in the analysis of longitudinal data in other fields.

In our dataset of GP records there are between two and fifteen repeated (eGFR) measurements per patient and these are used to develop parametric, semi-parametric and non-parametric regression models. In general, three main study questions are investigated;

1- What is the natural history of changes in the progression of CKD over time?

2- Are there any differences in changes in this progression between patients?

3- Can differences between patients be explained by patient level covariates such as age, gender and the presence of different co-morbidities?

### 7.1.1 Conclusion on review of existing literature and descriptive statistics

A review of existing literature revealed that the use of advanced statistical methods for the analysis of routinely collected longitudinal health data is not wide-spread (chapter 2). Before starting statistical modelling of any dataset it is wise to undertake some preliminary basic investigation of key variables within the data that are relevant to the study questions. Bearing in mind that the main outcome of interest is in the change over time in the response (repeated eGFR readings) and factors affecting this change, a summary of descriptive statistics is presented in chapter 3. Basic analyses are performed on the data in order to understand more about the nature of the dataset. The results indicated that the prevalence of CKD within our data is concurrent with estimates of CKD within the UK population as a whole. Further, examination of population demographics (age, gender, ethnicity, etc.) against national figures validated that our dataset is a good representation of the general population of the UK.

### 7.1.2 Conclusion on logistic regression analysis

Subsequently and as a starting point for complex statistical analysis logistic regression models are applied to the data for the purpose of finding which-co-morbidities have the greatest impact on diagnosis of CKD. In chapter four, odds ratios and effect sizes are used to assess the impact of co-morbidities on CKD, where the outcome of interest is binary,based on whether the patient has been diagnosed to have CKD or not. From this analysis, it has been proven that cardiovascular diseases, diabetes, anaemia and stroke are the most influential factors influencing in the occurrence of CKD in individual patients. This is consistent with the existing medical literature.

The logistic analyses also focus on the outcome 'rapid decline' (i.e. whether a patient has experienced rapid decline of CKD status) – two models are run based on the two definitions of rapid decline recognised in the clinical setting. Obviously these analyses begin to focus only on patients who have been diagnosed with CKD, a smaller group than before. These two models aim to identify which factors influence rapid decline (each definition is modelled separately). From the results we can conclude that all of the predictors considered are significantly important, and further that the primary factors affecting the diagnosis of CKD are different from the primary factors affecting the progression of CKD, a previously unrecognised fact within existing medical literature. For instance, cardiovascular diseases have the highest impact on the occurrence of CKD but diagnosis of anaemia has the highest impact on the progression of CKD. Further factors affecting the rate of decline differ between definitions; when progression is classified as rapid decline according to definition 1(more than 5ml decline in eGFR within a year), diagnoses of diabetes and anaemia were the only significant factors among those examined. However, when rapid decline or progression of CKD is based on definition 2 (i.e. more than 10 ml decline in eGFR within 5 years), all of the co-morbidities were significant.

This means that diabetes and anaemia are the only factors affecting rapid decline within the short term; whereas all of the factors are important in terms of influencing rapid decline of kidney function over a longer time period. This fact is not reported within the clinical CKD community and the results obtained show the additional understanding of progressive health conditions can be achieved through the analysis of longitudinal data.

### 7.1.3 Conclusion on parametric models

The next step in this research was to move towards investigating progression of CKD and how this varies between patients, by taking repeated observations of eGFR into consideration. In chapter 5, a series of different parametric models of increasing complexity from linear mixed models to generalized linear mixed models and polynomial linear mixed models were developed. These are suitable as both random and fixed effects are considered in the model design. This means that they allow explanation of total variation in the outcome (i.e. repeated eGFR measurements) as two components, variation between individuals (fixed effects) and variation within individuals (random effects). These models are found to be parsimonious and efficient, and enable inferences from the results obtained to be generalised and applied to the wider population. In LMMs, when there are no predictors at either level, the null model showed that 34.45% of the total unexplained variation is due to within-subject variation and 65.55% of the total variation is due to between-subject variation. When time is added at both levels (i.e. as level 1 and level 2 variable), the LMMs was improved by explaining 13.48% of with-in subject variation (of 34.45%, leaving 20.97% still unexplained) and 6% of between-subject variation (of 65.55%, leaving 59.55% still unexplained). Furthermore, addition of level 2 covariates and factors further improved the LMM by increasing explained with-in subject variation from 13.48% to 14.07% and explained between-subject variation from 6% to 17.87%.

However, as only patients with CKD are included, then the distribution of the outcome (eGFR readings) is unlikely to satisfy the normality assumption for linear mixed models. Therefore, appropriate transformations were used to bring the distribution of the transformed data closer to a normal distribution and make the data suitable for analysis using both LMMs and GLMMs. Firstly log transformations of repeated eGFR values are modelled as a linear function of co-morbidities and time (model 2, section 5.6.1.2, eq. (5.21)). This is considered as the baseline GLMM. This model showed that 31.62% of the total variation in eGFR readings is due to with-in subject variation whereas 68.38% of the total variation is due to between-subject variation. A further development (model 3, section 5.6.1.3, eq. (5.22)) is performed by running a GLMM model assuming a gamma distribution, this model improves on the baseline GLMM (model 2, eq. (5.21)) by explaining 15.59% of unexplained 31.62% with-in subject variation (leaving 16.03% still unexplained with-in subject variation) and by explaining 18.81% of unexplained 68.38% between-subject variation (leaving 49.57% still unexplained between-subject variation). Further, when a GLMM model using a log link function and assuming a gamma distribution is fitted (model 4, section 5.6.1.4, eq. (5.23)) the explained with-in subject variation increases from 15.59% to 18.59% and the explained between-subject variation from 18.81% to 39.32%. A final modification of model 4 uses a simpler covariance matrix (i.e. model 5, section 5.6.1.5, eq. (5.24)) and results in a further improvement in explained within subject variation from 39.32% to 42.74% but without making a great difference in the explanation of with-in subject variation. The results obtained from LMMs and GLMMs have showed that significant improvement is achieved in the explanation of both types of variations specifically between-subject variation when GLMMs are used instead of LMMs.

However, due to the great heterogeneity between patients, the linearity assumption between outcome and independent predictors required for these modelling techniques (LMM and GLMM) is unlikely to be valid. Therefore, both LMMs and GLMMs might be too restrictive and less robust against violations of the linearity assumption. For these reasons, polynomial models were constructed to look at the non-linear relationship between eGFR and time. It was concluded that the model obtained, based on various transformations, even with the addition of polynomial terms, gave a rather poorer model fit in terms of taking account for non-linear behaviour of eGFR over time.

### 7.1.4 Conclusion on semi-parametric and non-parametric models

A different approach is described in chapter 6, where semi-parametric GAMM models and Bayesian models were constructed. When the assumptions of the parametric models such as assuming a normal distribution for the outcome and assuming a linear relationship between outcome and independent variables are satisfied, parametric models are more powerful than semi-parametric or non-parametric models. However, when these assumptions of the parametric models are violated, semi-parametric or non-parametric models are more powerful than parameteric models. Hence, GAMM models described in chapter 6 are more robust against violations of the linearity assumption than the parametric models described in chapter 5. In GAMM modelling, penalised smoothing spline methods are used for continuous covariates, e.g. systolic blood pressure or age, to model the outcome variable as a non-linear function using the available data. In doing so, an appealing model is constructed that contains a good combination of the properties of mixed effects models (i.e. combining fixed linear component with random component) whilst also using smoothing splines to model non-linearity between the covariate and the outcome. When the best GLMM model described in chapter 5 is carried out by using GAMM modelling approach, it is found that more data points are needed in order to evaluate the interaction effects of different co-morbidities. Therefore, only single effects are calculated and the model results show 4.85% of the total variation in the outcome explained. This indicates that, in such cases, GLMM models are more powerful than GAMM models as they can explain a greater proportion of the variation in the outcome. The application of GAMM and Bayesian methodologies to model the complex, non-linear relationship between serum creatinine measures (SCr – substitute for eGFR) and covariates over time are used to to describe the progression of CKD and how defined factors such as age, gender and various co-morbidities influencing this progression. Factors included age, gender, CKD stage at diagnosis, age at CKD diagnosis and interactions of time with diagnoses of anaemia, diabetes, cardiovascular disease and mean values of systolic blood pressure. The results of the GAMM models show that there is a significant non-linear relationship between SCr and the independent covariates and heterogeneity between patients. When the outcome variable is changed to SCr in order to evaluate the effect of age, gender and various co-morbidities on the change in SCr and hence on the change in eGFR, in such models the results showed that explained variation is increased to 58.9% of total variation in GAMM models where a greater amount of variation is explained compared to GLMM models.

In GAMM models, since such models are from the frequentist perspective, the model coefficients for the effect of independent variables are estimated by using regression models such as REML approaches. Further in chapter 6, additional models are carried out from a

Bayesian perspective in order to see if alternative Bayesian models are available which will have similar power. The Bayesian perspective models can be either the full bayes models where the coefficients of are estimated by using MCMC approach; or mixed model representations where in relation to mixed models, empirical bayes estimation (i.e. REML) is used to evaluate the model coefficients. When these two models from the Bayesian perspective are compared with the model performed using frequestist approach, very similar results are achieved. From a Bayesian perspective the mixed model representations are preferable as an alternative models to GAMM models due to the high computational cost of the MCMC iteration technique and possible convergence problems in the full Bayes models.

## 7.2 Limitations of the dataset

While data collection within the health service is continually improving, the routine use of very large and complex datasets poses many challenges. Here we have used a secondary data source which was derived from raw GP records. The secondary nature of the data and the manipulations which were carried out on it prior to this investigation has limited the scope for clinical interpretation of the findings and any such findings should be viewed as indicative rather than conclusive. However it is reassuring that despite the limitations of the data the findings do reflect current knowledge within the CKD field and suggest additional knowledge that may be worthy of further investigation. The analysis of this reduced and condensed dataset has also highlighted the complexities involved in utilising routinely collected data. Here, even with using much condensed data, obtaining good fitting models proved to be difficult. We have had to make several assumptions in working with this data, for example, since the existence of co-morbidities are labelled prior the data collection from 129 participating practices, it was assumed that diagnoses of all the co-morbidities were made before the diagnosis of CKD for all patients. This is unlikely to be the case in real life and is something that could be overcome using a primary raw data source. However, the focus of this thesis is the application and assessment of the statistical and modelling methodologies rather than findings relating to clinical aspects of CKD. As data collection in the primary care setting improves and more reliable and cleaner data is obtained, fewer assumptions may be necessary and valid data analysis may be more readily achievable.

## 7.3 Final Overall Statements

In conclusion, the research presented in this thesis shows that there is much potential for the statistical analysis of routinely collected GP data to further understanding and knowledge of the natural progression of chronic disease. Methodologies that would suit the complexity of the medical questions that arise in this clinical area are available and developing. However such methods are very complex and would require skilled researchers and data analysts working alongside medical expertise to undertake meaningful investigations. Similarly further development of appropriate, user friendly software would facilitate further research and usability of such data and analysis techniques.

Despite the many obstacles encountered this research, the research shows that meaningful results can be obtained using GP records and can reveal trends and associations that may otherwise go unrecognised. For example in our case study into CKD, the general decline of eGFR over time was confirmed, however substantial heterogeneity between patients was found in this decline. Another observation was a possible slight increase in eGFR within a year after the diagnosis of CKD. This might be due to the initial effects of starting appropriate medications. Further, it was found that the relationship between eGFR and time was significantly non-linear both overall and in individual patients. In fully understanding the progression of CKD it is very important to take this non-linearity into account over larger timescales (e.g. several years); current research assumes linear decline.

When various covariates and influencing factors are taken into account to explain part of the total variation in the outcome, adjustments of the covariate included is found to significantly improve the proportion of the total variation of the outcome explained by the model. Analyses performed in this study also indicate that the random effects due to differences between repeated measurements within patients, the non-linear change in the outcome over time and patient level factors such as diagnoses of co-morbidities should be taken into consideration in the statistical model formulation. If these are considered, and non-linear mixed models are used in longitudinal analysis of the progression of CKD, then around 58.90 % of the total variation in the outcome is explained which can be compared against explanation of around 20% of total variation in the outcome by using linear mixed models.

## 7.4 Challenges and Future Research

The main challenges for this type of research lie in the complexity of the appropriate methodologies and the need for very skilled analysts to undertake their application. For example the outcome considered in the semi-parametric models developed in this research is transformed into the log domain to ensure that the distribution of the response is Gaussian, because the theoretical properties of smoothing splines are only appropriate for Gaussian data. Even when the data follow a Gaussian distribution, the REML estimator used to evaluate the smooth terms is not well established. The GAMMs and the alternative semi-parametric Bayesian models are found to give better fits to the data than the LMMs or GLMMs. However, the complexity of GAMMs and semi-parametric Bayesian models can make such models less efficient and hence such models need further theoretical investigation. Furthermore, in non-parametric Bayesian models, the Markov Chain Monte Carlo (MCMC) iteration approach used, might be computationally costly and could create convergence problems. Therefore, theoretical investigation should be performed on such models as well, to assess consistency, efficiency and robustness.

However, there is vast potential for the analysis of routinely collected computerized GP records, and other health data sources for research. They are particularly attractive as they are available at little or no cost to researchers and this study validates that datasets of this type can be used to investigate the progression of chronic diseases using suitable methodologies such as mixed-effect models. Since the dataset was shown to be a good representation of the UK population, the inferences made can also be applicable to the general population. Further research might also include consideration of the medications taken by the CKD patients, as this information is also routinely collected. This might allow further understanding of the impact of treatments on the progression of CKD, over and above the effects of co-morbidities. Other further work might be to consider changes in the status of co-morbidities and covariates over time and the effect this has on progression of CKD, which should be achievable using GP records.

The methods and techniques presented within this work illustrate the use of complex statistical modelling to the GP data and how they can lead to useful and meaningful interpretations. It also highlights that the process is not simple but with development should lead to easier applications. It is clear that routinely collected data contains much valuable information at minimal expense and as such should be more widely utilised both in health research and further afield.

# 8 References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM.* Oxford University Press: Oxford.

Albert, P. S. (1999) 'Longitudinal data analysis (repeated measures) in clinical trials', *Statistics in Medicine*, 18, pp. 1707-1732.

Ameresan, M. S. and Geetha, R. (2008) 'Early diagnosis of CKD and its prevention', *The Journal of the Association of Physicians of India*, 56, pp. 41-46.

Andrews, P. A. (2008) 'Early identification and management of chronic kidney disease in adults in primary and secondary care: a commentary on NICE Guideline No 73, Sept 2008', *The British Journal of Diabetes & Vascular Disease*, 8(6), pp. 257.

Astor, B., Matsushita, K., Gansevoort, R., van der Velde, M., Woodward, M., Levey, A. S. et al. (2011) 'Lower estimated glomerular filtration rate and higher albuminuria are associated with mortality and end-stage renal disease. A collaborative meta-analysis of kidney disease population cohorts', *Kidney Internationl*, 79(12), pp. 1331-1340.

Australia and New Zealand Dialysis and Transplant Registry (2011) 'ANZDATA 33rd annual report', *ANZDATA*, Aldelaide.

Azuero, A. Pisu, M., McNees, P., Burkhardt, J., Benz, R. et al. (2010) 'An Application of Longitudinal Analysis with Skewed Outcomes', *National Institute of Health*, 59(4), PP. 301-307.

Barri, Y. M. (2006) 'Hypertension and kidney disease: a deadly connection', *Current Cardiology Reports*, 8(6), pp. 411-417.

Bates, D. (2012) 'Computational methods for mixed models', *Vignette for lme4.*

Baver, J. H., Brooks, C. S. and Burch, R. N. (1982) 'Clinical appraisal of creatinine clearance as a measurement of glomerular filtration rate', *American Journal of Kidney Diseases*, 2(3), pp. 337-346.

Beaumont, R. (2012) 'Analysing repeated measures with Linear Mixed Models (Random Effects Model)', *Hlcourseweb*, 1.

Belitz, C., Bezger, A., Kneib, T. and Cang, S. (2012) 'BayesX Software for Bayesian Inference in Structured Additive Regression Models Version 2.0.1: *Methodology Manual*'.

Belitz, C., Bezger, A., Kneib, T. and Cang, S. (2012) 'BayesX Software for Bayesian Inference in Structured Additive Regression Models Version 2.0.1: *Reference Manual*'.

Belitz, C., Bezger, A., Kneib, T. and Cang, S. (2012) 'BayesX Software for Bayesian Inference in Structured Additive Regression Models Version 2.0.1: *Tutorials*'.

Berger, A. K., Duval, S., Manske, C., Vazquez, G., Barber, C., Miller, L. and Luepker, R. V. (2007) 'Angiotensin – converting enzyme inhibitors and angiotension receptor blockers in patients with congestive heart failure and chronic kidney disease', *American Heart Journal*, 153(6), pp. 1064-1073.

Bickel, R. (2007) *Chapter 5: Developing the Multilevel Regression Model*, Guilford Publications.

Black, C., Sharma, P., Scotland, G., McCullough, K., McGurn, D., Robertson, L., Fluck, N., MacLeod, A., McNamee, P., Prescott, G. and Smith, C. (2010) 'Early referral strategies for mangement of people with markers of renal disease: a systematic review of the evidence of clinical effectiveness, cross-effectiveness and economic analysis', *Health Technology Assessment (Winchester, England)*, 14(21), pp. 1-184.

Brenner, B., Meyer, T. and Hostetter, T. (1982) 'Dietary protein intake and the progressive nature of kidney disease: The role of hemodynamically mediated glomerular injury in the pathogenesis of progressive glomerular sclerosis in aging, renal ablation and intrinsic renal disease', *The New England Journal of Medicine*, 307(11), pp. 652-659.

Breslow, N. E. and Clayton, D. G. (1993) 'Approximate Inference in Generalized Linear Mixed Models', Journal of the American Statistical Association, 88, pp. 9-25.

Brown, H. and Prescott, R. (1999) *Applied Mixed Models in Medicine*, John Wiley & Sons: New York.

Bruch, C., Rothenburger, M., Gotzmann, M., Wichter, T., Scheld, H. H., Breithardt, G. and Gradaus, R. (2007) 'Chronic kidney disease in patients with chronic heart failure – Impact on intracardiac conduction, diastolic function and prognosis', *International Journal of Cardiology*, 118(3), pp. 375-380.

Burden, R. and Tomson, C. (2005) 'Identification, management and referral of adults with chronic kidney disease: concise guidelines', *Clinical Medicine (London, England)*, 5(6), pp. 635-642.

Burden, R., Thomson, C., Guideline Development Committee, Joint Specialty Committee on Renal Disease of the Royal College of Physicians of London and the Renal Association (2005) 'Identification, management and referral of adults with chronic kidney disease: concise guidelines', *Clinical Medicine*, 5(6), pp. 635-642.

Burns, R. B. and Burns, R. A. (2008) *Research Methods and Statistics Using SPSS.* 1st edn. SAGE Publications Ltd: London.

Chadban, S. J., Briganti, E. M., Kerr, P. G., Dunstan, D. W., Welborn, T. A., Zimmet, P. Z. et al. (2003) 'Prevalence of kidney damage in Australian adults: The AusDiab kidney study', *Journal of the American Society of Nephrology*, 14(7 Suppl 2), pp. S131-S138.

Chen, N., Wang, W., Huang, Y., Shen, P., Pei, D., Yu, H., Shi, H., Zhang, Q., Xu, J., Lv, Y. and Fan, Q. (2009) 'Community-based study on CKD subjects and the associated risk factors', *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association – European Renal Association*, 24(7), pp. 2117-2123.

Cockcroft, D. W. and Gault, M. H. (1976) 'Prediction of creatinine clearance from serum creatinine', *Nephron*, 16(1), pp. 31-41.

Coresh, J., Selvin, E., Stevens, L. A., Manzi, J., Kusek, J. W., Eggers, P. et al. (2007) 'Prevalence of chronic kidney disease in the United States', *Journal of American Medical Association*, 298(17), pp. 2038-2047.

Cox, D. R. and Snell, E. J. (1989) *Analysis of Binary Data.* 2nd edn. Chapman & Hall / CRC.

Crinson, I., Gallagher, H., Thomas, N. and de Lusignan, S. (2010) 'How ready is general practice to improve quality in chronic kidney disease? A diagnostic analysis', *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, 60(575), pp. 403-409.

Crowe, E., Halpin, D. and Stevens, P. (2008) 'Early identification and management of chronic kidney disease: summary of NICE guidance', *British Medical Journal*, 337.

Culleton, B., Larson, M., Parfrey, P., Kannel, W. and Levy, D. (2000) 'Proteinuria as a risk factor for cardiovascular disease and mortality in older people: A prospective study', *American Journal of Medicine*, 109(1), pp. 1-8.

Damsgaard, E., Froland, A., Jorgensen, O. and Mogensen, C. (1990) 'Microalbuminuria as predictor of increased mortality in elderly people', *British Medical Journal*, 300(6720), pp. 297-300.

Davidian, M. and Giltinan, D. M. (1993) 'Some general estimation methods for non-linear mixed models', *Journal of Biopharmaceutical Statistics*, 3, pp. 23-55.

Davidian, M. and Giltinan, D. M. (2003) 'Nonlinear models for repeated measurement data: An overview and update', *Journal of Agricultural, Biological, and Environmental Statistics*, 8, pp. 387-419.

Davis, C. S. (2002) *Statistical Methods for the Analysis of Repeated Measurements*. 1st edn. New York: Springer.

Dawson, J. D. (1994) 'Comparing treatment groups on the basis of slopes, areas-under-the-curve, and other summary measures', *Drug Information Journal*, 28, pp. 723-732.

De Boor (1978) *A Practical Guide to Splines*. Springer: New York.

De Lusignan, S. (2010) 'Computerised routinely collected primary care data: Essential for patient access to records, quality improvement and research', *Informatics in Primary Care*, 18(1), pp. 5-7.

De Lusignan, S., Chan, T., Stevens, P., O'Donoghue, D., Hague, N., Dzregah, B., Van Vlymen, J., Walker, M. and Hilton, S. (2005) 'Identifying patients with chronic kidney disease from general practice computer records', *Family Practice*, 22, pp. 234-241.

De Lusignan, S., Gallagher, H., Chan, T., Thomas, N., van Vlymen, J., Nation, M., Jain, N., Tahir, A., Du Bois, E., Crinson, I., Hague, N., Reid, F. and Harris, K. (2009) 'The QICKD study protocol: A cluster randomised trial to compare quality improvement interventions to lower systolic BP in chronic kidney disease (CKD) in primary care', *Implementation Science*, 4(1).

De Lusignan, S., Tomson, C., Harries, K., van Vlymen, J. and Gallagher, H. (2010) 'Creatinine Fluctuation Has a Greater Effect than the Formula to Estimate Glomerular Filtration Rate on the Prevalence of Chronic Kidney Disease', *Nephron – Clinical Practice*, 117(3), pp. c213-c224.

Diggle, P. J. (1988) 'An approach to the analysis of repeated measurements', *Biometrics*, 44, pp. 959-971.

Diggle, P. J., Heagerty, P. J., Liang, K. Y. and Zeger, S. L. (2002) *Analysis of Longitudinal Data*. Oxford University Press: Oxford.

Diggle, P. J., Liang, K. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. 1st edn. Oxford: Clarendon Press.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1996) *Analysis of longitudinal data*. Oxford University Press.

Dobson, A. J. (1990) *An Introduction to Generalized Linear Models*. Chapman and Hall: London.

Drüeke, T., Locatelli, F., Clyne, N., Eckardt, K., Macdougall, I., Tsakiris, D. et al. (2006) 'Normalization of hemoglobin level in patients with chronic kidney disease and anemia', *The New England Journal of Medicine*, 355(20), pp. 2071-2084.

Du Bois, D. and Du Bois, E. F. (1916) 'A formula to estimate the approximate surfact area if height and weight be known', *Archives of Internal Medicine*, 17, pp. 863.

Eilers, P. H. C. and Marx, B. D. (1996) 'Flexible smoothing with B-splines and penalties', *Statistical Science*, 11(2), pp. 89-121.

El Nahas, A. M., Masters-Thomas, A., Brady, S. A., Farrington, K., Wilkinson, V., Hilson, A. J., Varghese, Z. and Moorhead, J. F. (1984) 'Selective effect of low protein diets in chronic renal diseases', *British Medical Journal (Clinical Research Ed.)*, 289(6455), pp. 1337.

ERA EDTA Registry (2011) 'ERA EDTA registry annual report 2009', *ERA EDTA Registry*, Amsterdam.

Fahrmeir, L. and Lang, S. (2004) 'Bayesian inference for generalized additive mixed models based on markov random field priors', *Applied Statistics*, 50, pp. 201-220.

Field, A. (2013) *Discovering Statistics Using IBM SPSS Statistics*. 4th edn. SAGE Publications Ltd: London.

Fieuws, S., Verbeke, G. and Molenberghs, G. (2007) 'Random-effects models for multivariate repeated measures', Statistical Methods in Medical Research, 16, pp. 387-398.

Filler, G., Brokenkamp, A., Hofmann, W., Le Bricon, T., Martinez-Bru, C., and Grubb, A. (2005) 'Cystatin C as a marker of GFR – history, indications, and future research', *Clinical Biochemistry*, 38(1), pp. 1-8.

Fitzmaurice, G. and Molenberghs, G. (2009) 'Advances in longitudinal data analysis: An historical perspective', *Longitudinal Data Analysis*, pp. 1.

Fitzmaurice, G. M. (2009) *Longitudinal data analysis*. Chapman & Hall / CRC.

Fitzmaurice, G., Laird, N. and Ware, J. (2004) 'Applied longitudinal analysis', *New York: Wiley*, pp. 187-234.

Foley, R., Parfrey, P., Harnett, J., Kent, G., Murray, D. and Barre, P. (1996) 'The impact of anemia on cardiomyopathy, morbidity, and mortality in end-stage renal disease', *American Journal of Kidney Diseases*, 28(1), pp. 53-61.

Foley, R., Parfrey, P., Sarnak, M. (1998) 'Epidemiology of cardiovascular disease in chronic renal disease', *Journal of the American Society of Nephrology*, 12(Suppl), pp. S16-S23.

Forbes, G. and Bruining, G. (1976) 'Urinary creatinine excretion and lean body mass', *American Journal of Clinical Nutrition*, 29(12), pp. 1359-1366.

Fouque, D. and Aparicio, M. (2007) 'Eleven reasons to control the protein intake of patients with chronic kidney disease', *Nature Clinical Practice Nephrology*, (7), pp. 383-392.

Fouque, D., Laville, M. and Boissel, J. P. (2006) 'Low protein diets for chronic kidney disease in non diabetic adults', *Conchrane Database of Systematic Reviews (online)*, (2).

Fouque, D., Laville, M., Boissel, J. P., Chifflet, R., Labeeuw, M. and Zech, P. Y. (1992) 'Controlled low protein diets in chronic renal insufficiency: meta-analysis', *British Medical Journal*, 304(6821), pp. 216.

Frankel, A., Brown, E. and Wingfield, D. (2005) 'Management of chronic kidney disease', *British Medical Journal*, 330(7499), pp. 1039.

Frison, L. and Pocock, S. J. (1992) 'Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design', *Statistics in Medicine*, 11, pp. 1685-1704.

Froissart, M., Rossert, J., Jacquot, C., Paillard, M. and Houillier, P. (2005) 'Predictive performance of the modification of diet in renal disease and cockcroft-gault equations for estimating renal function', *Journal of the American Society of Nephrology*, 16(3), pp. 763-773.

Gallagher, D., Visser, M., De Meersam, R. E., Sepulveda, D., Baumgartner, R. N., Pierson, R. N. et al. (1997) 'Appendicular skeletal muscle mass: Effects of age, gender and ethnicity', *Journal of Applied Physiology*, 83(1), pp. 229-239.

Gallagher, H., de Lusginan, S., Harris, K. and Cates, C. (2010) 'Quality-improvement strategies for the management of hypertension in chronic kidney disease in primary care: a systematic review', *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, 60(575), pp. 258-265.

Gansevoort, R. T. and de Jong, P. E. (2009) 'The case for using albuminuria in staging chronic kidney disease', *Journal of the American Society of Nephrology*, 20, pp. 465.

Garrett, B. M. (2007) 'Review: a low protein diet delays end stage renal disease or death in chronic kidney disease', Evidence-Based Nursing, 10(1), pp. 10-16.

Gibbons, R. D., Hedeker, D. and Dutoit, S. (2010) 'Advances in analysis of longitudinal data', Annual Review of Clinical Psychology, 6, pp. 79-107.

Glanz, S. A. and Slinker, B. K. (2001) *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill: New York.

Glassock, R. J. and Winearls, C. (2009) 'Ageing and the glomerular filtration rate: truths and consequences', *Transactions of the American Clinical and Climatological Association*, 120, pp. 419-428.

Goldstein, H. (1995) *Multilevel Statistical Models*, Edward Arnold: London.

Goldstein, H. (2009) 'Handling attrition and non-response in longitudinal data', *Longitudinal and Life Course Studies*, 1(1).

Gonçalves Torres, M., Cardoso, L. G., de Abrev, V., Sanjuliani, A. F. and Francischetti, E. A. (2009) 'Temporal relation between body mass index and renal function in individuals with hypertension and excess body weight', *Nutrition*, 25(9), pp. 914-919.

Gordon, J., Kopp, J. (2011) 'Off the beaten renin-angiotension-aldosterone system pathway; New perspectives on antiproteinuric theraphy', *Advances in Chronic Kidney Disease*, 18(4), pp. 300-311.

Graubard, B. I. and Korn, E. L. (1994) 'Regression analysis with clustered data', *Statistics in Medicine*, 13, pp. 509-522.

Graves, J. W. (2008) 'Diagnosis and management of chronic kidney disease', *Mayo Clinic Proceedings*, 83(9), pp. 1064-1069.

Green, P. J. (1987) 'Penalized likelihood for general semiparametric regression models', *International Statistical Review*, 55, pp. 245-259.

Hallan, S. I., Coresh, J., Astor, B. C., Asberg, A., Powe, N. R., Romundstad, S. et al. (2006) 'International comparison of the relationship of chronic kidney disease prevalence and ESRD risk', *Journal of the American Society of Nephrology*, 8, pp. 2275-2284.

Harville, D. A. (1977) 'Maximum likelihood approached to variance component estimation and to related problems', *Journal of the American Statistical Association*, 72, pp. 320-338.

Hastie, T. and Tibshirani, R. (1986) 'Generalized additive models (with discussion)', *Statistical Science*, 1, pp. 297-318.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman & Hall.

Hedeker, D. (2005) Generalized Linear Mixed Models. *Encyclopaedia of Statistics in Behavioural Science*, John Wiley & Sons Ltd.

Hedeker, D. and Gibbons, R. (2006) *Longitudinal Data Analysis*, Wiley Series in Probability and Statistics.

Hedeker, D. and Gibbons, R. D. (2006) *Longitudinal data analysis*. New York:Wiley.

Hemmelgarn, B., Manns, B., Lloyd, A., James, M., Klarenback, S., Quinn, R. et al. (2010) 'Relation between kidney function, proteinuria, and adverse outcomes', *Journal of Medical Association*, 303(5), pp. 423-429.

Hosmer, D. W. and Lemeshow, S. (1989) *Applied Logistic Regression*. 1st edn. Wiley: New York.

Hotelling, H. (1931) 'The generalization of student's ratio', *Annals of Mathematical Statistics*, 2, pp. 360-378.

Humphrey, P. (2009) *Introduction to the Practice of Statistics*. 6th edn. W. H. Freeman and Company: New York.

Huynh, H. and Feldt, L. S. (1970) 'Conditions under which mean square ratios in repeated measurement designs have exact F-distributions', *Journal of the American Statistical Association*, 65, pp. 1582-1589.

Ibrahim, H. and Weber, M. (2010) 'Weight loss: A neglected intervention in the management of chronic kidney disease', *Current Opinion in Nephrology and Hypertension*, 19(6), pp. 534-538.

Jafar, T., Stark, P., Schmid, C., Landa, M., Maschio, G., de Jong, P. et al. (2003) 'Progression of chronic kidney disease: The role of blood pressure control, proteinuria, and angiotension-converting enzyme inhibition: A patient-level meta-analysis', *Annals of Internal Medicine*, 139(4), pp. 244-252.

Jennrich, R. I. and Schluchter, M. D. (1986) 'Unbalanced repeated measures models with structured covariance matrices', *Biometrics*, 42, pp. 805-820.

Jones, C., Roderick, P., Harris, S. and Rogerson, M. (2006) 'Decline in kidney function before and after nephrology referral and the effect on survival in moderate to advanced chronic kidney disease', *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association – European Renal Association*, 21(8), pp. 2133-2143.

Jones, R. H. (1993) *Longitudinal Data with Serial Correlation: A State-Space Approach*. Chapman and Hall: London.

Jones, R. H. and Boadi-Boateng, F. (1991) 'Unequally spaced longitudinal data with AR(1) serial correlation', *Biometrics*, 47, pp. 161-175.

Kannel, W. (1985) 'Lipids, diabetes and coronary heart disease: Insights from the Framingham study', *American Heart Journal*, 110(5), pp. 1100-1107.

Kannel, W., Stampfer, M., Castelli, W. and Verter, J. (1984) 'The prognostic significance of proteinuria: The Framingham study', *American Heart Journal*, 108(5), pp. 1347-1352.

Kaufmann, H. (1987) 'Regression models for nonstationary categorical time series: Asymptotic estimation theory', *Annals of Statistics*, 15, pp. 79-98.

Keselman, H. J., Algina, J. and Kowalchuk, R. K. (2001) 'The analysis of repeated measures designs: A review', *British Journal of Mathematical and Statistical Psychology*, 54, pp. 1-20.

Khatri, C. G. (1966) 'A note on a MANOVA model applied to problems in growth curves', Annals of the Institute of Statistical Mathematics, 18, pp. 75-86.

Kidney Disease Outcomes Quality Initiative (2002) 'Clinical practice guidelines for chronic kidney disease: Evaluation, classification and stratification', *American Journal of Kidney Diseases*, 39(1), pp. S46.

Korn, E. L. and Whittemore, A. S. (1979) 'Methods for analysing panel studies of acute health effects of air pollution', *Biometrics*, 35, pp. 795-802.

Kraemer, H. C. and Blasey, C. (2004) 'Centring in regression analyses: a strategy to prevent errors in statistical inference', *International Journal of Methods in Psychiatric Research*, 13(3), pp. 141-151.

Kraut, J. and Kurtz, I. (2005) 'Metabolic acidosis of CKD: Diagnosis, clinical characteristics and treatment', *American Journal of Kidney Diseases*, 45(6), pp. 978-993.

Król, E., Rutkowski, B., Czarnak, P., Kraszewska, E., Lizakowski, S., Szubert, R., Czekalski, S., Sułowicz, W. and Wiecek, A. (2009) 'Early detection of chronic kidney disease: results of the PolNef study', *American Journal of Nephrology*, 29(3), pp. 264-273.

Kwok, O., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R. and Yoon, M. (2008) 'Analyzing Longitudinal Data with Multilevel Models: An Example with Individuals Living with Lower Extremity Intra-articular Fractures', *National Institute of Health*, 53(3), pp. 370-386.

Laird, N. M. and Ware, J. H. (1982) 'Random-effects models for longitudinal data', *Biometrics*, 38, pp. 963-974.

Laird, N. M., Lange, N. and Stram, D. (1987) 'Maximum likelihood computations with repeated measures: Application of the EM algorithm', *Journal of the American Statistical Association*, 82, pp. 97-105.

Lamb, E. J., Tomson, C. R. and Roderick, R. J. (2005) 'Clinical Sciences Reviews Committee of the Association for Clinical Biochemistry', *Annals of Clinical Biochemistry*, 42(5), pp. 321-345.

Lee, Y., Chiu, H., Su, H., Yang, J., Voon, W., Lin, T., Lai, T., Lai, W. and Sheu, S. (2008) 'Lower hemoglobin concentrations and subsequent decline in kidney function in an apprently healthy population aged 60 year and older', *Clinica Chimica Acta; International Journal of Clinical Chemistry*, 389(1-2), pp. 25-30.

Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N. and Roth, D. (1999) 'A more accurate method to estimate glomerular filtration rate from serum creatinine: A new prediction equation. Modification of diet in renal disease study group', *Annals of Internal Medicine*, 130(6), pp. 461-470.

Levey, A. S., Coresh, J., Greene, T., Stevens, L. A., Zhang, Y. L., Hedriksen, S. et al. (2006) 'Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate', *Annals of Internal Medicine*, 145(4), pp. 247-254.

Levey, A. S., de Jong, P. E., Coresh, J., El Nahas, M., Astor, B. C., Matsushita, K. et al. (2011) 'The identification, classification, and prognosis of chronic kidney disease: A KDIGO controversites conference report', *Kidney International*, 80(1), pp. 17-28.

Levey, A. S., Perrone, R. D. and Madias, N. E. (1988) 'Serum creatinine and renal function', *Annual Review of Medicine*, 39, pp. 465-490.

Levey, A. S., Stevens, L. A., Schmid, C. H., Zhang, Y. L., Castro, A. F., Feldman, H. I. et al. (2009) 'A new equation to estimate glomerular filtration rate', *Annals of Internal Medicine*, 150(9), pp. 604-612.

Levin, A., Thompson, C., Ethier, J., Carlisle, E., Tobe, S., Mendelssohn, D. et al. (1999) 'Left ventricular mass index increase in early renal disease: Impact of decline in hemoglobin', *American Journal of Kidney Diseases*, 34(1), pp. 125-134.

Leyland, A. H. (2004) 'A review of multilevel modelling in SPSS', *MRC Social and Public Health Unit*.

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

Liang, K-Y., Zeger, S. L. and Qaqish, B. (1992) 'Multivariate Regression Analyses for Categorical Data', *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1), pp. 3-40.

Lin, X. and Zhang, D. (1999) 'Inference in generalized additive mixed models using smoothing splines', *Journal of the Royal Statistical Society, Series B*, 61, pp. 381-400.

Lindstrom, M. J. and Bates, D. M. (1988) 'Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data', *Journal of the American Statistical Association*, 83, pp. 1014-1022.

Longford, N. T. (1987) 'A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects', *Biometrika*, 74(4), pp. 817-827.

Longford, N. T. (1994) 'Logistic regression with random coefficients', *Computational Statistics and Data Analysis*, 17, pp. 1-15.

Lv, J. Liu, J. S. (2013) 'Model selection principles in misspecified models', *Journal of Royal Statistical Society, Series B: Statistical Methodology*.

MacKenzie, F. (2006) 'UK NEQAS for eGFR estimations: Guidance for creatinine slopes and intercepts'.

Matthews, J. N. S. (1993) 'A refinement to the analysis of serial data using summary measures', *Statistics in Medicine*, 12, pp. 27-37.

McCullagh, P. and Nelder, J. A. (1989) *Generalized linear models*. Chapman & Hall.

McCulloch, C., Searle, S. and Neuhaus, J. (2008) *Generalized, linear and mixed models*. 2nd edn. John Wiley & Sons, Inc.: Hoboken, NJ.

McLean, R. A., Sanders, W. L. and Stroup, W. W. (1991) 'A unified approach to mixed linear models', *The American Statistician*, 45, pp. 54-64.

Menard, S. W. (1995) *Applied Logistic Regression Analysis*. 1st edn. SAGE Publications: London.

Miller, A. E., Cziraky, M. and Spinler, S. A. (2006) 'ACE inhibitors versus ARBs: comparison of practice guidelines and treatment selection considerations', *Formulary*, 41(6), pp. 274.

Molenberghs, G. and Verbeke, G. (2005) *Models for discrete longitudinal data*. Springer.

Molenberghs, G., Verbeke, G. and Demétrio, C. G. B. (2007) 'An extended random-effects approach to modelling repeated, overdispersed count data', *Lifetime Data Analysis*, 13, 513-531.

Nagelkerke, N. J. D. (1991) 'Miscellanea; A note on general definition of the coefficient of determination', *Biometrika*, 78(3), pp. 691-692.

National Collaborating Centre for Chronic Conditions (2008) 'Chronic kidney disease: National clinical guideline for early identification and management in adults in primary and secondary care', *Royal College of Physicians*, London.

Navaneethan, S. D., Yehnert, H., Moustarah, F., Schreiber, M. J., Schaver, P. R. and Beddhu, S. (2009) 'Weight loss interventions in chronic kidney disease: a systematic review and meta-analysis', *Clinical Journal of the American Society of Nephrology*, 4(10), pp. 1565.

Neuhaus, J. M. (1992) 'Statistical methods for longitudinal and clustered designs with binary responses', *Statistical Methods in Medical Research*, 1, pp. 249-273.

Neuhaus, J. M. and Jewell, N. p. (1990) 'Some comments on Rosner's multiple logistic model for clustered data', *Biometrics*, 46, pp. 523-531.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991) 'A comparison of cluster-specific and population-averaged approached for analysing correlated binary data', *International Statistical Review*, 59, pp. 25-35.

Paccagnella, O. (2006) 'Centring or not centring in multilevel models? The role of the group mean and the assessment of group effects', *Evaluation Review*, 30(1), pp. 66-85.

Perkovic, V., Verdon, C., Ninomiya, T., Barzi, F., Cass, A., Patel, A. et al. (2008) 'The relationship between proteinuria and coronary risk: A systematic review and meta-analysis', *PLos Med*, 5(10), pp. e207.

Peugh, J. L. and Enders, C. K. (2005) 'Using the SPSS mixed procedure to fit cross-sectional and longitudinal multilevel models', *Educational and Psychological Measurement*, 65(5), pp. 717-741.

Pfeffer, M., Burdmann, E., Chen, C., Cooper, M., de Zeeuw, D., Eckardt, K. et al. (2009) 'A trial of darbepoetin alfa in type 2 diabetes and chronic kidney disease', *The New England Journal of Medicine*, 361(21), pp. 2019-2032.

Pocock, S. J. (1983) *Clinical Trials: A Practical Approach*. John Wiley and Sons: New York.

Potthoff, R. F. and Roy, S. N. (1964) 'A generalized multivariate analysis of variance model useful especially for growth curve problems', *Biometrika*, 51, pp. 313-326.

Randers, E. and Erlandsen, E. J. (1999) 'Serum cystatin C as an endogenous marker of the renal function – a review', *Clinical Chemistry and Laboratory Medicine*, 37(4), pp. 389-395.

Rao, C. R. (1965) 'The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves', *Biometrika*, 52, pp. 447-458.

Rao, C. R. (1967) 'Least squares theory using an estimated dispersion matrix and its application to measurement of signals', *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 355-372.

Rao, C. R. (1996) 'Covariance adjustment and related problems in multivariate analysis', *P. R., editor, Multivariate Analysis, Academic Press*: New York, pp. 87-103.

Reddy, K., Stablein, D., Taranto, S., Stratta, R., Johnston, T., Waid, T. et al. (2003) 'Long-term survival following simultaneous kidney-pancreas transplantation versus kidney transplantation alone in patients with type 1 diabetes mellitus and renal failure', *American Journal of Kidney Diseases*, 41(2), pp. 464-470.

Roderick, P., Atkins, R., Smeeth, L., Mylne, A., Nitsch, D., Hubbard, R. et al. (2009) 'CKD and mortality risk in older people: A community-based population study in the United Kingdom', *American Journal of Kidney Diseases*, 53(6), pp. 950-960.

Roth, M., Roderick, P. and Mindell, J. (2011) 'Health survey for England 2010 volume 1, chapter 8; kidney disease and renal function', *Health and Social Care Information Centre.*

Rowe, J. W., Andres, R., Tobin, J. D., Norris, A. H. and Shock, N. W. (1976) 'The effect of age on creatinine clearance in men: A cross-sectional and longitudinal study', *The Journals of Gerontology*, 31(2), pp. 155-163.

Ruggenenti, P., Perticucci, E., Cravedi, P., Gambara, V., Costantini, M., Sharma, S. K., Perna, A. and Remuzzi, G. (2008) 'Role of remission clinics in the longitudinal treatment of CKD', *Journal of the American Society of Nephrology: JASN*, 19(6), pp. 1213-1224.

Ruspini, E. (2002) *Introduction to Longitudinal Research.* 1st edn. London: Routledge.

Rutter, M., Prais, H., Charlton-Meny's, V., Gittins, M., Roberts, C., Davies, R. et al. (2011) 'Protection against nephropathy in diabetes with atorvastatin (PANDA): A randomized double-blind placebo-controlled trial of high-vs. low dose atovastatin', *Diabetic Medicine*, 28(1), pp. 100-108.

Scottish Intercollegiate Guidelines Network (2008) 'Diagnosis and management of chronic kidney disease', *Scottish Intercollegiate Guidelines Network*, Edinburgh.

Scottish Intercollegiate Guidelines Network (2010) 'Management of diabetes. A national clinical guideline', *Edinburgh: Scottish Intercollegiate Guidelines Network*, SIGN Guideline 116.

Scottish renal registry (2011) 'Scottish renal registry report 2010', *NHS National Services Scotland*, Edinburgh.

Shek, D. T. L. and Ma, C. M. S. (2011) 'Longitudinal Data Analyses Using Linear Mixed Models in SPSS: Concepts, Procedures and Illustrations', *TheScientificWorld Journal*, 11, pp. 42-76.

Shen, J. (2011) *Additive Mixed Modelling of HIV Patient Outcomes Across Multiple Studies*, PhD Thesis, University of California, Los Angeles.

Singer, J. D. and Willett, J. B. (2003) *Applied Longitudinal Data Analysis: Modelling Change and Event Occurrence*, Oxford University Press: New York.

Singh, A., Szczech, L., Tang, K., Barnhart, H., Snapp, S., Wolfson, M. et al. (2006) 'Correction of anemia with epoetin alfa in chronic kidney disease', *The New England Journal of Medicine*, 355(20), pp. 2085-2098.

Stevens, L., Coresh, J., Feldman, H., Greene, T., Lash, J., Nelson, R. et al. (2007) 'Evaluation of the modification of diet in renal disease study equation in a large diverse population', *Journal of American Society of Nephrology*, 18(10), pp. 2749-2757.

Stevens, P. E., O'Donoghue, D. J., de Lusignan, S., van Vlymen, J., Klebe, B., Middleton, R. et al. (2007) 'Chronic kidney disease management in the United Kingdom: NEORICA project results', *Kidney International*, 72(1), pp. 92-99.

Sullivan, L. M., Dukes, K. A. and Losina, E. (1999) 'Tutorial in Biostatistics: An Introduction to Hierarchical Linear Modelling', *Statistics in Medicine*, 18, pp. 855-888.

Tiwari, P. and Shukla, G. (2011) 'Approach of Linear Mixed Model in Longitudinal Data Analysis using SAS', *Journal of Reliability and Statistical Studies*, 4(1), pp. 73-84.

Tonelli, M., Jose, P., Curhan, G., Sacks, F., Braunwald, E. and Pfeffer, M. (2006) 'Proteinuria, impaired kidney function, and adverse outcomes in people with coronary disease: analysis of a previously conducted randomised trial', *British Medical Journal*, 332(7555), pp. 1426.

Tuma, N. B. (2013) 'Hand book of data analysis: Modelling Change', *SAGE Research Methods*.

Twisk, J. (2003) *Applied Longitudinal Data Analysis of Epidemiology – A Practical Guide*. 1$^{st}$ edn. Cambridge University Press: Cambridge.

UK renal registry (2010) 'The thirteenth annual report', *UK Renal Association*, Bristol.

Umlauf, N., Kneib, T., Lang, S. and Zeileis, A. (2015) 'Package 'R2BayesX' An R Interface to estimate structured additive regression (STAR) models with BayesX'. *CRAN*. https://cran.r-project.org/web/packages/R2BayesX/R2BayesX.pdf. [Accessed on: 10.02.2014].

US Renal Data System (2011) 'USRDS 2011 annual data report: Atlas of chronic kidney disease and end-stage renal disease in the United States', *National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases*, Bethesda, MD, USA.

Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E. and Tuero-Herrero, E. (2011) 'Selecting the best unbalanced repeated measures model', *Behavioural Research*, 43, pp. 18-36.

Vassalotti, J. A., Stevens, L. A. and Levey, A. S. (2007) 'Testing for chronic kidney disease: A position statement from the National Kidney Foundation', *American Journal of Kidney Diseases*, 50(2), pp. 169-180.

Verbeke, G., Fieuws, S., Molenberghs, G. and Davidian, M. (2012) 'The analysis of multivariate longitudinal data: A review', *Statistical Methods in Medical Research*, 23, pp. 42.

Verbeke, G., Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. Springer-Verlag: New York.

Vonesh, E. F. and Chinchilli, V. M. (1997) *Linear and nonlinear models for the analysis of repeated measurements*. Marcel Dekker, Inc.

Wahba, G. (1990) *Spline models for observational data*. SIAM: Philadelphia.

Wahi, M., Keane, W., Kasiske, B., Svendsen, K. and Grimm, R. (1997) 'Proteinuria is a risk factor for mortality over 10 years of follow-up. MRFIT research group multiple risk factor intervention trial', *Kidney Internatіonal*, 63, pp. S10.

Ware, J. H. (1985) 'Linear models for the analysis of longitudinal studies', *The American Statistician*, 39, pp. 95-101.

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

Wedderburn, R. W. M. (1974) 'Quasilikelihood functions, generalized linear models and the Gauss-Newton method', Biometrika, 61, pp. 439-447.

White, S. L., Polkinghorne, K. R., Cass, A., Shaw, J. E., Atkins, R. C. and Chadban, S. J. (2009) 'Alcohol consumption and 5-year onset of chronic kidney disease: the AusDiab study', *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association – European Renal Association*, 24(8), pp. 2464-2472.

Wishart, J. (1938) 'Growth rate determination in nutrition studies with the bacon pig, and their analysis', *Biometrika*, 30, pp. 16-28.

Wood, S. (2006) *Generalized Additive Models: An Introduction with R*. Chapman & Hall/ CRC.

Wood, S. (2015) 'Package 'mgcv': Mixed GAM computation Vehicle with GCV/AIC/REML Smoothness Estimation'. *CRAN.* https://cran.r-project.org/web/packages/mgcv/mgcv.pdf [Accessed on: 10.02.2014].

Wood, S. and Scheipl, F. (2015) 'Package 'gamm4': Generalized additive mixed models using mgcv and lm4'. *CRAN.* https://cran.r-project.org/web/packages/gamm4/gamm4.pdf [Accessed on: 10.02.2014].

Yudkin, J., Forrest, R. and Jackson, C. (1988) 'Microalbuminuria as predictor of vascular disease in non-diabetic subjects. Islington diabetes survey', *Lancet*, 2(8610), pp. 530-533.

Zeger, S. L. (1988) 'Commentary', *Statistics in Medicine*, 7, pp. 161-168.

Zeger, S. L. and Karim, M. R. (1991) 'Generalized linear models with random effects: A Gibbs sampling approach', *Journal of the American Statistical Association*, 86, pp. 79-86.

Zeger, S. L. and Liang, K. Y. (1992) 'An overview of methods for the analysis of longitudinal data', *Statistics in Medicine*, 11, pp. 1825-1839.

Zeger, S. L., Liand, K. Y. and Albert, P. S. (1988) 'Models for longitudinal data: A generalized estimating equation approach', *Biometrics*, 44, pp. 1049-1060.

Zhang, p., Song, P. X. K., Qu, A. and Greene, T. (2008) 'Efficient Estimation for Patient-Specific Rates of Disease Progression Using Nonnormal Linear Mixed Models', *Biometrics*, 64, pp. 29-38.

Zimmerman, D. L. (2000) 'Viewing the correlation structure of longitudinal data through a PRISM', *The American Statistician*, 54, pp. 310-318.

(2008) *NICE Clinical Guideline 73. Early identification and Management of Chronic Kidney Disease in Adults in Primary and Secondary Care.*

(2010) *UK National Kidney Federation – Glossary of Renal Terms 2007.*

# 9 Appendix: Publications

Yarkiner, Z., O'Neil, R. and Hunter, G. (2012) 'Developing longitudinal models for monitoring chronic diseases in computerised GP records', *6th CSDA International Conference on Computational and Financial Economics (CFE 2012) and 5th International Conference of the ERCIM (European Research Consortium for Informatics and Mathematics) Working Group on Computing and Statistics (ERCIM 2012)*, Oviedo: Spain.

Yarkiner, Z., O'Neil, R. and Hunter, G. (2013) 'Predictors of decline in eGFR in patients with CKD in the UK: Findings from the longitudinal study of routinely collected GP records', *2nd International Conference and Exhibition on Biometrics and Biostatistics*, Chicago: USA.

Yarkiner, Z., O'Neil, R., Hunter, G. and Bidgood, P. (2013) 'Application of Linear Mixed Models to Routinely Collected General Practice Data: A case study in chronic kidney disease (CKD) in the UK', *7th Annual International Conference on Annual International Conference on Mathematics Education & Statistics Education, Mathematics and Statistics*, Athens: Greece.

Yarkiner, Z., Hunter, G., O'Neil, R. and de Lusignan, S. (2013) Application of Mixed Models for Investigating Progression of Chronic Disease in a Longitudinal Dataset of Patient Records from General Practice', *Journal of Biometrics and Biostatistics*, S9: 001.

Yarkiner, Z., Hunter, G., O'Neil, R. and de Lusignan, S. (2013) 'Analysis of Complex, Routinely-obtained Longitudinal Data from Medical General Practice Records: A case study on the progression of chronic kidney disease', *S.Co. 2013 Complex Data Modelling and Computationally Intensive Statistical Methods for Estimation and Prediction, Politecnico di Milano*, Milan: Italy.