UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

University of Pretoria
Faculty of Health Sciences
School of Health Systems and Public Health

A Comparison of Survival Analysis, Threshold Regression and Linear Mixed Models in a Longitudinal Diabetes Clinic Study (2009 – 2013) at Kalafong Hospital with Nephropathy as Outcome

In Partial Fulfilment for the degree
M.Sc (Biostatistics)

Author: Lynda Olinger

Student Number: 12352579

Contact Details:
357 Chelsea Crescent, Garsfontein
Tel: 012 993-2495
Cell: 074-863-0036
Email: olinger274@gmail.com

Supervisor: Mrs L.Dzikiti
Co-Supervisor: Prof P.Rheeder

29 May 2014

Abstract

Background: This study compares three methodologies appropriate for the analysis of longitudinal time-to-event data. The Cox model is well researched and frequently used. Threshold regression, however, is relatively new and there are few articles describing its application in biomedical statistics. A linear mixed model provides an alternative interpretation of a continuous outcome rather than time to an event. A longitudinal study of the time to onset of diabetic nephropathy, a common complication of Diabetes Mellitus, is used to compare the three models with respect to their explanatory and predictive abilities and utilitarian value to researchers.

Methods: The study entails a secondary data analysis of 1160 retrospective patient records, collected at a diabetic clinic at Kalafong Hospital, Pretoria. Model selection was based on current literature, backward elimination of insignificant variables ($p>0.2$) and the Akaike and Bayesian Information Criterion. Survival and hazard functions and ratios were determined for the survival data. Risk categories in the Cox model evaluated discrimination, while threshold regression predicted survival probabilities for specific patient profiles. The linear mixed model predicted albumin-creatinine ratio values, a marker for the diagnosis of diabetic nephropathy.

Results: The Cox model, stratified by glucose control, gender, hypertension, type of diabetes and smoking status, had an AIC of 81 and was the most parsimonious model. Threshold regression, with an AIC of 1428, indicated duration of diabetes as a significant factor in the process of health deterioration. Individual variation in weight and total cholesterol amongst patients was accounted for by the linear mixed model, with an AIC of 3755.

Conclusion: All three regression models provided valuable insight into underlying risk factors of diabetic nephropathy and should form part of a multi-faceted approach to analysing longitudinal survival data.

In loving memory of my brother, Billy Brenchley

His support, encouragement and belief in me will never be forgotten.

Contents

1. Introduction

The purpose of statistical modelling is to extract meaningful information from data collected in surveys and experiments. Important facets of modelling include description of the data, an explanation of relationship between outcome and predictor variables, and the prediction of outcomes. The choice of model is guided by the type of data collected, and in this study longitudinal survival, or time-to-event, data is analysed. Longitudinal studies, with repeated measurements upon the same subject over time, have certain challenges in the statistical analysis of the data. These challenges include correlation between the successive measurements, time dependent covariates and a high dropout rate of patients resulting in missing data. The models discussed in this paper take these challenges into account. Three models, namely the Cox Proportional Hazards (PH), a threshold regression and a linear mixed model (LMM), are compared in the analysis of a longitudinal study of the time to onset of diabetic nephropathy, a common complication of Diabetes Mellitus (DM).

The Cox PH regression model is a well-known and frequently used survival analysis method of analysing time-to-event data. In contrast, threshold regression is relatively new and less has been written of its application in biomedical statistics. It has proven valuable in providing a deeper insight into underlying causes of a stochastic process with a first hitting time or threshold value, such as health deterioration.[1] A linear mixed model (LMM) has a different approach to the analysis of longitudinal data. The relationship between a continuous dependent variable and independent predictors is quantified on a population level, while taking individual variation of patients or subjects into account.[2]

In the research data used to compare the three models, it is the progression of diabetes mellitus (DM) and the onset of diabetic nephropathy (DN) that is of interest to medical practitioners. Potential risk factors, the time to onset of DN as well as defining characteristics of the disease's progression are important considerations. Although measuring different aspects of the disease, the three approaches to analysing survival and longitudinal data proposed in this study provide such analysis. Survival analysis can be used to analyse and predict the probability and time to onset of DN, while threshold regression will provide a more detailed analysis of this same outcome. LMMs, on the other hand, can be used to predict values of an outcome variable, such as the albumin-creatinine ratio (ACR), a marker for DN, while taking into account independent covariates.

It is recognised that survival analysis, threshold regression and mixed models measure different outcomes in the data, and that a direct comparison is limited. However, there is value in evaluating the usefulness of these three approaches to analysing longitudinal survival and continuous data. Comparison of these models will thus be based upon how well each model describes the data in the context of its particular outcome, as well as the predictive power of the models and their ability to produce valid, effective information for the clinic and broader medical community

2. Literature Review

A brief background of DM and DN is followed by an overview of research in survival analysis, threshold regression and LMMs in the analysis of longitudinal data. Various methods of analysing data with missing observations are also noted.

2.1 Diabetes Mellitus and Diabetic Nephropathy

Diabetes Mellitus (DM) is a lifestyle disease that was responsible for 5.1 million deaths in the world in 2013, and is estimated to currently affect 382 million people worldwide. More than half the adults in the world with diabetes are between 40 and 59 years of age.[3]

The prevalence of diabetes worldwide has increased as a result of an ageing population and lifestyle changes associated with urbanisation and westernisation. This is evident in the increased number of people suffering from type 2 diabetes, which is often associated with obesity. The International Diabetes Foundation (IDF) estimates a 109% increase in the next 20 years in the number of adults with DM in Africa, from approximately 19.8 million in 2013 to 41.5 million in 2035. Furthermore, 76% of deaths caused by DM in Africa were in people younger than 60 years. In particular, South Africa is estimated to currently have a national prevalence of 8.3%, or 3.2 million people, with DM. [4]

DM has a high complication burden, higher in Africa than in the developed world.[5] These complications include macrovascular disease (coronary artery disease, peripheral vascular disease and stroke) and microvascular damage (diabetic retinopathy and nephropathy).[6] In particular, DN is a clinical syndrome which includes symptoms such as persistent albuminuria (urinary albumin excretion rate greater than 300mg / 24hr), a decline in glomerular filtration rate (GFR) and hypertension.[7] A raised level of urine albumin is a key marker in detecting a decline in renal function and development of DN.

The albumin-creatinine ratio (ACR) provides threshold values defining the progression of DN, and is obtained by dividing the level of urine albumin (mg/Dl) by the urine creatinine level (g/Dl).[8] A positive correlation exists between the 24 hour albumin excretion rate and the ACR, and both can be used to diagnose raised levels of urinary protein. ACR has the advantage that a 24 hour collection of urine is not necessary, and two random urinary samples can be as reliable when screening for micro- and macro-albuminuria in Type 1 pregnant diabetics.[9]

Threshold values denoting normal, micro- and macro-albuminuria are influenced by gender, and are defined as[10]:

Normal: ACR <2.5mg/mmol (men) and <3.5mg/mmol(women),

Micro-albuminuria: 2.5≤ACR≤30mg/mmol (men), 3.5≤ACR≤30mg/mmol (women),

Macro-albuminuria (overt nephropathy): ACR>30mg/mmol

DN is a major cause of end-stage renal disease, the most common cause of death amongst diabetic patients.[11] Rarely occurring in the first five years of diabetes, it is more commonly seen 10 to 15 years after the onset of disease. Normalising glycaemic levels, strictly controlling blood pressure, treating dyslipidaemia and administering angiotensin converting enzyme (ACE) inhibitors or angiotensin receptor blockers (ARB) can decrease the ACR, and consequently slow down the progress of DN. Some studies question whether aggressive glycaemic control is associated with improved mortality risks in DM patients, and recommend an individualised approach to glycaemic control in these patients.[12,13] It is thus important for clinicians to know the risk factors associated with DN and to identify the population at risk before GFR declines and microalbuminuria is present.[14]

Many risk factors are associated with DN and include arterial blood pressure, HbA1c levels, lipid profiles (HDL, LDL, triglycerides, total cholesterol), waist-hip ratio, age, duration of DM, body mass index (BMI), baseline HbA1c, baseline urine ACR, hypertension, gender, smoking status, race and glomerular filtration rate (GFR) amongst others.[14-17] A study by Gall et al.[13] on the risk factors for the development of DN indicated the presence of retinopathy, increased serum cholesterol concentration, haemoglobin A_1C , age, male gender and an increased baseline log urinary albumin excretion rate as significant risk factors. An increased risk of premature death for diabetic patients suffering from microalbuminuria (urinary albumin excretion rate between 30 and 299 mg/24h) compared to diabetic patients with normal albuminuria, was also observed in this study. The known duration of diabetes, arterial blood pressure, serum creatinine concentration, pre-existing coronary heart disease and a history of smoking were not significant risk factors for the development of DN in the study.[13] A family history of DN as well as the presence of albuminuria have been associated with an early decline in the functioning of GFR, an early indicator of DN.[14] Studies have also shown that race and ethnicity can play a significant role in the frequency of renal disease, with a higher prevalence occurring in Native Americans, Mexican Americans and African Americans.[18,19] Although many prevalence studies have been conducted on DN in black Africans [20-23], very few studies have focused on a longitudinal clinical follow-up of DN in an African population. This study is therefore relevant in this regard.

2.2 Survival Analysis

Survival analysis is a collection of statistical methods used to analyse data for which the outcome variable is time to an event occurring. The event is not limited to death, and can also be the onset of a disease, such as DN.[24] Both the Cox PH and threshold regression models are suitable for analysing survival data, but it is possible that threshold regression may add a further valuable dimension in explaining the data at hand.

If *T* = survival time, the probability of survival beyond time *t* is represented by the survivor function, *S(t)* = *P(T>t)*. In contrast, the hazard function measures the instantaneous potential of the event (i.e. death/failure/disease) occurring at time *t* per unit time, given that the individual has survived to time *t,* and can be expressed as:[24]

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t}.$$

The Cox PH model describes a hazard at time *t* as the product of a baseline hazard (a function of *t*) and an exponential component of explanatory variables that are time independent. The hazard function, which measures the probability that "failure" occurs at time t_j, given that the event has not occurred by time t_{j-1}, is given by:[24]

$$h(t, X) = h_0(t)e^{[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k]},$$

where $X = (X_1, X_2, \ldots X_k)$ are *k* explanatory variables, t represents time and $\beta_{i,}$ $i = 1,2 \ldots k$ are *k* coefficients of the explanatory variables.[24]

The hazard ratio comparing two individuals or groups, *X* and *X'*, can be formulated as:

$$HR(X) = \frac{h(t,X\prime)}{h(t,X)} = \frac{h_0\prime(t)}{h_0(t)} exp\left[\sum_{i=1}^{k} \beta_i (X_i{}' - X_i)\right]$$

The proportional hazards model assumes that this hazard ratio is constant over time and has the advantage of not having to specify a baseline hazard function, $h_0(t)$.[24] Under the proportional hazards assumption, the hazard ratio is reduced to:

$$HR(X) = \frac{h(t,X\prime)}{h(t,X)} = exp\left[\sum_{i=1}^{k} \beta_i (X_i{}' - X_i)\right]$$

It is only necessary to estimate the β's in the exponential part of the equation to determine the effect of the explanatory variables on time to the event. [24]

However, in a repeated measures study, some covariates are time dependent and this violates the proportional hazards assumption. Such data should rather be analysed with the extended Cox model, which allows for time-dependent covariates.[24]

The formula for the extended Cox model is given by:[24]

$$h\big(t, X(t)\big) = h_0(t)\, exp\left[\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t)]\right]$$

where $X(t) = X_1, X_2, \ldots X_{p1}, X_1(t), X_2(t), \ldots X_{p2}(t)$ are p_1 time-independent and p_2 time-dependent explanatory variables, and β_i and $\delta_j$are the corresponding coefficients for the explanatory variables.

The hazard ratio for the Cox extended model can be written as:[24]

$$HR(t) = \frac{h(t,X\prime(t))}{h(t,X(t))} = exp\left[\sum_{i=1}^{p_1} \beta_i (X_i{}' - X_i) + \sum_{j=1}^{p_2} \delta_j\ [X_j{}'(t) - X_j(t)]\ \right]$$

where $X(t)$ and $X'(t)$ are two sets of predictors, with p_1 time independent and p_2 time dependent predictors, and β_i and $\delta_j$are the corresponding coefficients.

An important consideration in the analysis of survival data is that the data is often censored. Individuals may experience the event before the study begins, but still be included in the study, which results in left-censored data. If these individuals are deliberately excluded from the study because they have already experienced the event of interest beforehand, the data is said to be left truncated, and no further information is collected from these patients.[25] Right censoring of data occurs if participants drop out of the study or the study is concluded and participants have not yet experienced the event, i.e. the time of event in the future is unknown. Censoring complicates analysis of survival data, especially left-truncated or delayed entry data. Very often left truncated data will introduce a bias.[25,26] For example, patients presenting with DN at the first visit may be excluded from the study, and their shorter survival times will not be taken into account in the calculation of overall time to onset of DN. To accommodate data that is highly censored, parametric survival models and joint modelling have been suggested.[24,26] Although robust, especially when data is heavily censored, parametric models should only be used when the particular distribution of the hazard function is known, for example Weibulll or negative binomial.[27] In comparison, the Cox regression models do not require knowledge of the distribution of the hazard function, and are also robust in a variety of data settings.[24]

Literature on the Cox PH model, and its variations, is widely available. Topics such as non-proportional hazards, [28,29] time-dependent covariates, [12,30,31] parametric survival models [32,33] and external validation of a Cox prognostic model [34-36] have been well researched. In a study[12] examining the association between glycaemic control and extended haemodialysis survival in diabetic patients, both the traditional and time-dependent Cox models were utilised. The time-dependent model was slightly more sensitive than the traditional model, and highlighted a greater risk at the extreme ends of HbA1c levels. Another study by Boberg, et al.[37] indicated that the time-dependent model had greater predictive power in the prognosis of primary sclerosing cholangitis than the traditional model.

2.3 Linear Mixed Models

Mixed modelling refers to statistical models that have both a fixed and random effects component. Fixed effects include continuous or categorical covariates, such as baseline measurements and gender, for the whole population of analysis units. However, a model with random effects takes into account the additional randomness present in the data due to effects which vary across multiple visits for each patient, such as weight, creatinine and HbA1c levels. The fixed effects coefficients measure the magnitude of the fixed covariates' effects on the dependent outcome. In contrast, the random effects covariates in the model represent deviations, particular to a specific patient, from the overall fixed effects of the whole population.[2,38]

It is not possible to determine coefficients for the random effects, but rather a Best Linear Unbiased Predictor (BLUP) is predicted. BLUPs can be defined as the expected value of the random effects, for a specific level of a random factor, given the observed outcome values. BLUPs are not fixed parameters, but are considered random variables with a multivariate normal distribution. They are smaller than the estimated effects would be if the random factors were considered fixed, and are therefore also known as shrinkage estimators. Also sometimes referred to as empirical BLUPs (EBLUPs), they are based upon the estimated variance and covariances of the variables.[2] Different covariance structures can be more finely selected for LMMs and are summarised in table 1 below.[39]

Table 1 Covariance Structures for a Linear Mixed Model

Covariance Structure	
Identity	Equal variance for all random effects, covariances are zero
Independent	Unique variances for all random effects, covariances are zero
Exchangeable	Equal variance for all random effects, common pairwise covariance
Unstructured	No restrictions on the variances and covariances

A general formula for the measurement of the outcome variable, *Y*, in a LMM for patient *i* at time *t,* can be defined as follows:[2]

$$Y_{ti} = \beta_1 X_{(1)ti} + \beta_2 X_{(2)ti} + \cdots + \beta_p X_{(p)ti} \quad + \quad u_{1i} Z_{(1)ti} + u_{2i} Z_{(2)ti} + \cdots + u_{qi} Z_{(q)ti}$$

(Fixed) *(Random)*

X and *Z* are *p* fixed and *q* random covariates respectively. $\beta_1, \beta_2, \ldots, \beta_p$ and $u_{1i}, u_{2i}, \ldots, u_{qi}$ are *p* fixed and *q* random effects, associated with *X* and *Z*, respectively.[2]

In matrix form, a LMM can be expressed as follows:[2]

$$Y_i = X_i \beta + Z_i u_i + \varepsilon_i$$

$$u_i \sim N_q(0, \varphi)$$

$$\varepsilon_i \sim N_{n_i}(0, \sigma^2 \gamma_i)$$

where

Y_i= n_i × 1 response vector for observations of the ith subject

X_i = n_i × p design matrix for the p fixed covariates for the observations of the ith subject

β= p × 1 vector of fixed-effect parameters

Z_i= n_i × q design matrix for the q known covariates of the random effects for the ith subject

u_i = q × 1 vector of random effects for the ith subject

ε_i = n_i × 1 vector of errors for observations in the ith subject

φ = q × q covariance matrix for the random effects in u_i

$\sigma^2\gamma_i$ = n_i × n_i covariance matrix for the errors of the ith subject

Furthermore, one can distinguish between a ‘random intercept only’ model and a ‘random coefficients’ model. The former describes a deviation for a given subject from the overall fixed intercept of the model, but individual slopes remain parallel to the overall model. A random coefficients model includes a random intercept as well as a random slope, and represents an individual deviation from the intercept and slope of the overall model.[2]

A primary assumption of a fixed effects model is violated when analysing longitudinal data, namely that all the observations are independent of each other. Repeated measurements on an individual are correlated, and thus a mixed model with a random effects component is more appropriate.[40] General LMMs are an extension of the linear regression model, but include random effects. Generalised LMMs can be used when the data is not continuous or normally distributed, and adds random effects to generalised linear models such as logistic or Poisson regression models.[41] It is assumed that the residuals of a LMM are normally distributed, but are not necessarily independent or constant.[2]

Mixed models have been shown to be a better analysis option for longitudinal data than more traditional statistical methods such as ANOVA, ANCOVA and MANCOVA for several reasons. Time dependent covariates can be included in mixed models and the data does not have to be balanced or collected at even points. In addition, covariance structures of the data are more accurately modelled. However, limitations of LMM's include the assumptions of multivariate normality of the random terms and that missing data are MAR which is not the case with many longitudinal studies.[42]

The residual vector of ε_i is assumed to have a normal distribution with zero mean, and covariance matrix $\sigma^2\gamma_i$. Several covariance structures are also available for this covariance matrix, including a diagonal structure if residuals within measurements of a patient are independent of each other, a compound symmetry structure if equal correlation of the residuals can be assumed, and a first-order autoregressive structure where residuals closer in time to each other have higher correlation than measurements further apart in time. It is assumed that the error vectors of patients are independent of each other, as well as independent of the BLUP vectors u_i.[2] In this study the diagonal residual matrix is assumed.

LMM have been studied in varied contexts [43-46] and research on the development of diagnostics for LMMs is well recorded.[47-49] In a further article, Zucker et al. compared the power of summary measures, mixed model and survival analysis in an analysis of a repeated-measures trial, and observed a significant decrease in efficiency with survival-based analysis compared to when continuous measurements are analysed. [50]

2.4 Threshold Regression

Threshold regression models an underlying health process, not always observable, that tends towards a threshold value. It is based on the theory that an individual's health follows a stochastic process and at a defined point, can reach a point of failure. This point is the threshold, also known as a first-hitting time.[1] An example of such a process is the progression of nephropathy in a diabetic patient from first diagnosis to end stage renal disease. Threshold values for urinary albumin/creatinine ratios (ACR) used to diagnose the development of nephropathy, as mentioned earlier, are:[10]

Normal: ACR <2.5mg/mmol (men) and <3.5mg/mmol(women),

Micro-albuminuria: 2.5 – 30mg/mmol (men), 3.5 – 30mg/mmol (women),

Macro-albuminuria (overt nephropathy): >30mg/mmol.

There are two components to a threshold regression model, namely a baseline value ($\ln(y_0)$) representing the initial health status of a patient, and mu (μ) which measures the change in the patient's health status over time. Covariates associated with $\ln(y_0)$ are typically baseline values, whereas covariates associated with mu can represent both baseline values and covariates affecting the health process.[51]

The model can be formulated as follows:[52]

$$S = \inf\{t: Y(t) \in B\},$$

$$\ln(y_0) = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + .. + \gamma_k Z_k$$

$$\text{mu}(\mu) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 .. + \beta_k Z_k$$

where Y(t) is the stochastic process of health deterioration, B is the threshold or medical endpoint denoting a critical health or disease state that is triggered when this value is reached, and S is the time it takes for the process to first reach this threshold level. Z_1, Z_2, …, Z_k are k covariates affecting the initial health status($\ln(y_0)$) and drift process (μ), and γ_i and β_i , $i = 1$ …. k, represent regression coefficients for $\ln(y_0)$ and μ respectively.[52,53]

It is also possible to apply threshold regression to a latent health process, where the transition of a patient's health to a final point of failure is less definable. For instance, the development of lung cancer is unobservable and too complex to be measured in stages. The endpoint of such a latent process is the diagnosis of primary lung cancer. In such cases, a proxy function for the latent process can be constructed from the covariates in the model. This proxy function, also known as a marker process, follows the progress of the latent, or parent, health process and provides some insight into causal factors affecting the health process.[1,54]

Threshold regression has several advantages. It does not require an assumption of proportional hazards and allows for the analysis of data unevenly spaced in time, which commonly occurs with longitudinal data.[1,53] A further significant advantage of threshold regression is the insight it can provide into causal factors and underlying hazard patterns in the data.[52] At best, the Cox PH model provides only a single hazard ratio for the time period under study. However, threshold regression allows the change in hazard ratios over time to be observed.[53]

Several papers explore the application of threshold regression to survival data.[1,52,54,55] A case study by the same authors highlights the benefits of threshold regression compared to Cox regression, finding it a feasible alternative to the Cox PH model.[53] Additionally, an earlier study in 2007 by Whitmore and Su provides a detailed description of threshold regression modelling low birth weights. [56]

2.5 Missing Data

The long-term nature of a longitudinal study often results in subjects missing occasional examinations or dropping out of the study completely. Bias is introduced when the dropouts or missing observations differ significantly from those observations remaining in the study. For instance, in medical studies, it is often the respondents who are more seriously ill that drop out of a study. If this is not taken into account, results may incorrectly overestimate the effectiveness of medical treatment. Too often ad-hoc measures are applied to data with missing observations, without considering the reasons underlying the absence, resulting in biased parameter estimations.[57]

Rubin classified three types of missing data mechanisms.[58] Missing Completely at Random (MCAR) data occurs when the probability that a value is missing is random and does not depend on any observation in the data, such as an administration decision not to measure a certain health-related covariate. When observed values in the data influence the probability of a value being missing, the data is classified as Missing at Random (MAR). An example of this is when a participant refuses to answer questions on sexual behaviour because of religious beliefs. If religious belief is included in the study as a covariate, it can be used to predict the probability of missing values in other questions. The final classification is one of data “Not Missing at Random (NMAR)” where the probability of a value missing depends on the missing values themselves, for example, adolescents not stating how many cigarettes they smoke in a study on tobacco use amongst adolescents.[59]

Many methods exist to estimate missing values. Single imputation methods, where missing values are replaced by a single value, include Last Observation Carried Forward (LOCF), mean substitution, regression substitution and complete case (CC) analysis. However these methods are not reliable and often result in underestimated standard errors, an increased Type 1 error and biased estimates.[60] Unfortunately, many researchers use these less reliable methods of data imputation because of their ease of implementation and ignorance of other more statistically accurate methods that are available. Multiple imputation (MI) on the other hand, replaces each missing value with two or more imputed values, resulting in multiple complete datasets. These datasets are then analysed and the results combined to obtain a final estimate. MI not only portrays the uncertainty caused by missing observations more accurately than simple imputation, but also introduces randomness into the model.[41,60,61] Although most MI assumes data is multivariate normal, it is fairly robust and can be applied to moderately non-normal data and data with a high volume of missing observations if N>200.[41,60] MI provides valid results for all three missingness mechanisms. However, if the data is MAR, those variables that affect the missing data must be included in the imputation model, and if the data is NMAR, it should be correctly modelled in the imputation procedure.[41,61]

Other authors have proposed Full Information Maximum Likelihood (FIML) as an alternative to MI in using observed responses to estimate missing values.[59,62] Both methods have advantages and are asymptotically efficient. MI is easier to apply than FIML and although computationally more intense than some of the simpler methods of imputation, software exists to assist with these computations. [60,61] FIML, however, provides a consistent answer every time the analysis is performed, compared to MI which results in a different parameters, standard errors and test statistics each time it is run. [62]

Imputation of missing values is only reliable if a moderate proportion of data is missing. Marshall et al.[63] compared 5 different methods of imputing missing covariates in a Cox PH model, and concluded that CC analysis and single imputation are only reliable if 10% or less of the data is missing. Even if as little as 5% of the data is missing, simple methods of imputation, such as LOCF, listwise deletion, casewise deletion and mean substitution should be avoided or used with caution.[59] Furthermore, MI produced biased results with 50% MAR data.[63]

Many of the imputation procedures assume data is either MCAR or MAR. With MCAR data, no bias is present in the available data, but the analysis will have reduced power and larger standard errors because of a smaller sample size.[59] However, if the data is non-ignorable i.e. NMAR, a modelling option is a pattern mixture model. Patterns of missingness in data are observed, for example dropouts after first measurement never return. Parameters are varied for each of the patterns and estimates obtained are combined and weighted according to the number of observations in each pattern. Further identifying restrictions are required to estimate the missing values and parameters and include assumptions on how information on missing observations is obtained e.g. CC, available case or neighbouring case missing values.[41]

Siddique, et al.[57] also emphasise that it is important to understand why the data is missing and apply a model that is consistent with this reason. Any analysis of data with missing observations must preserve the relationship between the variables by taking into account the subject’s responses prior to dropout or absence. It must also correct any non-response bias occurring in the data as well as take the uncertainty caused by missing values into account. They propose three valid approaches that accomplish this with longitudinal data – a mixed effects regression model, multiple imputation of missing values, and a pattern-mixture model.[57,64]

3. Aim and Objectives

The aim of this dissertation is to compare the three methods of analysing survival and longitudinal data i.e. survival analysis, threshold regression and linear mixed models. The primary objective is to determine the predictive power and utilitarian value of the three methods to researchers. Secondary objectives include identifying significant covariates associated with ACR levels and the time to onset of DN in diabetic patients and, where relevant, evaluating different methods of addressing missing observations in a dataset.

4. Methods

4.1 Study Design and Setting

This study is a secondary analysis of data that was routinely collected from 2009 to 2013 at the Diabetic Clinic at Kalafong Hospital, a tertiary public hospital in the suburb of Atteridgeville, west of Pretoria, South Africa. The collection of data from patients is an ongoing observational cohort study.

4.2 Study Population

The study population consists of diabetic patients visiting the Diabetic Clinic at Kalafong Hospital. The clinic serves mostly middle to lower socio-economic groups with no medical insurance, and the majority of the patients are African. The Diabetes Clinic sees an average of 800 diabetic patients a year, who are usually referred to the clinic because of a diagnosis of diabetes, or the presence of diabetic complications. Diabetic patients presenting with nephropathy at the first visit are included in the database, thus limiting the bias that could have occurred with their exclusion.

4.3 Measurements

The study consisted of several repeated measurements on patients. Each patient is allowed four basic visits per year, and each visit has a different focus – feet, cardiovascular, optometry and a final general visit. At each visit to the clinic serum glucose, arterial blood pressure and weight were measured. A urine dipstick was also taken. Biannual laboratory measurements included haemoglobin A1C concentration, serum creatinine and potassium, lipid profiles (total cholesterol, HDL, LDL and triglycerides) and urine ACR. If the patient required further medical treatment, data for up to seven visits were recorded.

ACR is the primary outcome of this analysis, and is a continuous variable, used to also categorise patients according to their level of albuminuria viz. normal, micro- and macro-albuminuria. Threshold levels defining the three categories of albuminuria have been mentioned earlier in this dissertation (see Par 2.1 and 2.4). Other variables taken into consideration in the study were height, GFR, age, duration of diabetes and categorical variables gender, race, type of diabetes, hypertension and smoking status. Another categorical variable, glucose control, was based upon HbA1c levels, with HbA1c<8% considered good control, 8%≤HbA1c≤11% moderate control and HbA1c>11% poor glucose control.

4.4 Data Analysis

Permission to proceed with research was obtained from the Academic Programme Committee of the School of Health Systems and Public Health and the Student Ethics Committee of the Faculty of Health Sciences, University of Pretoria. Specific permission to use the data was obtained from the Chief Executor Officer of Kalafong Hospital. The data was collected under the supervision of Professor D. van Zyl from the University of Pretoria.

All available clinic records were included in the analysis. Informed consent had been obtained from the patients upon attending the clinic. No identifying information of patients was included in the final report. STATA v.12 was used for all statistical analyses.

The original dataset comprised several repeated measurements on 1160 patients, routinely collected over a period of five years (2009 – 2013) at the Diabetic Clinic at Kalafong Hospital in Pretoria. A few outliers were observed and investigated further, and descriptive statistics were determined for the whole dataset. Since the variable ACR was very skew, the change in median ACR values, rather than the mean, was calculated over the five year period. Changes in mean HbA1c from 2009 to 2013 were also recorded. Missing values were evaluated to determine whether the data was MCAR, MAR or NMAR, and analysed accordingly. All variables were tested for normality, by means of histograms and normal quantile graphs. Non-normal variables were transformed with a log transformation.

Significant associations between categorical variables, level of albuminuria (normal, micro- and macro- as defined in par 2.4) and glucose control, gender, type of diabetes, hypertension, race, and smoking status, were tested with a chi-squared test. Student's t tests were used to test whether significant differences existed between the HbA1c and ACR of males and females, hypertensive and non-hypertensive patients, and type 1 and 2 patients.

Many risk factors are associated with DN, and from 30 possible variables identified in the literature,[14-17] significant covariates were initially selected by means of univariate Cox and threshold regression models if P<0.25. A top-down approach was used with the LMM, and variables known to be associated with the development of DN were included as fixed effects before determining a random effects structure.[2] The Akaike Information Criterion

(AIC) and Bayesian Information Criterion (BIC) identified which models within a method were relatively superior. Where possible, likelihood ratio tests confirmed the inclusion of covariates. The final model was determined by means of backward elimination of these covariates, removing variables with a significance level $P > 0.15$. Some variables with higher *P* values were retained because excluding them from the model raised the AIC and BIC scores significantly, implying a poorer fit of the model.

Evaluating the performance of a model is essential if the model is to be of value in predicting outcomes, which in a clinical setting are often associated with decisions concerning treatment and surgery. Ideally the models should be tested on independent external datasets, but internal validation was provided by splitting the dataset into a development (67%, n=10 283) and validation dataset (33%, n=5136). The observations were sorted according to hospital identification number. The first and last third of the patients were allocated to the development dataset, and the middle third to the validation dataset. There was very little difference in the summary and demographic statistics for each of these groups, confirming the similarity of patients in each group. For each approach the model was derived from the development dataset, and tested in the validation dataset. Discrimination, the extent to which a model can accurately separate high and low risk patients, was evaluated by creating risk categories based on predicted hazard ratios.[36]

4.4.1 Cox Proportional Hazards

Time to the onset of DN was modelled by means of the Cox PH model, stratified according to gender, glucose control, hypertension, type of diabetes and smoking status. Failure in this study was defined as developing DN, which occurred when a patient’s ACR exceeded 2.5mg/mmol for men and 3.5 mg/mmol for women. However, it was possible for a patient to fall back below these cut-off values for DN on subsequent visits. Therefore patients were retained in the risk pool until they no longer attended the clinic, and not removed from the study once a “failure” had occurred. This ensured that data of patients fluctuating between normal and micro-albuminuria on subsequent visits was still included in the analysis. It was assumed that diabetic patients entered the risk pool for developing nephropathy at the time of diagnosis of Diabetes Mellitus. Although some patients may have had type 2 diabetes for several years before being diagnosed, it was the best available estimate of length of time at risk for these patients. Thus, this was a survival model with multiple records, delayed entries and multiple failures, all of which were taken into account during analysis.

Kaplan-Meier survival curves and Nelson-Aalen cumulative hazard curves were plotted. This provided an indication of the survival and hazard rates of all patients, as well as a comparison of these rates across gender, glucose control, hypertension, type of diabetes and smoking status. Log-rank tests determined whether a significant difference existed between the survival rates across the groups mentioned above. The PH assumption was tested by means of Schoenfeld residuals and log-log plots. A prognostic index obtained from the linear predictor of the Cox model was used to identify four risk categories based on the 25th, 50th and 75th percentiles[32]. Kaplan Meier curves of the different risk categories confirmed the model’s discrimination ability.

4.4.2 Threshold Regression

The same definition of failure, entry into and exit from the risk pool was used in the threshold regression model as in the Cox PH model. Univariate analysis of the data included likelihood ratio tests to determine whether the variable was included in the initial baseline component (ln (y_0)) or in the drift process as well (μ). Hazard ratios for gender and smoking status of a specific patient profile were obtained at 5, 10, 15 and 20 years. In addition, survival probabilities for specific patient profiles were determined and illustrated graphically.

4.4.3 Linear Mixed Models

The final model was selected on the basis of AIC and BIC scores and, where possible, likelihood ratio tests. A formula for the log (ACR) which included both fixed and random effects was determined. Diagnostics to test the assumptions of a LMM were carried out viz. the normality of the residuals. The data was also assessed for heteroscedasticity by means of plotting the standardised residuals against the predicted outcome and relevant covariates. The normality of the BLUPs was tested, and finally the fit of the model was evaluated by comparing the observed frequencies to the predicted frequencies.

4.4.4 Outliers

A number of outliers were observed during the analysis. These were investigated, and where possible original records were checked. Errors in data capturing were corrected, but many of the values were as recorded in clinic records and laboratory reports. Models were fitted with and without extreme outliers, but very little difference to the AIC, significant *P* values and coefficients were noted. The only consistent difference observed was smaller variances of the random effects in the LMM. Analyses were therefore carried out on the dataset with the outliers.

5. Results

All variables were assessed for normality, and ACR, baseline ACR, triglycerides and creatinine were found to be highly skewed to the right. A log-transformation normalised these four variables. Histograms and normal quantile graphs confirmed that all other variables were normally distributed.

Univariate analysis indicated significant associations between levels of albuminuria and levels of glucose control (*P*<0.001), gender (*P*<0.001), smoking status (*P*=0.01), race (*P*=0.003), hypertension (*P*<0.001) and type of diabetes (*P*<0.001). Significant differences were also observed in the ACR (*P*<0.001) and HbA1c levels (*P*<0.001) between hypertensive and non-hypertensive patients, type 1 and type 2 patients. No significant differences in ACR (*P*=0.01) and HbA1c (*P*=0.01) were noted across gender.

Of the 10 283 observations (visits) in the development dataset, only 9465 were used in survival analysis of the data as some patients exited the survival analysis at their first visit. The 10 283 visits were obtained from 775 patients. Similarly, the validation dataset consisted of 5136 observations, of which only 4412 from 287 patients were utilised in the survival analysis.

5.1 Demographic Results

The mean age of patients included in the study was 55.1 years (SD=15.5), with a range of 13 to 98 years. The sample was predominantly black African (90%) with most patients diagnosed as hypertensive (77%). Most patients had good glucose control (37.2%), 34.2% had moderate control and 28.6% had poor control. Baseline measurements of each patient were included in the analysis if the variable had been measured at their first visit to the clinic. The baseline and demographic characteristics of the whole dataset are summarised in Table 2 and Table 3.

Table 2 Baseline Characteristics

	n=No. Patients	$\overline{X}$ (SD)
Albumin/Creatinine Ratio[1] (median/IQR)	888	2.3 (0.90-10.34)
HbA1c	1064	9.72 (3.2)
Duration of Diabetes Mellitus (years)	779	11.73 (8.07)
LDL	940	2.83 (1.02)
HDL	976	1.24 (0.56)
Triglycerides[1](median/IQR)	972	1.5 (1-2.2)
Total Cholesterol	986	4.83 (1.22)
Creatinine[1](median/IQR)	1025	83 (68-105)
BMI	937	30.73 (6.87)
Glomerular Filtration Rate (GFR – MDRD)	970	0.51 (0.31)
Systolic Blood Pressure	1124	141.1 (27.1)
Diastolic Blood Pressure	1124	85.6 (15.6)

[1] Not log transformed

Table 3 Demographic Characteristics

	n=No. Patients		%
Gender	1142	Male	37.7
		Female	62.3
Race	1145	Black	90.7
		Coloured	0.4
		Indian	4.1
		White	4.8
Type of Diabetes	766	Type 1 Diabetes	32.4
		Type 2 Diabetes	67.6
Smoking Status	1040	Never smoked	70.3
		Stopped < 1 year ago	2.0
		Stopped > 1 year ago	18.2
		Currently smoking	9.5
Hypertension	1095	Yes	77.0
		No	23.0
Level of Albuminuria	889	Normal [1]	55.1
		Microalbuminuria [2]	30.4
		Macroalbuminuria *(ACR>30mmol/mg)*	14.5
Glucose Control	1063	Good (HbA1c<8)	37.2
		Moderate (8≤HbA1c≤11)	34.2
		Poor (HbA1c>11)	28.6

[1] *Males: ACR<2.5mg/mmol; Females: ACR<3.5mg/mmol*

[2] *Males: ACR 2.5mg -30mg/mmol; Females: ACR 3.5mg-30mg/mmol*

Longitudinal data, by its very nature, is useful in observing a change in mean values over time. ACR is the primary outcome variable, and HbA1c is an indication of glucose control in the patient. The change in the median ACR and mean HbA1c levels, over the years 2009 to 2013, is given below. Figure 1 illustrates the decrease in median ACR values across this time period. Mean HbA1c levels decreased between 2009 and 2010, implying better glucose control amongst these patients, but a sharp increase was noted in 2012. Upon further inspection, it appears that this increase was mostly amongst patients with poor glucose control. (See Figure 1 below)

Table 4 Change in median ACR and mean HbA1c (2009 – 2013)

Year	Median ACR	Mean HbA1c
2009	2.6	9.3
2010	2.1	8.3
2011	2.1	8.8
2012	1.7	9.9
2013	1.7	8.4

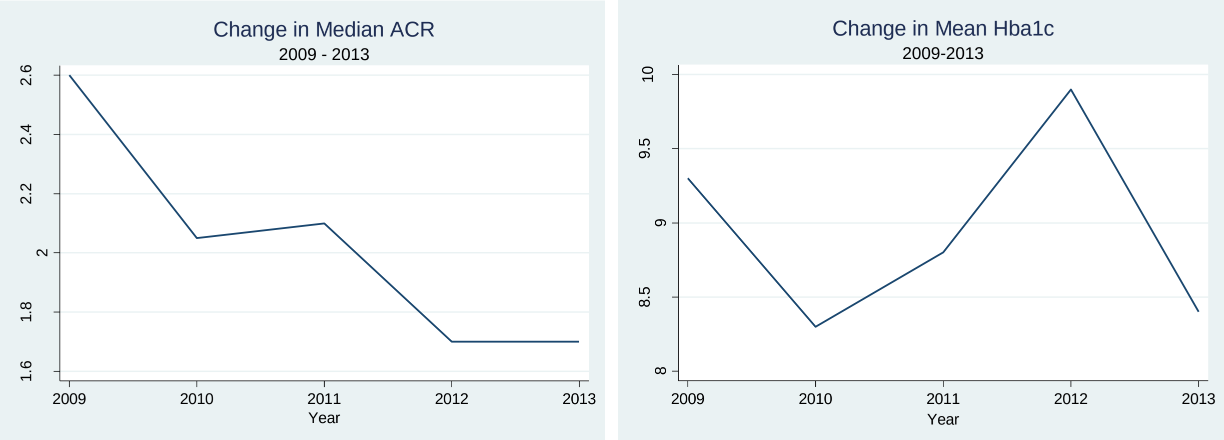


Figure 1 Changes in Median ACR and Mean HbA1c (2009-2013)

Table 5 Changes in Median ACR according to Levels of Glucose Control (2009-2013)

Year	Median ACR			Mean HbA1c		
	Good	Moderate	Poor	Good	Moderate	Poor
2009	1.7	2.8	3.8	6.9	9.4	13.7
2010	1.5	2.5	3.6	6.6	9.2	12.7
2011	1.2	2.9	2.4	6.7	9.3	12.5
2012	1.1	1.7	3.6	7.0	9.3	13.0
2013	1.2	2.6	3.7	6.9	9.2	12.4

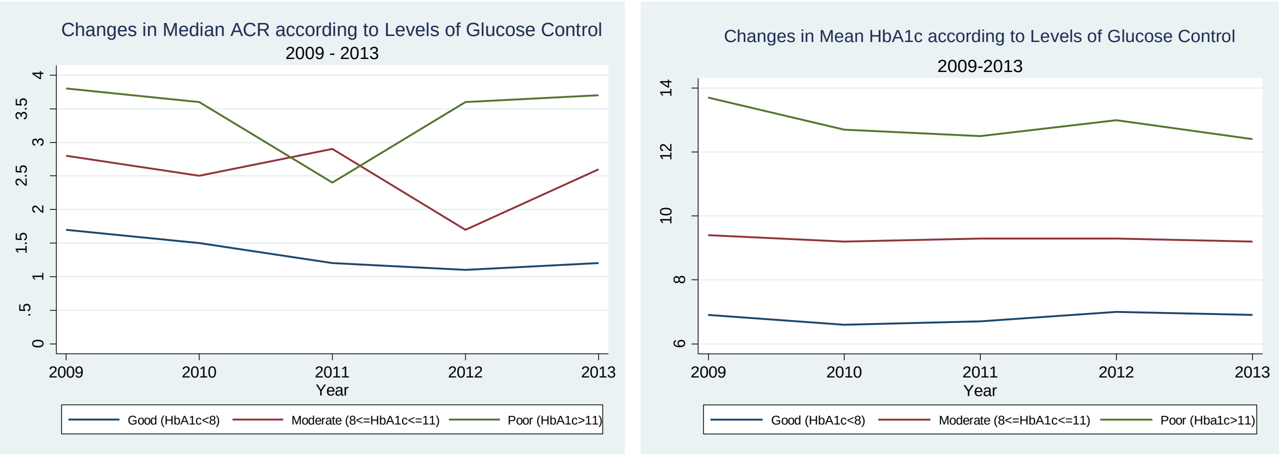


Figure 2 Changes in Median ACR and Mean HbA1c according to Levels of Glucose Control

Changes in median ACR according to levels of glucose confirm that patients with good glucose control maintained the lowest levels of ACR throughout the five years. Patients with poor glucose control improved ACR levels quite sharply in 2011, but then deteriorated again in 2012.

5.2 Missing Data

The important issue of missing data needed to be addressed before proceeding with the analysis. As mentioned previously, parameter estimates are biased if the reason subjects drop out is related to the data being collected, and cannot be ignored.[57] A large percentage of measurements were missing from this study, as much as 74% for some covariates. However, the blood tests for HbA1c, ACR, lipid profiles and creatinine levels were only requested twice a year by doctors as a standard procedure for all patients. Therefore, if one considers only the four basic annual visits, of which two should contain measurements for these variables, the percentage of missing values is reduced to approximately 40% (see Table 6). 14727 basic visits (i.e. visits 1 to 4) were recorded. Assuming that 50% of the visits should contain values for HbA1c, ACR, lipid profile and creatinine levels, the expected number of values that should have been recorded is 7363 for each of these variables. It is then possible to determine the proportion of actual missing values for these variables.

Table 6 Percentage of missing observations

	Number of Observed Visits with Measurements	% of Total Visits Missing	% of Basic Four Visits Missing
Urine ACR	6159	61.5	16.4
HbA1c	6859	57.2	6.8
Total Cholesterol	4170	73.7	43.4
LDL	4014	74.9	45.0
HDL	4147	74.1	44.0
Triglycerides	4141	74.1	43.8
Creatinine	4419	72.4	40.0
Potassium	4378	72.7	40.5

Furthermore, the missing values should be MCAR since the reason data is missing is not related to the outcome or any of the independent covariates in the study. One way of determining whether data is MCAR is to evaluate the proportion of missing values in each category of the data. If the data is MCAR these proportions should be equal across the categories.[62] Percentages of missing values across gender, type of diabetes, presence of hypertension, smoking status and race are presented in Table 7 below. The distribution of missing observations was found to be fairly equal across all groups. With the exception of missing ACR values of type 1 and 2 diabetic patients (*P*=0.013) and patients with and without hypertension (*P*=0.005), χ^2 tests indicated no significant associations at the 5% level of significance, between the proportion of missing values and the other categories. In spite of the significant association, the actual percentages of missing ACR data across type of diabetes and hypertension were still similar viz. 62.7% missing in Type 1 vs 60.3% missing in Type 2, and 63.6% missing in hypertensives vs 60.9% missing in non-hypertensives.

Table 7 Percentages of Missing Observations across Gender, Type of Diabetes, Hypertension, Smoking Status and Race

	Urine ACR	HbA1c	Total Cholesterol	LDL	HDL	Tri-glycerides	Creatinine	Potassium
Gender								
Male	61.4	57.1	73.6	75.5	73.9	73.9	72.5	72.7
Female	61.6	57.2	73.8	74.6	74.2	74.3	72.4	72.7
P	0.75	0.85	0.79	0.24	0.66	0.57	0.98	0.98
Type of Diabetes								
Type 1	62.7	58.2	74.1	75.3	74.2	74.2	72.7	72.9
Type 2	60.3	57.0	74.0	75.1	74.3	74.4	72.9	73.2
P	**0.013**	0.22	0.93	0.81	0.94	0.83	0.77	0.75
Hypertension								
Yes	63.6	58.2	73.7	75.1	74.3	74.3	72.2	72.3
No	60.9	56.9	73.8	75.0	74.1	74.2	72.6	72.9
P	**0.005**	0.19	0.98	0.91	0.81	0.86	0.66	0.53
Smoking Status								
Never smoked	61.5	74.0	74.0	75.0	74.5	74.5	72.8	73.1
Stopped >1 year ago	60.7	74.5	74.5	76.0	74.4	74.3	73.1	73.4
Stopped <1 year ago	61.3	72.8	72.8	75.4	73.8	73.8	71.5	72.1
Currently smoking	62.8	73.5	73.5	75.2	73.8	73.9	72.3	72.5
P	0.60	0.90	0.90	0.69	0.94	0.94	0.90	0.91
Race								
White	64.4	57.2	72.6	75.6	72.4	72.4	70.8	71.1
Black	61.5	57.2	73.8	74.8	74.2	74.3	72.5	72.7
Coloured	62.3	54.1	73.8	73.7	73.8	73.8	65.6	67.2
Indian	60.3	57.2	72.3	78.8	73.2	73.3	73.4	73.5
P	0.49	0.97	0.89	0.11	0.76	0.77	0.45	0.61

When one considers possible reasons for missing data in this study, one can conclude that it is because blood tests were not requested, or if ordered, the patient did not have the blood test, possibly due to costs, time constraints, or fear. Another reason for missing measurements occurs when the patient does not come to the clinic because of illness, transport constraints or relocation from the district. Of these reasons, only the possibility of a patient being too ill to attend the clinic is possibly related to variables in the study. Upon examining the data, only 593 visits of the total 16012 visits (3.7%) were recorded as missing i.e. the patient did not attend the clinic.

It was decided that MI would produce unreliable results because of the large percentage of values that would need to be imputed. Pattern theory cannot be applied to the data as it is not NMAR, and no pattern of missingness was identified within the missing observations. A further hindrance to estimating missing values in this particular dataset is that both the dependent and independent variables have missing data. Therefore, since the pattern of missing data was considered largely MCAR, the available data was analysed without further imputation of missing values.

5.3 Cox Proportional Hazards

The final Cox PH model fitted to the data, seen below in Table 8, had an AIC of 81 and a BIC of 90. The Cox model assumes equal baseline hazards unless otherwise specified.[24,33] Stratification accounts for the survival differences within a category, and the model was stratified according to gender, levels of glucose control, presence of hypertension, type of diabetes and smoking status.

Table 8 Cox Stratified Regression Model

	Coefficient (95% CI)	Hazard Ratio (95% CI)	p>|z|
Log (Baseline ACR)	0.37 (0.15 – 0.60)	1.45 (1.16 – 1.82)	0.003
Total Cholesterol	0.69 (0.24 – 1.14)	1.99 (1.27 – 3.12)	0.001

Stratified by gender, glucose control, hypertension, type of diabetes and smoking status

The Cox model fitted to the data indicates that patients with high total cholesterol are twice as likely to develop DN, than patients who have lower levels. High log (baseline ACR) also increases the risk of DN by almost 50%.

The probability of the diabetic patients in this study not developing DN by analysis time *t,* is illustrated in the Kaplan-Meier survival curve in Figure 3. The estimated probability of survival at 5 and 10 years is 0.09 and 0.0044 respectively (See Appendix 11.2). It is clear that 10 years after diagnosis, the risk of developing DN is almost certain for these patients. As all patients referred to the diabetes clinic have DM, and many present with complications, it is very probable that almost the whole study population will at some point develop DN.

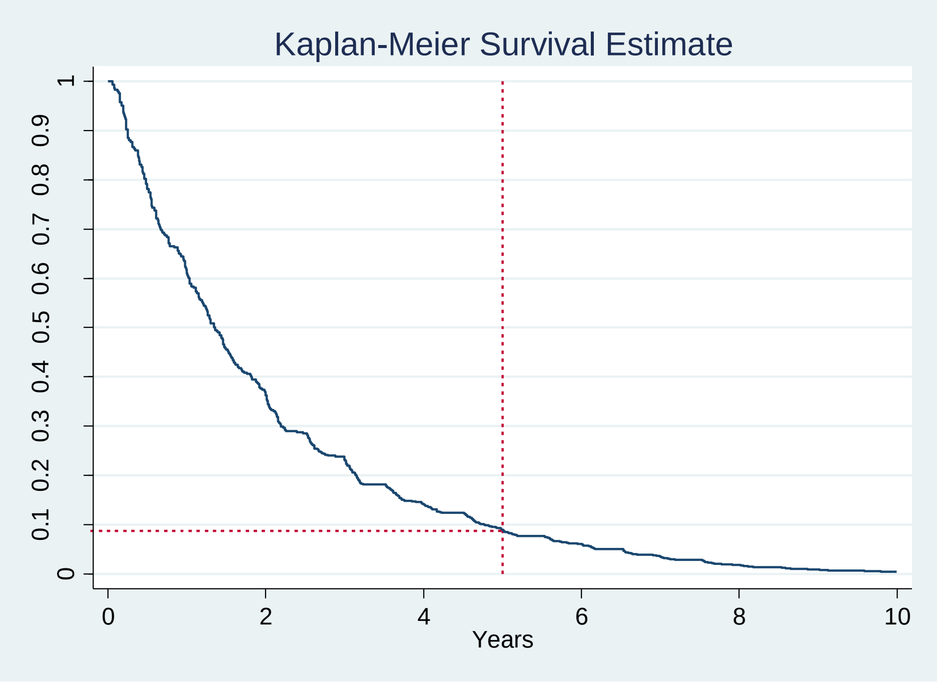
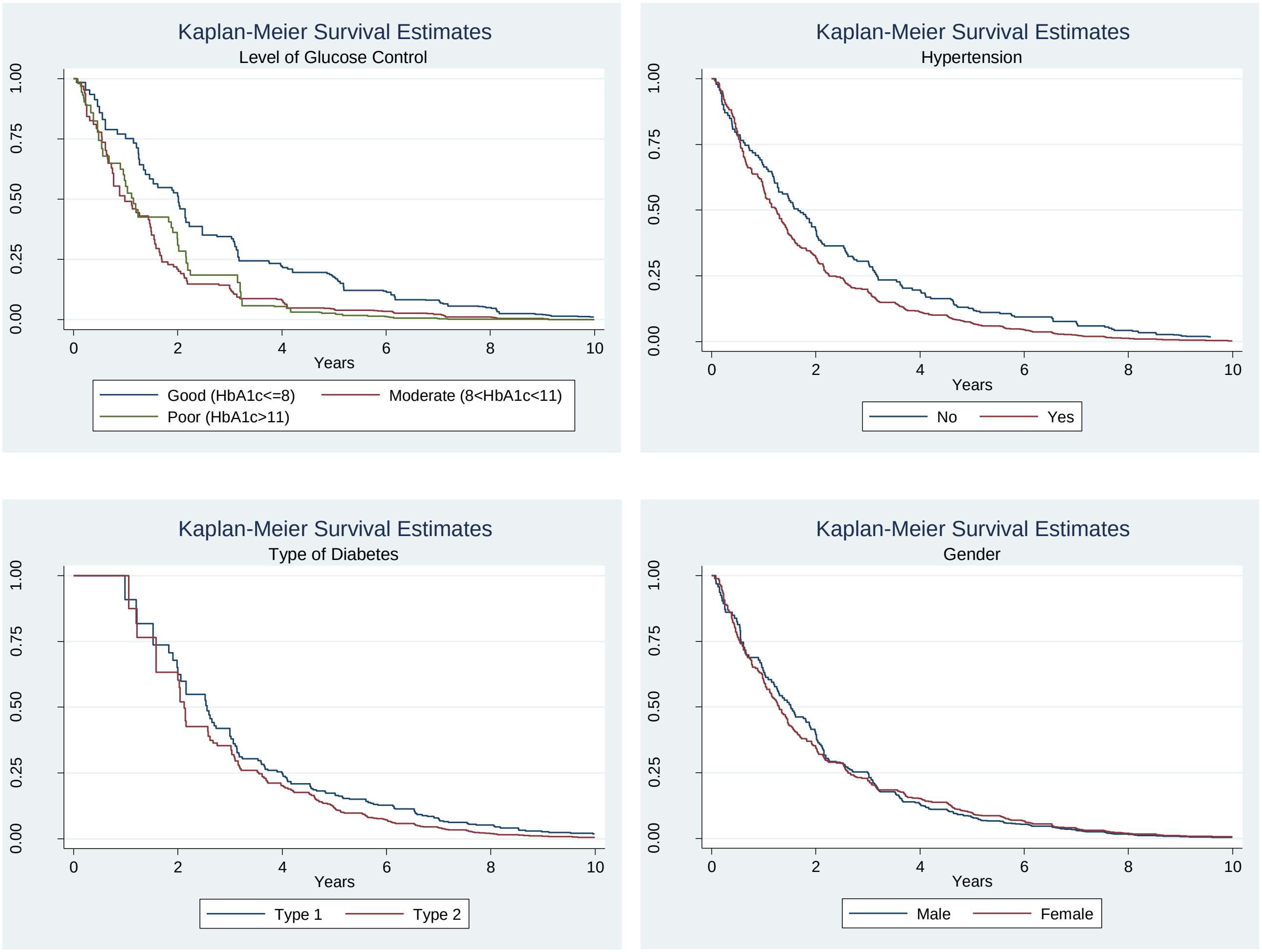


Figure 3 Overall Kaplan-Meier survival curve

Survival probabilities across categories can also be estimated. The Kaplan-Meier curves in Figure 4 below illustrate survival functions according to levels of glucose control, hypertension, type of diabetes and gender. The closer the survival curve is to the y-axis, the lower the survival rates of that category. Intersecting survival curves are an indication that the PH assumption may not hold for that variable. From the graphs below one can see that good glycaemic control and the absence of hypertension play a role in delaying the development of DN. Also, the probability of survival is marginally better for females and type 1 DM patients.


Kaplan-Meier Survival Estimates
Level of Glucose Control

Kaplan-Meier Survival Estimates
Hypertension

Kaplan-Meier Survival Estimates
Type of Diabetes

Kaplan-Meier Survival Estimates
Gender

Figure 4 Kaplan-Meier Survival Curves according to Level of Glucose Control, Hypertension, Type of Diabetes and Gender.

The risk of failure, or hazard rate, is closely related to the survival function, and is illustrated by the Nelson-Aalen cumulative hazard graph. The cumulative hazard rate, or total amount of risk accumulated to time *t*, can be interpreted as the number of failures or hazards (raised ACR levels indicating nephropathy) one would expect in a given time period, rather than a probability of a hazard occuring.[65] The hazard functions in Figure 5 provide a clearer picture of the risks faced across the four categories. A higher cumulative hazard curve indicates a higher number of expected failures. Clearly, female patients with good glucose control and no hypertension should experience fewer hazards than male patients with poor glucose control and hypertension. However, the differences in gender only seem to become noticeable approximately 8 years after diagnosis of diabetes. Patients able to control their glucose levels have a distinctly lower hazard rate, as do non-hypertensive patients. Finally, patients with type 2 DM have slightly higher cumulative hazards than type 1 patients. (See Appendix 11.1 Figure 18 for Kaplan-Meier and Nelson-Aalen curves according to smoking status).

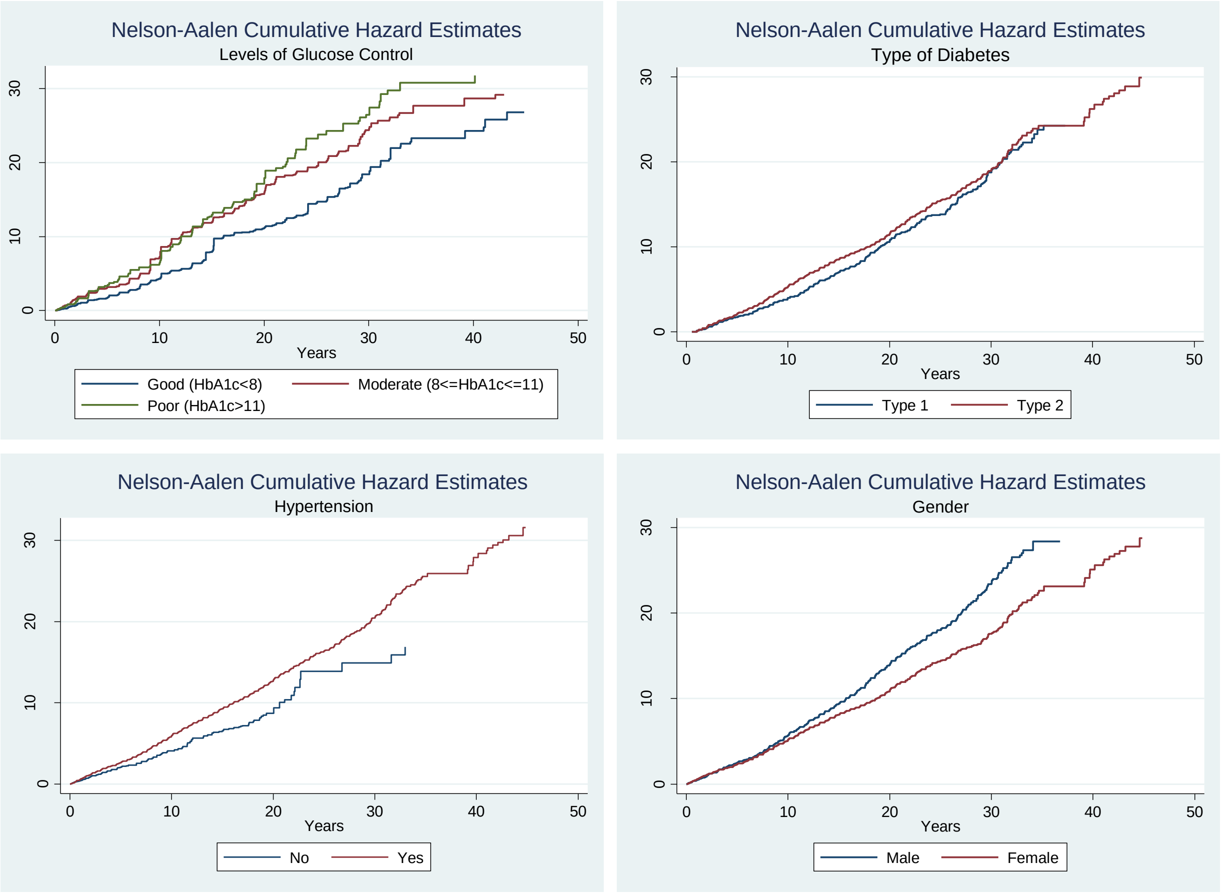


Figure 5 Cumulative Hazard Estimates by Level of Glucose Control, Type of Diabetes, Hypertension and Gender

Log-rank tests carried out on gender (*P*<0.001) and glucose control (*P*<0.001) indicated a significant difference between the survival rates of male and female patients and the different levels of glucose control. There was also a significant difference in survival rates of those with hypertension and those without (P<0.001). However, no significant difference in the survival rates according to type of diabetes (*P*=0.5) and smoking status (*P*=0.06) was noted.

A visual assessment of the proportional hazards assumption of the Cox survival model can be observed in the log-log plots of Figure 6. Parallel lines for the categories confirm the assumption of proportional hazards.

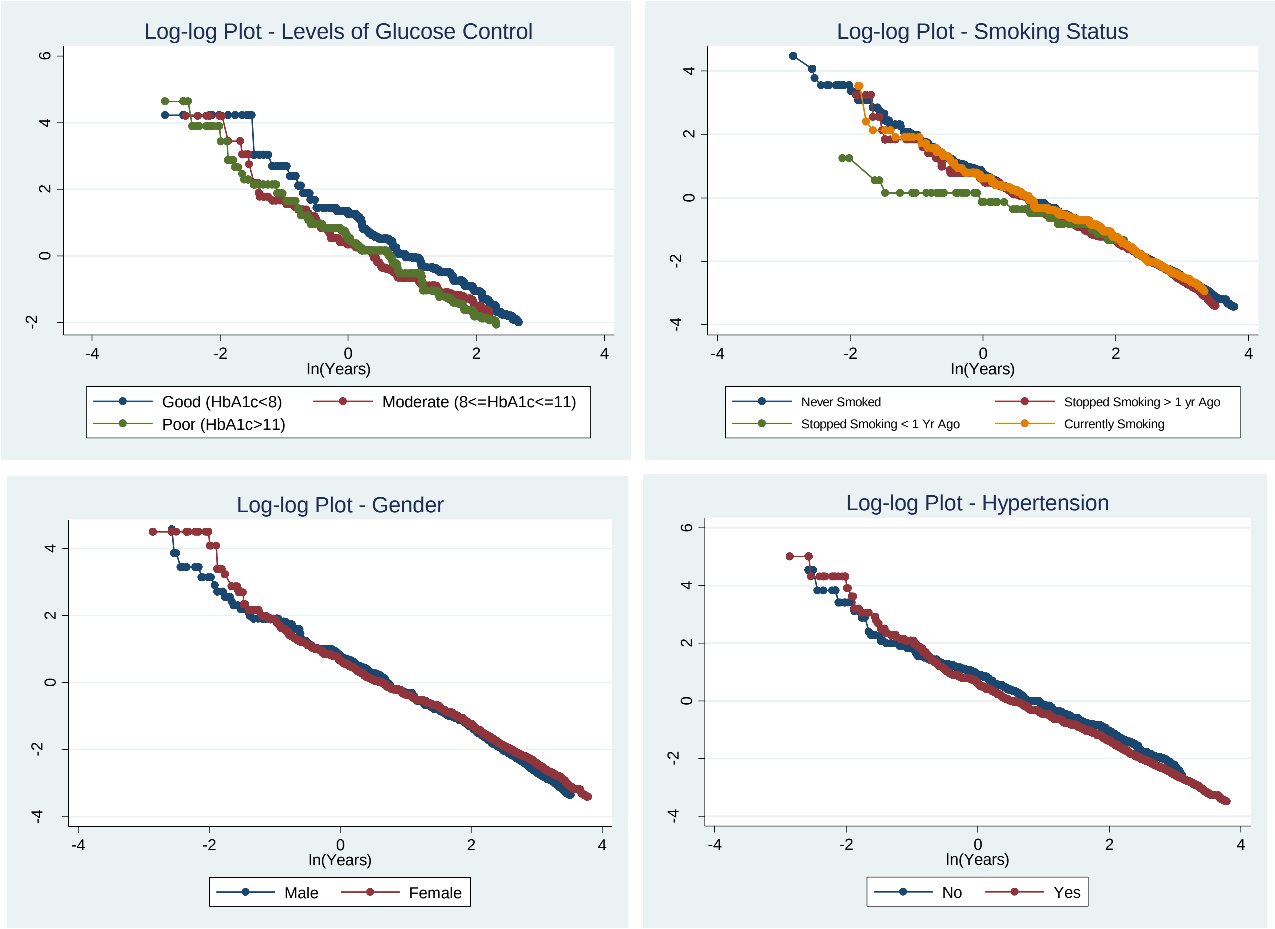


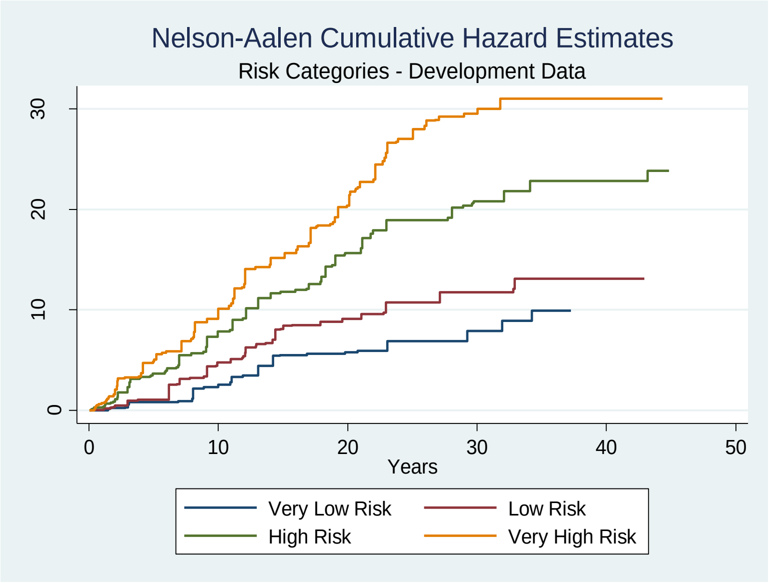
Figure 6 Testing the proportional hazards assumption by Levels of Glucose Control, Smoking Status, Gender and Hypertension

A further test of the proportional hazards assumption was based on the Schoenfeld residuals. The null hypothesis of proportional hazards cannot be rejected with the significance values seen in Table 9 below.

Table 9 Critical Values for Test of Proportional Hazards Assumption

	P
Log(Baseline ACR)	0.85
Total Cholesterol	0.15
Global Test	**0.35**

A prognostic index was obtained from the linear predictions of the model. Four risk groups, namely Very Low Risk, Low Risk, High Risk and Very High Risk, were created from the 25th, 50th and 75th percentiles of the prognostic index. There was very good discrimination between the four risk groups, evidenced by the clearly separated hazard curves in Figure 7 below. Fitting the same model to the validation data also showed fairly good discrimination, but was less likely to discriminate between the four risk categories in the first 8 years.



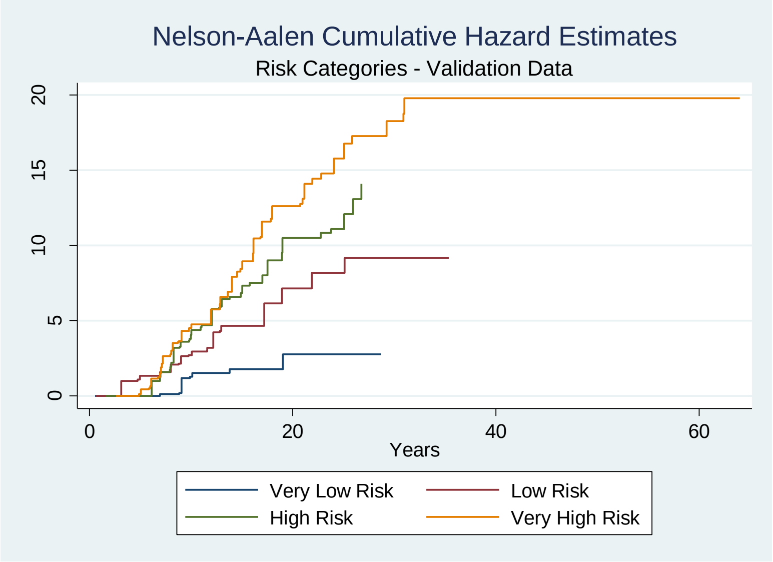


Figure 7 Discrimination between the four risk categories illustrated by Nelson-Aalen cumulative hazard curves for the development and validation datasets

Finally, a comparison of the observed survival probabilities of the Kaplan-Meier curve with the predicted survival curves of the Cox models in Figure 8 below, again confirmed the PH assumption for gender, hypertension and glucose control. The close fit of the Cox PH model to the observed survival probabilities is also very clear.

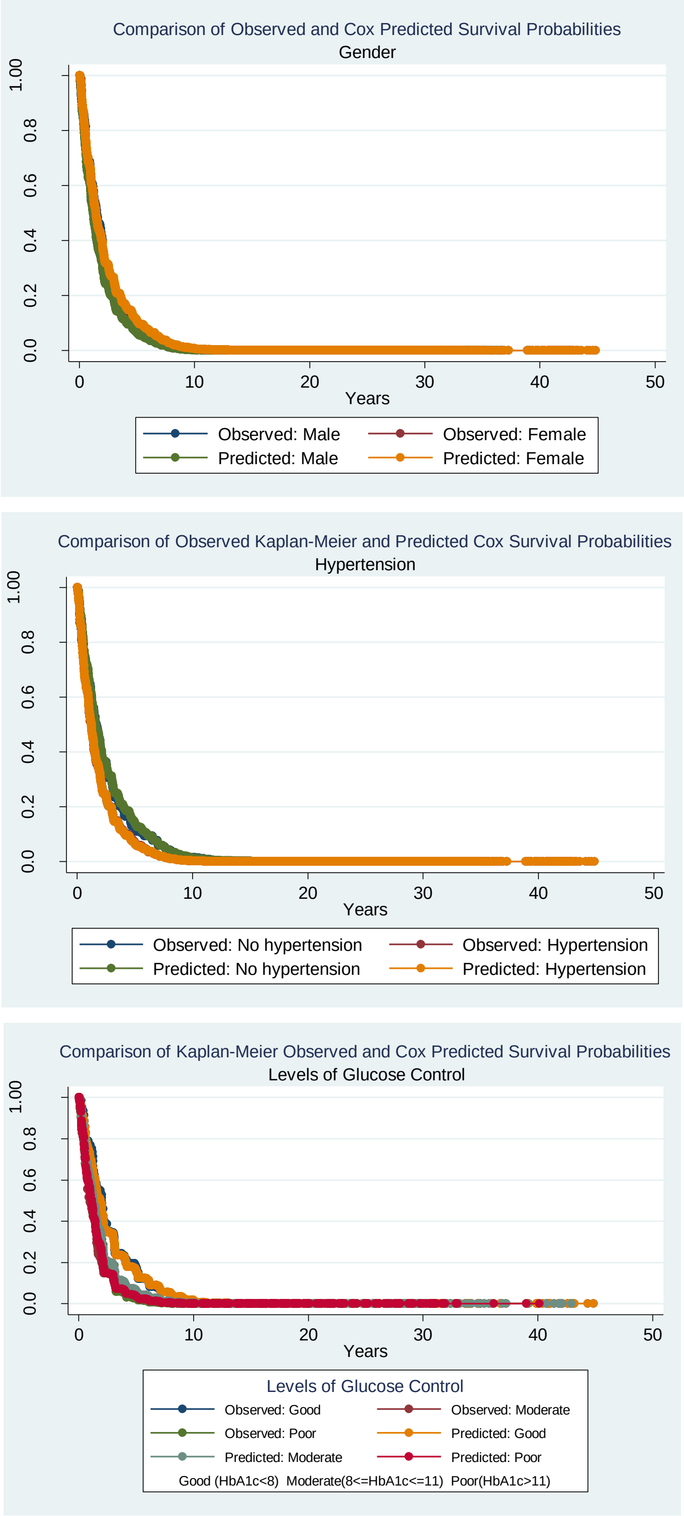


Figure 8 Comparison of Kaplan-Meier Observed and Cox Predicted Survival Probabilities by Gender, Hypertension and Level of Glucose Control

5.4 Threshold Regression

The threshold regression model, described below in Table 10, included several predictor variables. Initial health state was influenced primarily by smoking status, baseline log(ACR) and baseline HbA1c levels, all risk factors known to be associated with DN. These covariates contribute negatively to the initial health status, as indicated by the negative coefficients of their parameters. From this, one can deduce that a smoker with higher baseline log(ACR) and HbA1c levels, is closer to the threshold of developing DN. Gender and type of diabetes were also included in the model as covariates. A positive coefficient for gender implies that females tend to have a healthier initial state than males (Gender = 0 males, 1 females). Similarly, the positive coefficient of type of diabetes indicates that DM Type 2 patients begin the process of health deterioration from a healthier point than that of Type 1 patients.

In this model, the process of health deterioration (**μ**) was influenced mostly by the duration of DM, which is measured from the date DM was diagnosed in the patient to the date of clinic visits. If this information was not available, it was assumed DM was diagnosed at the first visit to the diabetes clinic. An interesting effect is illustrated by the inclusion of DM duration in the model. A positive coefficient for **μ** covariates indicates that the rate of health decline is slower, whereas a negative coefficient, which one would expect to see more frequently, is evidence that the covariate contributes to a steeper decline in health. Thus, the longer the duration of DM, the slower the deterioration towards DN is in the patient. This makes sense, as the earlier diabetes is detected in a patient the better a patient can manage their glucose, lipids and blood pressure, thus slowing down the progression of DN.

The remaining covariates affecting the process of health decline, viz. weight, potassium and triglyceride levels, contribute to a more rapid decline towards DN. An overweight patient with high levels of potassium and triglycerides will cross the threshold for DN quicker than a patient with normal values of these covariates. The AIC of the threshold model was 1428 and BIC 1497. Applying the same model to the validation data yielded an AIC of 835 and a BIC of 897.

Table 10 Threshold Regression Model

	Coefficient	95% CI	p>|z|
Ln (y_0)			
Gender	0.09	-0.016 – 0.197	0.095
Smoking Status ***(Reference: Never Smoked)***			
Stopped >1 year ago	-0.01	-0.13 – 0.10	0.85
Stopped <1 year ago	-0.35	-0.61 – -0.09	0.01
Currently Smoking	-0.16	-0.30 – -0.01	0.04
Type of Diabetes	0.09	0.00 – 0.18	0.06
Log(Baseline Albumin-Creatinine Ratio)	-0.06	-0.08 – -0.04	<0.001
Baseline HbA1c	-0.01	-0.03 – 0.00	0.06
Constant	2.12	1.89 – 2.34	<0.001
µ			
Duration of Diabetes	0.02	0.015 – 0.024	<0.001
Weight	-0.001	-0.002 – 0.0002	0.11
Log(Potassium)	-0.08	-0.220 – 0.067	0.30
Log(Triglycerides)	-0.02	-0.059 – 0.024	0.40
Constant	-0.30	-0.556 – -0.039	0.02

Threshold regression calculates hazard ratios for very specific patient profiles.[51] Table 11 describes hazard ratios calculated for gender and smoking status at 5, 10, 15 and 20 years for the following profile:

- Female patient
- Currently smoking
- Type 2 diabetes
- Weighing 80kg

- Baseline HbA1c=9
- Duration of DM=10 years
- Log(triglycerides)=2.65 *(Triglycerides=14.15)*
- Log(potassium)=1.5 *(Potassium=4.48)*
- Log(baseline ACR)=1 *(ACR=2.72)*

Hazard ratios for gender confirm that diabetic females have a lower risk of developing nephropathy than diabetic males throughout the 20 years. Non-smoking patients also have a distinctly lower risk compared to smoking patients. For instance, 5 years after diagnosis, female smokers are 50% less likely to develop DN than their male counterparts, whereas non-smoking females are 64% less likely to develop DN than non-smoking males. These percentages decrease with time however, and at 20 years the female smoking / non-smoking patients are only 8% and 11%, respectively, less likely than male smoking/non-smoking patients to develop DN. At 5 years after diagnosis smokers with the above profile are 5.13 times more likely to develop nephropathy than patients who have never smoked, diminishing to 1.21 times more likely after 20 years.

Table 11 Threshold Regression: Hazard ratios according to gender and smoking status

Time	Hazard Ratio – Females (Non-Smoker) (95% CI)	Hazard Ratio – Females (Smoker) (95% CI)	Hazard Ratio – Currently Smoking (95% CI)
	Reference: Males – Non-Smoker	*Reference: Males – Smoker*	*Reference: Never Smoked*
5 years	0.36 (0.08 – 1.22)	0.50 (0.13 – 1.23)	5.13 (0.57– 49.46)
10 years	0.69 (0.41 – 1.12)	0.78 (0.49 – 1.05)	1.79 (0.80 – 3.60)
15 years	0.83 (0.65 – 1.03)	0.88 (0.71 – 1.02)	1.35 (0.95 – 1.92)
20 years	0.89 (0.77 – 1.04)	0.92 (0.81 – 1.02)	1.21 (0.96 – 1.51)

The hazard ratio curves accompanying Table 11 are presented below.

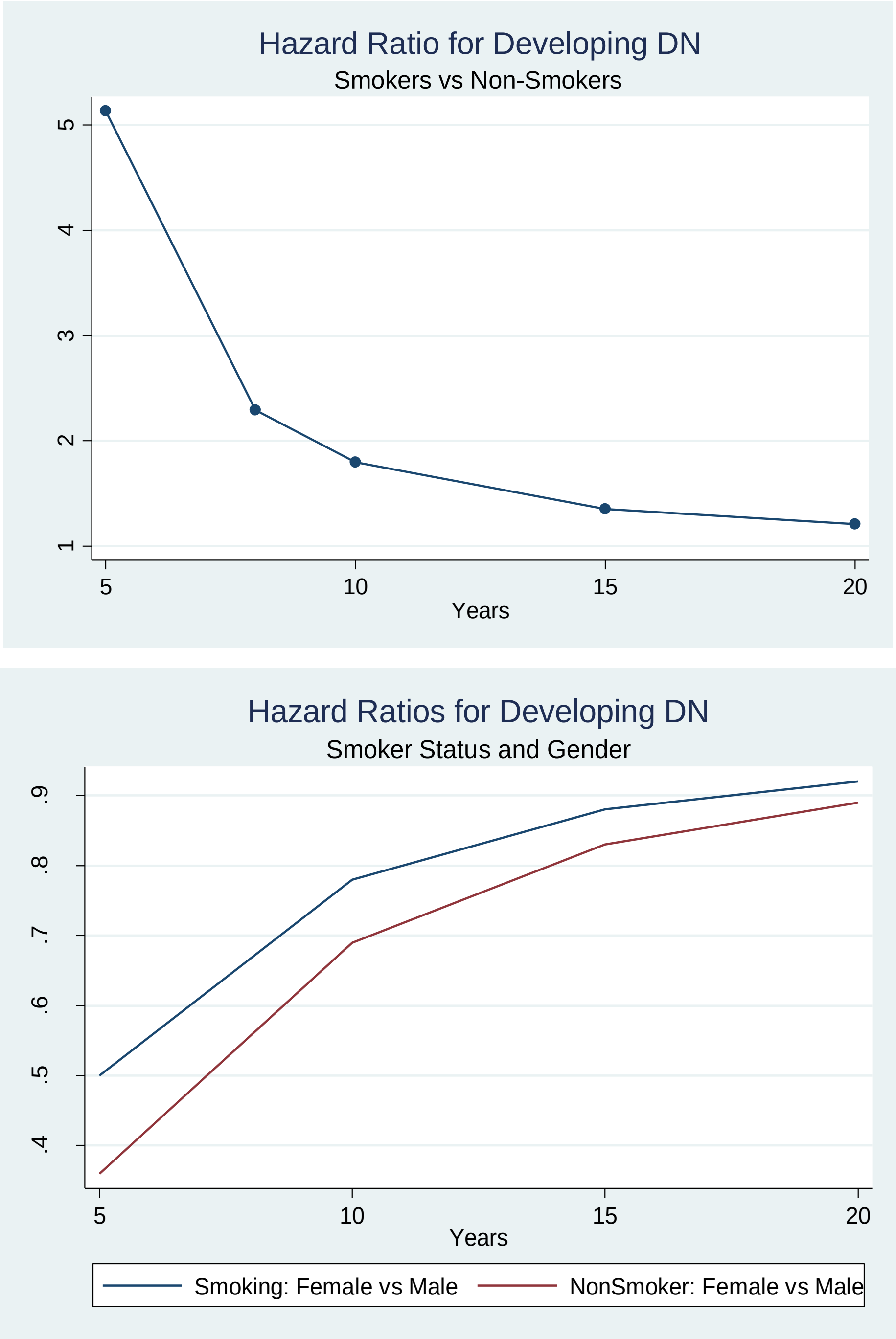


Figure 9 Hazard Ratios according to Smoking Status and Gender for Patient Profile: Weighs 80kg, Type 2 DM, Baseline HbA1c=9, Duration of DM=10 years, Log(baseline ACR)=1, Log(triglycerides)=2.65, Log(potassium)=1.5

Survival predictions for specific patient profiles are also possible.[51] For the same profile used to determine hazard ratios above, the predicted survival functions are contrasted below according to weight, baseline HbA1c and log (ACR) levels and duration of DM. Overweight women had lower survival rates, illustrating the importance of maintaining a healthy weight in lowering the risk of diabetic nephropathy. Higher baseline HbA1c and log (ACR) levels also indicated lower survival probabilities, as did patients with a shorter duration of DM.

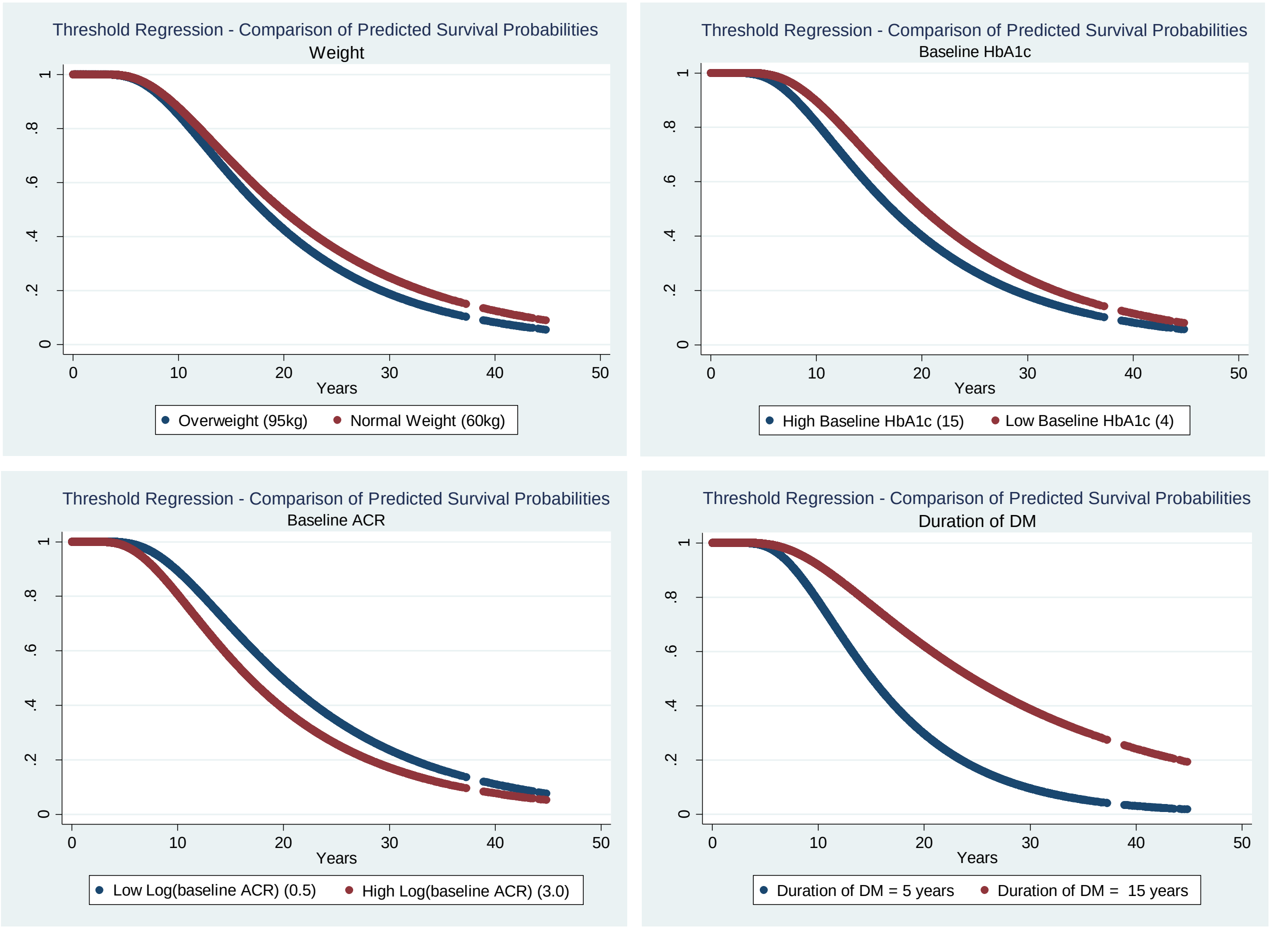


Figure 10 Predicted Threshold Regression Survival Functions according to Weight, Baseline HbA1c and Log(ACR) and Duration of DM for Patient Profile: Female, 80kg, Smoker, Type 2 DM, Moderate glucose control (HbA1c=9), Duration=10 years, Log(baseline ACR)=1, Log(triglycerides)=2.65, Log(potassium)=1.5

The requirement of such specific patient profiles for predictions in threshold regression hindered the creation of risk categories comparable to those created with the Cox model. It was therefore difficult to compare levels of discrimination between the two models.

5.5 Linear Mixed Model

A LMM with random coefficients was fitted to the data. Log (baseline ACR), log (creatinine), hypertension, poor glucose control and systolic blood pressure were significant fixed effects in the model. Random effects measuring the individual deviation from the population average included smoking status, total cholesterol and weight. The variances and confidence interval components of total cholesterol and weight are very small, and therefore the data was also modelled without these two random effects in order to see the effect of their omission on the model. Although minimal change to the coefficients was noted, the AIC worsened from 3755 to 3864 and BIC from 3818 to 3914, and it was decided to retain the random effects of total cholesterol and weight. AIC and BIC for the model fitted to the validation data was 2430 and 2475 respectively.

Table 12 Linear Mixed Model

Log (ACR)	Coefficient	95% CI	*P*>|z|
Fixed Effects			
HDL	0.086	(-0.108 – 0.279)	0.385
Baseline HbA1c	0.029	(-0.004 – 0.062)	0.083
Baseline log(ACR)	0.623	(0.576 – 0.669)	<0.001
Log(Creatinine)	0.273	(0.093 – 0.453)	0.003
Systolic Blood Pressure	0.009	(0.006 – 0.012)	<0.001
Type of Diabetes	0.047	(-0.149 – 0.242)	0.640
Glucose Control *(Ref Good)*			
Moderate	0.026	(-0.143 – 0.196)	0.760
Poor	0.371	(0.134 – 0.609)	0.002
Hypertension	0.246	(-0.032 – 0.524)	0.083
Number of Hypoglycaemic Episodes	0.015	(-0.011 – 0.041)	0.271
Constant	-2.739	(-3.757 - -1.721)	<0.001
Random Effects – Variance			
Identity: Smoking Status	0.309	(0.111 – 0.864).	.
Total Cholesterol	7.24e -17	.(4.06e-19 – 1.29e-14)	.
Weight	6.72e-13	.	.
Constant`	0.010	(3.08e-15 – 3.45e+10)	.
Variance: Residual	1.159	(1.052 – 1.278).	.

Using the above estimates of the fixed-effect parameters and random effects BLUPs it is possible to write a formula for the predicted log (ACR) at visit *t* for patient *i* as follows:

$$\begin{aligned}\log\widehat{(ACR)} = {} & -2.739 + 0.086\, HDL + 0.029\, Baseline\ HbA1c + 0.623 \log(Baseline\ ACR) \\ & + 0.273 \log(Creatinine) + 0.009\, Systolic\ Blood\ Pressure + 0.054\, Type \\ & + 0.026\, Glucose\ Control(Mod) + 0.371\, Glucose\ Control(Poor) \\ & + 0.246\, Hypertension + \ 0.015\, No.Hypoglycaemic\ Episodes \\ & + \ u_{1i}\ \ Smoking\ Status\ + u_{2i}\, Total\ Cholesterol + \ u_{3i}\, Weight\ + \ 0.01\end{aligned}$$

u_{1i} , u_{2i} and u_{3i} represent the predicted individual BLUPs of the random effects of smoking status, total cholesterol and weight for patient *i*. [2]

Predictive margins of log (ACR), according to levels of glucose control, yield the following table:

Table 13 Predictive Margins According to Levels of Gender and Glucose Control

Levels of Glucose Control	Predicted Log(ACR)	ACR	*P*	95% CI
Male				
Good (HbA1c<8)	1.09	2.98	<0.001	(0.96 – 1.22)
Moderate (8<=HbA1c<=11)	1.12	3.06	<0.001	(0.98 – 1.25)
Poor (HbA1c>11)	1.46	4.32	<0.001	(1.27 – 1.66)
Female				
Good (HbA1c<8)	0.89	2.43	<0.001	(0.76 – 1.01)
Moderate (8<=HbA1c<=11)	0.91	2.49	<0.001	(0.78 – 1.05)
Poor (HbA1c>11)	1.26	3.52	<0.001	(1.06 – 1.45)

The cut-off ACR values for micro-albuminuria, and thus DN, are an ACR greater than 2.5 for men, and 3.5 for women. The ACR for men is above this cut-off level in all three categories of glucose control. The risk of DN is lower for women, with only female patients with poor glucose control falling above the cut-off level for micro-albuminuria. Figure 11 below clearly indicates the higher risk of DN for males and patients with poor glucose control.

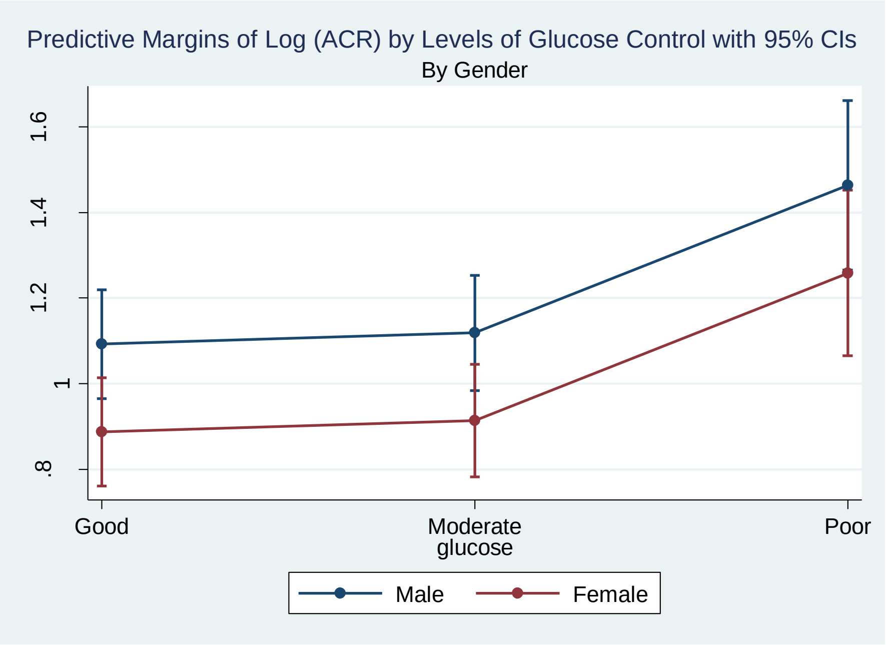


Figure 11 Margins Plot of Predicted Log(ACR) at Different Levels of Glucose Control and Gender

Diagnostic tests on the fixed effects confirmed that residuals are normally distributed with the exception of some outliers that skew the QQ plot slightly at the lower end (See Figure 12). These outliers were investigated, but values were correct according to clinic records and laboratory reports. Omitting them had little effect upon the value of coefficients or significance of variables included in the model, and they were therefore kept in the dataset.

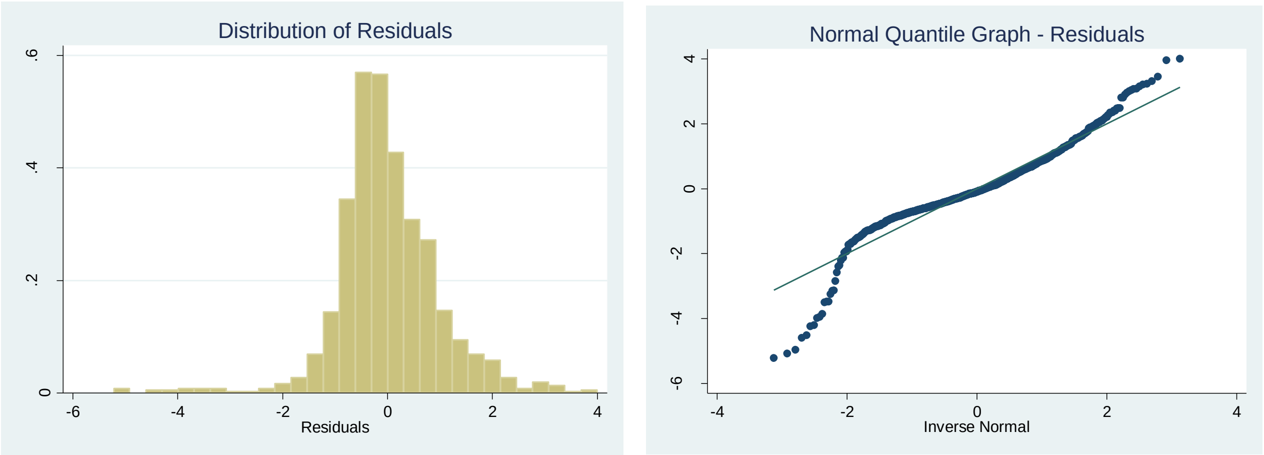


Figure 12 Histogram and Normal Quantile Graphs of LMM Residuals

By plotting standardised residuals against the predicted outcome and fixed effects, a fairly constant variance was observed in all variables. Fitted residual plots for covariates log (baseline ACR), baseline HbA1c, HDL and systolic blood pressure are illustrated below. (See Appendix 11.1 Figure 19 for graphs of remaining covariates.)

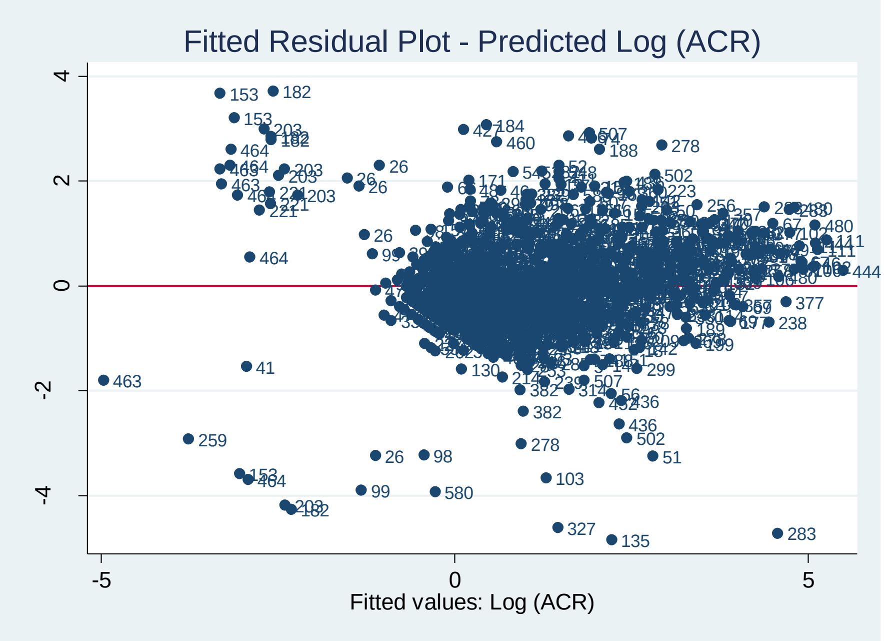


Figure 13 Fitted Residual Plot - Standardised Residuals against Predicted Log(ACR)

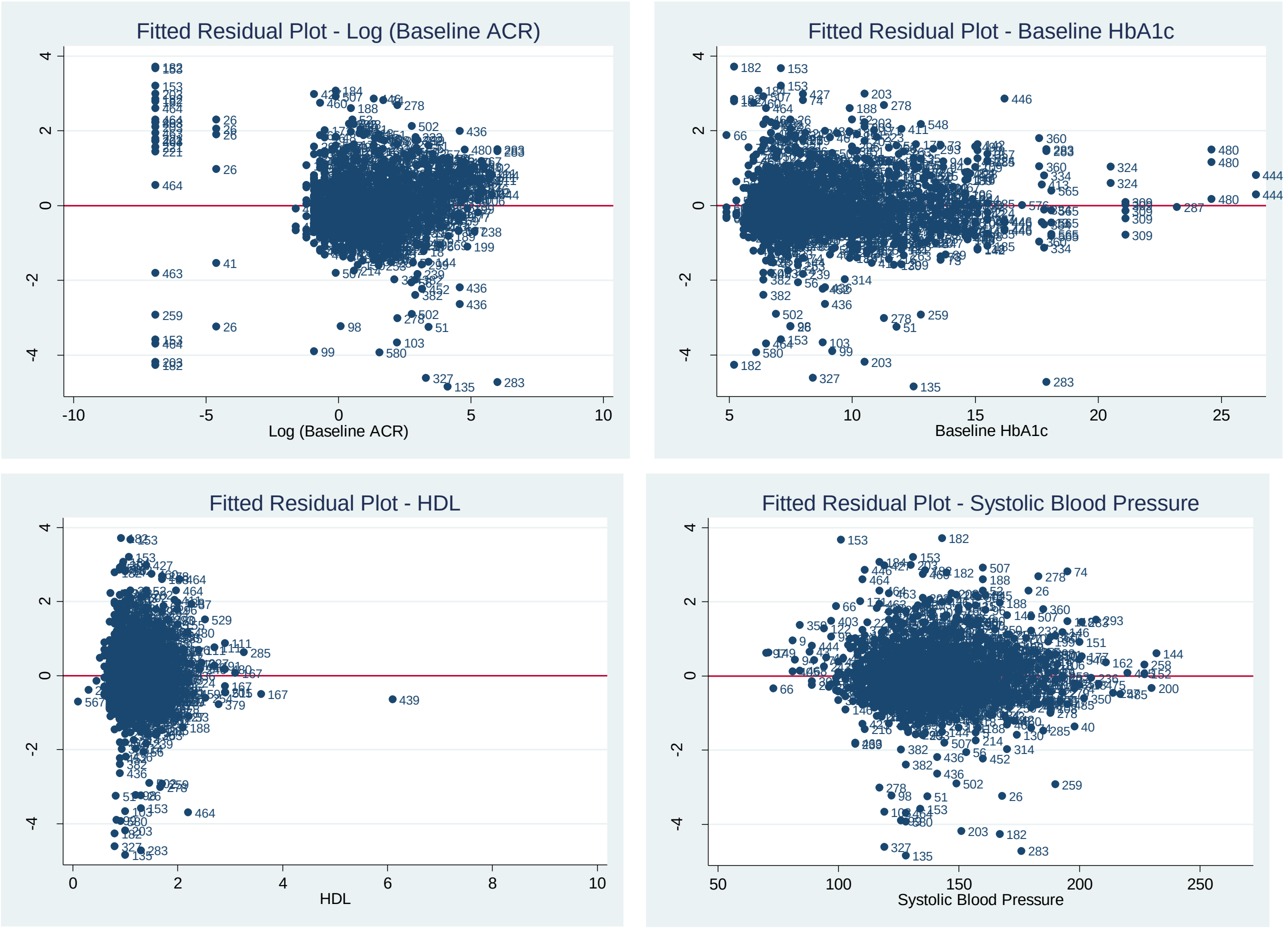


Figure 14 Fitted Residual Plots of Standardised Residuals against Log (Baseline ACR), Baseline HbA1c, HDL and Systolic Blood Pressure

The distribution of the random effects (BLUPs) was also normal. (See Appendix 11.1 Figure 20 for histograms illustrating this.)

Finally, the agreement between the observed and predicted ACR values was checked. If the LMM model has good predictive abilities there should be a correlation between these two sets of values along the 45° line in Figure 15 below. There appears to be a good measure of agreement between the observed and predicted values. A similar pattern was observed when the LMM model was applied to the validation dataset.

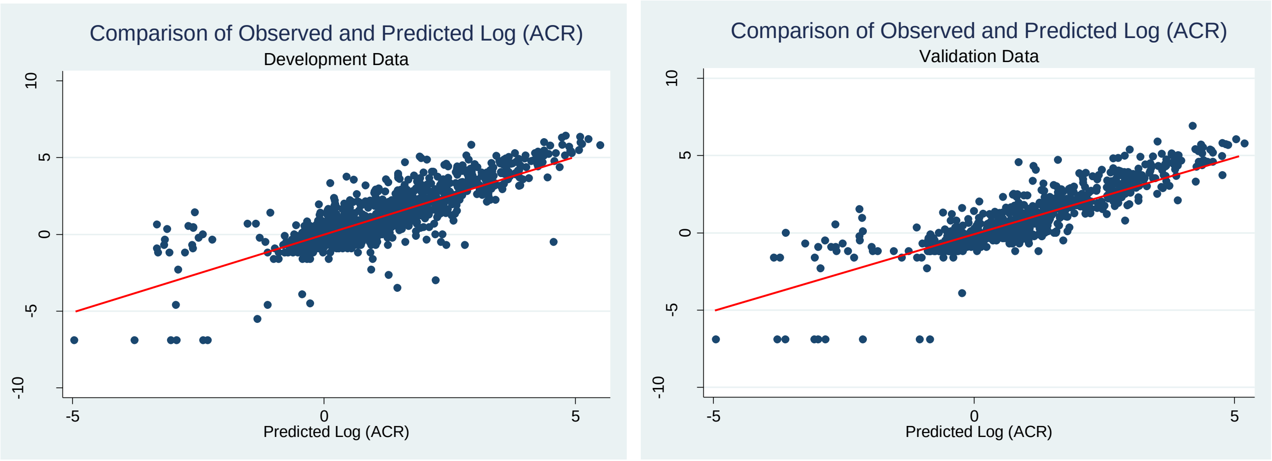


Figure 15 Comparison of Observed and Predicted Log (ACR) in Development and Validation Data Reference line is observed=predicted.

There is a tendency for the model to under-estimate the higher values of Log (ACR) as depicted in Figure 16 below, where the observed and predicted ACR values are compared across the three levels of glucose control. The LMM model is better at predicting values in the good glucose control category than the remaining two categories. There is a tendency for the model to underestimate ACR values, especially in the moderate level of glucose control.

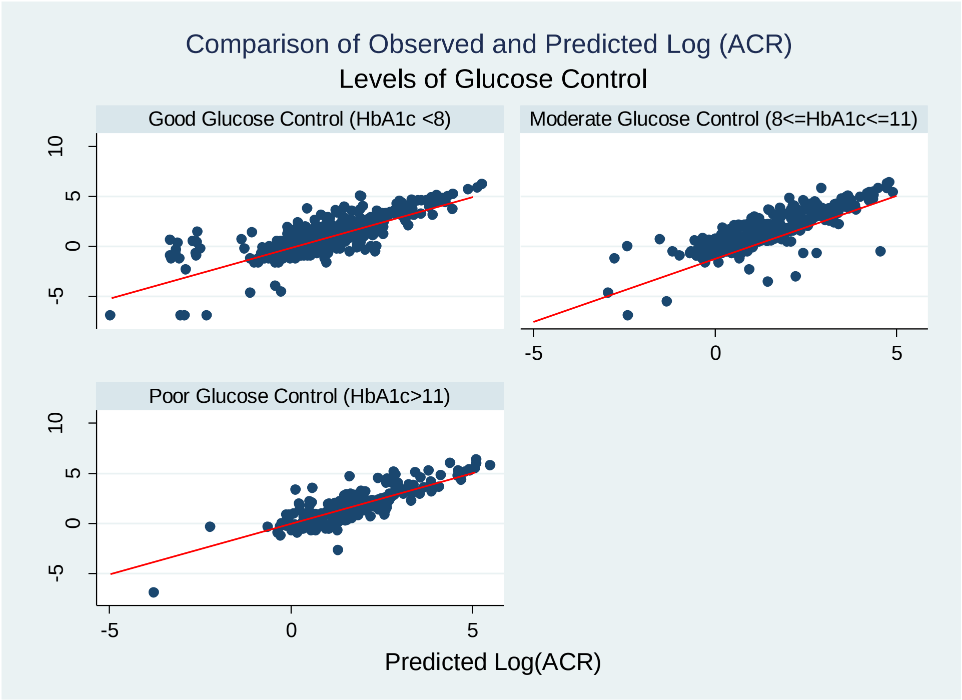


Figure 16 Comparison of Observed and Predicted Log (ACR) according to Levels of Glucose Control. Reference line is observed = predicted.

Examining the number of observations allocated to the three different DN categories by the LMM model in Table 14 below, 143 (73 + 68 + 2) observations were allocated to lower DN categories, and 88 (81 + 6+1) observations to higher DN categories. Again, the LMM tends to underestimate the degree of DN in a patient. A significant difference was observed between the observed and predicted risk categories (*P<0.001)*.

Table 14 Reclassification of DN Categories

Observed DN Categories	**Predicted DN Categories**		
	0	**1**	**2**
0	621	81	1
1	73	249	6
2	2	68	70

Once again, a similar pattern was observed between the observed and predicted DN categories of the validation dataset, with 87 observations reclassified into a lower risk category, and 52 observations reclassified into a higher risk category. There was also a significant difference between the observed and predicted frequencies of microalbuminuria categories (*P*<0.001). (See Appendix 11.1 Table 16 for validation data results).

Figure 17 below also clearly illustrates that the model is less able to predict the higher extremes of ACR values.

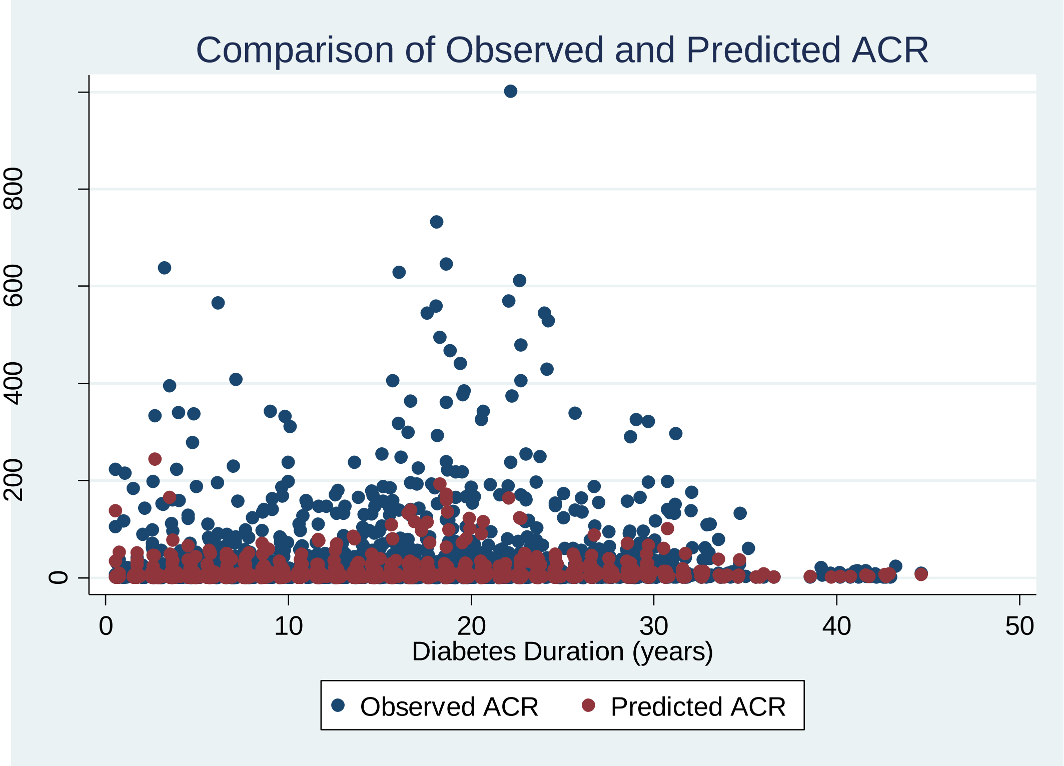


Figure 17 Comparison of Observed and LMM Predicted ACR

6 Discussion

In comparing the results of the stratified Cox regression, threshold regression and LMM it is immediately obvious that the Cox PH model is the most parsimonious model. However, this is not always the case, as other studies of threshold regression models have shown the opposite.[50] In the analyses for this dissertation, both the threshold regression and LMM included more explanatory variables, with threshold regression providing a deeper insight into factors affecting the deterioration of a diabetic patient's health and the LMM including individual variation in patients' smoking status, total cholesterol and weight. This agrees with the findings of other studies on the value added to the interpretation of the data by threshold regression modelling.[52,53] All the models indicated a significant influence of high baseline ACR on the risk of developing DN. Lipids played a role as well, with the Cox model and LMM including total cholesterol as a covariate.

A very close fit was observed when comparing the predicted survival curves of the stratified Cox model to the observed Kaplan-Meier survival probabilities across gender, glucose control, hypertension, type of diabetes and smoking status. In contrast, the threshold regression model was only able to provide predicted scenarios for very specific patient profiles, making comparison with the Cox PH model difficult. This limitation of the threshold regression model has not been observed in other publications. The LMM was able to predict values of Log (ACR) with some accuracy, although a tendency to underestimate the outcome was observed.

The Cox model demonstrated good discrimination, with the model able to distinctly separate the estimated cumulative hazards for the four risk categories. It was not possible to determine Harrell's C as a measure of discrimination, as the dataset was left censored because of late entries into the study. Very little literature addressing this issue was found, and may be an area of further research. The models performed equally well in the validation data set. However, external validation of the data is recommended before the models can be generalised to other study populations.

Different limitations were experienced in the fitting of the models. Although the Cox model allows for stratification, it limits further interpretation of the stratifying variable. For instance, the Cox model stratified by gender, glucose control, hypertension, type of diabetes and smoking status does not provide hazard ratios for these five variables. Because the threshold regression model does not assume proportional hazards, it is able to provide not only hazard ratios for these variables, but also the change in ratios over time (see Table 11). One is able to observe that the hazard ratio for smoking vs never smoked is not proportional as it changes from 5.13 to 1.21 over the 20 years. To a lesser degree, the hazard ratio for gender also changes, with women having a slightly lower risk of DN than men. Although providing such extensive information is an advantage, threshold regression is only able to provide hazard ratios for categorical variables. In comparison to the Cox model's hazard ratios of 1.45 for log (baseline ACR) and 1.99 for total cholesterol (see Table 8), similar ratios can only be determined by the threshold regression model if this data is categorised. Again, these limitations were seldom mentioned in the literature comparing the Cox PH model and threshold regression.

The LMM has the advantage of flexibility in defining a specific variance/covariance structure, but the complexity of computations can result in the iterative procedure not converging and standard errors and confidence intervals for the random effects not being calculated. The model needs to be simplified or re-specified should this occur, lengthening the process of model selection.[2]

A further limitation to the analysis of the data in this study is the possibility of referral bias. Patients were referred to the clinic because they were either diagnosed with DM or were experiencing complications of DM. The inception cohort of patients may thus have been more seriously ill than the general diabetic population, and this would result in the risk of DN being overstated.[66] Also, patients with higher ACR levels are at a greater risk of DN, and may tend to visit the clinic more frequently. This also resulted in more data being collected from the more sickly patients, than those who were managing their DM well.

The comparison of the three models is summarised in Table 15.

Table 15 Summary Comparison of the Cox PH Model, Threshold Regression and Linear Mixed Model

	Stratified Cox PH Model	**Threshold Regression**	**Linear Mixed Model**
Prediction	Very good discrimination. Easy to create risk categories. Very good fit to the observed survival probabilities	Predictions only for a specialised patient profile possible. This is difficult if model has many covariates, and limits comparability across models.	Average prediction was observed in this study. Risk categories determined from the predicted outcome. The LMM was better at predicting lower levels of glucose control.
Advantages	Well–developed software, easy to apply and interpret. Many features such as graphs and descriptive functions of survival probabilities and hazard ratios available Do not have to specify a baseline hazard function.	Provides insight into underlying causal factors and risk patterns. Doesn’t require proportional hazards. Can accommodate data unevenly spaced in time. Explains the change in hazard ratios over time.	Greater flexibility in defining the covariance/variance structure. Allows unequal number of measurements per patient collected at varied time points. Accounts for individual variation in data. All available observations for each patient are used.
Disadvantages	Censoring can complicate and limit analyses available. No hazard ratios available for strata variables.	Not well known. Software and features limited.	Residuals must be multivariate normal. Data must be MAR.
AIC (BIC)	81 (90)	1428 (1497)	3755 (3818)
Significant covariates (*P* <0.1)	Log(Baseline ACR) Total Cholesterol **Strata:** Gender Glucose Control Hypertension Type of Diabetes Smoking Status	**Initial State** Log(Baseline ACR) Baseline HbA1c Gender Type of Diabetes Smoking Status **Process** Duration of DM	**Fixed Effects** Log(Baseline ACR) Log(Creatinine) Systolic Blood Pressure Glucose Control **Random Effects** Smoking Status Total Cholesterol Weight

7 Conclusion

The objective of this study was to compare the Cox PH survival model, threshold regression and LMM with respect to their predictive power and utilitarian value to researchers. Both the Cox PH model and LMM proved to have good predictive power, whereas threshold regression was limited in this regard. All three models are useful in analysing longitudinal or time to event data and identified significant covariates contributing to the onset of DN in diabetic patients. The interpretation of results is further enhanced by the availability of a variety of survival and hazard graphs and ratios, marginal plots and diagnostic techniques.

Although the Cox model has been the traditional method of analysing time-to-event data, a threshold regression model also presents a flexible approach to such analysis, explaining the underlying health process in greater detail. As researchers become more familiar with the threshold regression model and further research and development of the theory is undertaken, the threshold regression model presents itself as a viable alternative to the Cox PH model, especially when the PH assumption is violated. Where the PH assumption holds, however, the Cox model is preferred as a model for survival data, as it has been the tried and tested model for time-to-event data for many years, and has well-developed software applications and support.

The LMM provided an alternate analysis and interpretation of the data, identifying significant variables contributing to a raised ACR level, which, in turn, is a marker for DN. The model made provision for the individual variation in weight and total cholesterol amongst patients, and is a valid analysis option for datasets with significant between subject variations.

Using all three models to explain the data provided a broader understanding of the underlying and contributing factors to the onset of diabetic nephropathy. A multi-faceted approach to the analysis of time-to-event data is highly recommended, and where feasible, all models should be applied to the data.

8 Abbreviations

PH	Proportional Hazards
LMM	Linear Mixed Model
DM	Diabetes Mellitus
DN	Diabetic Nephropathy
ACR	Albumin-Creatinine Ratio
GFR	Glomerular Filtration Rate
ACE	Angiotensin Converting Enzyme
ARB	Angiotensin Receptor Blocker
ANOVA	Analysis of Variance
ANCOVA	Analysis of Covariance
MANCOVA	Multivariate Analysis of Covariance
MCAR	Missing Completely at Random
MAR	Missing at Random
NMAR	Not Missing at Random
MI	Multiple Imputation
LOCF	Last Observation Carried Forward
CC	Complete Case

9 Acknowledgements

I wish to thank Mrs Loveness Dzikiti, Professor Paul Rheeder, Professor Brendan Girdler-Brown from the University of Pretoria, as well as Professor Matthew Chersich from the Wits Institute for Reproductive Health and HIV and Professor Piet Becker from the Medical Research Council for their availability, advice and guidance in completing this dissertation. Thank you, also, to Professor Danie Van Zyl for allowing access to the data collected at the diabetic clinic. Finally, thanks are also due to my family for their endless support and patience at not having a wife and mother for several months.

10 References

1. Lee M-LT, Whitmore GA. Threshold regression for survival analysis: modelling event times by a stochastic process reaching a boundary. Stat Sci. 2006;21(4):501-13.
2. West BT, Welch KB, Galecki AT. Linear mixed models: a practical guide using statistical software. Boca Raton: Chapman & Hall; 2007.
3. International Diabetes Federation [Internet]. IDF diabetes atlas, 6th edn – the global burden. [updated 2014; cited 2014 May 8]. Available from: www.idf.org/sites/default/files/EN_6E_Ch2_the_Global_Burden.pdf.
4. International Diabetes Federation [Internet]. IDF diabetes atlas, 6th edn – Africa at a glance. [updated 2014; cited 2014 May 8]. Available from: www.idf.org/sites/default/files/DA6_Regional_factsheets.pdf.
5. Sobngwi E, Mauvais-Jarvis F, Vexiau P, Mbanya JC, Gautier JF. Diabetes in Africans. Diabetes Metab J. 2001;27(6):628-34.
6. International Diabetes Federation [Internet]. Complications of diabetes. [updated 2014; cited 2014 May 8]. Available from: http://www.idf.org/complications-diabetes.
7. Walker PD, Shah SV. Glomerular diseases. In: Carpenter CCJ, Griggs RC, Losclzo J, editors. Cecil's essentials of medicine. 6th ed. Philadelphia: Saunders; 2004. p. 621-38.
8. National Institute of Health - National Kidney Disease Educational Program [Internet]. Urine albumin to creatinine ratio. [updated March 2010; cited 2014 May 28]. Available from:http://nkdep.nih.gov/resources/quick-reference-uacr-gfr-508.pdf
9. Justesen TI, Damm P, Peterson JLA, Mathiesen ER, Ekbom P. Albumin-to-creatinine ratio in random urine samples might replace 24-h urine collections in screening for micro- and macroalbuminuria in pregnant women with type 1 diabetes. Diab Care. April 2006;29(4):924-925.
10. Gross JL, De Azevedo MJ, Silveiro SP, Canani LH, Caramori ML, Zelmanovitz T. Diabetic nephropathy: diagnosis, prevention, and treatment. Diabetes Care. 2005 Jan;28(1):176-88.
11. Mbanya JC, Sobngwi E. Diabetes microvascular and macrovascular disease in Africa. J Cardiovasc Risk. 2003;10(2):97-102.

12. Williams M, Lacson Jr E, Wang W, Lazarus JM, Hakim R. Glycemic control and extended hemodialysis survival in patients with diabetes mellitus: comparative results of traditional and time-dependent cox model analyses. Clin J Am Soc Nephrol. 2010;5:1595-601.

13. Gall MA, Hougaard P, Borch-Johnsen K, Parving HH. Risk factors for development of incipient and overt diabetic nephropathy in patients with non-insulin dependent diabetes mellitus: prospective, observational study. BMJ. 1997 Mar 15;314:783-8.

14. Gunzler D, Bleyer AJ, Thomas RL, O'Brien A, Russell GB, Sattar A, et al. Diabetic nephropathy in a sibling and albuminuria predict early GFR decline: a prospective cohort study. BMC Nephrol. 2013 Jun 17;14(1):124

15. Al-Rubeaan K, Youssef AM, Subhani SN, Ahmad NA, Al-Sharqawi AH, Al-Mutlaq HM, et al.Diabetic nephropathy and its risk factors in a society with a type 2 diabetes epidemic: a Saudi national diabetes registry-based study. PloS One. 2014;9(2):e88956 doi:10.1371/journal.pone.0088956. [cited 2014 April 6]. Available from: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0088956.

16. Chakkarwar VA. Smoking in diabetic nephropathy: sparks in the fuel tank? World J Diabetes. 2012 Dec 15;3(12):186-195.

17. Bouaziz A, Zidi I, Zidi N, Mnif W, Zinelabidine HT. Nephropathy following type 2 diabetes mellitus in Tunisian population. West Indian Med J. 2012 Dec;61(9):881-9.

18. Barnett PS, Braunstein GD. Diabetes mellitus. In: Carpenter CCJ, Griggs RC, Losclzo J, editors. Cecil's essentials of medicine. 6th ed. Philadelphia: Saunders; 2004.

19. Van Dijk C, Berl T. Pathogenesis of diabetic nephropathy. Rev Endocr Metab Disord. 2004;5:237-48.

20. Jaffiol C. The burden of diabetes in Africa: a major public health problem. Bull Acad Natl Med. 2011 Jun;195(6)1239-53.

21. Gill GV, Mbanya JC, Ramaiya KL, Tesfaye S. A sub-Saharan African perspective of diabetes. Diab. 2009 Jan;52(1):8-16.

22. Choukem SP, Dzudie A, Dehavem M, Halle MP, Doualla MS, Luma H, et al. Comparison of different blood pressure indices for the prediction of prevalent diabetic nephropathy in a sub-Saharan African population with type 2 diabetes. Pan Afr Med J. 2012;11:67.

23. Gning SB, Thiam M, Fall F, Ba-Fall K, Mbaye PS, Fourcade L. Diabetes mellitus in sub-Saharan Africa: epidemiological aspects and management issues. Med Trop (Mars). 2007 Dec;67(6):607-11.

24. Kleinbaum DG. Survival Analysis: a self-learning text. New York: Springer-Verlag; 1996.

25. Klein JP, Moeschberger ML. Survival analysis – techniques for censored and truncated data. 2nd ed. New York: Springer-Verlag; 2003.

26. Su YR, Wang JL. Modelling left-truncated and right-censored survival data with longitudinal covariates. Ann. Stat. 2012;40(3):1465-88.

27. Sparling YH, Younes N, Lachin JM. Parametric survival models for interval-censored data with time-dependent covariates. Biostatistics. 2006;7(4):599-614.

28. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. Stat Med. 2009;28(19):2473-89.

29. Putter H, Sasako M, Hartgrink HH, van de Velde CJ, van Houwelingen JC. Long-term survival with non- proportional hazards: results from the Dutch Gastric Cancer Trial. Stat Med. 2005;30:24(18):2807-21.

30. Bolard P, Quantin C, Esteve J, Faivre J, Abrahamowicz M. Modelling time-dependent hazard ratios in relative survival: application to colon cancer. J Clin epidemiol. 2001;54(10):986-96.

31. Bellera CA, MacGrogan G, Debled M, Tunon de Lara C, Brouste V, Mathoulin-Pélissier S. Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. BMC Med Res. 2010;10:20.

32. Köhler HF, Kowalski LP. A critical appraisal of different survival techniques in oral cancer patients. Eur Arch Otorhinolaryngol. 2012;269(1):295-301.

33. Kasza J, Wraith D, Lamb K, Wolfe R. Survival analysis of time-to-event data in respiratory health research studies. Respirology. 2014;19(4):483-92.

34. Schmid M, Potapov S. A comparison of estimators to evaluate the discriminatory power of time-to-event models. Statist Med. 2012;31:2588–609.

35. Chen HC, Kodell RL, Cheng KF, Chen JJ: Assessment of performance of survival prediction models for cancer prognosis. BMC Med Res Meth. 2012;12:102.

36. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Meth. 2013;13:33.

37. Boberg KM, Rocca G, Egeland T, Bergquist A, Broomé U, Caballeria L, et al. Time-dependent Cox regression model is superior in prediction of prognosis in primary sclerosing cholangitis. Hepatology. 2002 Mar;3;5(3):652-7.

38. Brown H, Prescott R. Applied mixed models in medicine. 2nd ed. Chichester: Wiley & sons; 2006.

39. StataCorp. Stata Longitudinal data / panel data reference manual-release 12. Xtmixed – multilevel mixed effects linear regression. Texas: Stata Press; 2011.

40. Hedeker D. Generalised linear mixed models. In: Everitt BS, Howell D, editors. Encyclopaedia of statistics in behavioral science. John Wiley & Sons; 2005.

41. Hayat MJ, Hedlin H. Modern statistical modelling approaches for analysing repeated-measures data. Nurs Res. 2012 June;61(3):188-94.

42. Petkova E, Teresi J. Some statistical issues in the analyses of data from longitudinal studies of elderly chronic care populations. Psychosom Med. 2002;64:531-47.

43. Minalu G, Aerts M, Coenen S, Versporten A, Muller A, Adriaenssens N, et al. Application of mixed-effects models to study the country-specific outpatient antibiotic use in Europe: a tutorial on longitudinal data analysis. J Antimicrob Chemother. 2011;66(6):79-87.

44. Bondell HD, Krishna A, Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. Biometrics. 2010 Dec;66(4):1069–77.

45. Bandyopadhya D, Lachos VH, Castro LM, Dey DK. Skew-normal/independent linear mixed models for censored responses with applications to HIV viral loads. Biom J. 2012;54(3):405-25.

46. Bouwmeester W, Twisk JWR, Kappen TH, Van Klei WA, Moons KGM, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. BMC Med Res Meth. 2013;13:19.

47. Zewotir T, Galpin J. Influence diagnostics for linear mixed models. J. Data Science. 2005;3:153-77.

48. Nobre JS, Singer JDM. Residual analysis for linear mixed models. Biom J. 2007;49(6):863-75.

49. Mun J, Lindstrom MJ. Diagnostics for repeated measurements in linear mixed effects models. Stat Med. 2013;32(8):1361-75.

50. Zucker DM, Manor O, Gubman Y. Power comparison of summary measure, mixed model and survival analysis methods for analysis of repeated-measures trials. J Biopharm Stat. 2012;22:519-34.

51. Xiao T, Whitmore GA, He X, Lee MLT. Threshold regression for time to event analysis: the stthreg package. Stata. 2012;12(2):257-83.

52. Lee MLT, Whitmore GA. Proportional hazards and threshold regression: their theoretical and practical connections. Lifetime Data Anal. 2010;16:196-214.

53. Lee MLT, Whitmore GA, Rosner B. Benefits of threshold regression: a case-study comparison with Cox proportional hazards regression. In: Rykov VV et al, editors. Mathematical and statistical models and methods in reliability: applications to medicine, finance and quality control. Basel: Birkhäuser; 2010:359-70.

54. Lee MLT, Whitmore GA, Rosner BA. Threshold regression for survival data with time-varying covariates. Stat Med. 2010;29:896-905.

55. Lee MLT, Whitmore GA, Laden F, Hart JE, Garshick E. A case-control study relating railroad worker mortality to diesel exhaust exposure using a threshold regression model. J Stat Plan Inference. 2009; 139(5):1633-42.

56. Whitmore GA, Su Y. Modeling low birth weights using threshold regression: results for US birth data..Lifetime Data Anal. 2007;13:161-90.

57. Siddique J, Brown CH, Hedeker D, Duan N, Gibbons RD, Miranda J, et al. Missing data in longitudinal trials – part B, analytic issues. Psychiatr Ann. 2008 Dec 1;38(12):793-801.

58. Rubin, DB. Inference and missing data. Biometrika. 1976;63(3):581-92.

59. Little TD, Jorgenson TD, Lang KM, Moore EW. On the joys of missing data. J Pediatr Psychol. 2014;39(2):151-62.

60. Bell ML, Fairclough DL .Practical and statistical issues in missing data for longitudinal patient reported outcomes. Stat Methods Med Res. [Online] 2013 Feb 19. (Cited 2013 July 11) Available from: http://smm.sagepub.com/content/early/2013/02/14/0962280213476378

61. Wayman JC. Multiple imputation for missing data: what is it and how can I use it? Paper presented at Annual Meeting of the American Educational Research Association; 2003; Chicago: USA.

62. Allison PD. Handling missing data by maximum likelihood. Paper presented at SAS Global Forum; 23 April 2012; Orlando: USA.

63. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Medical Research Methodology. 2010;10:112.

64. Grittner U, Gmel G, Ripatti S, Bloomfield K, Wicki M. Missing value imputation in longitudinal measures of alcohol consumption. Int J Methods Psychiatr Res. 2011 Mar;20(1):50-61.

65. Describing the distribution of failure times. [Internet] [Cited 2013 Sept 14]. Available from: http://www.fordham.edu/economics/mcleod/SUAS_Chapter2B.pdf

66. Emory University School of Medicine [Internet]. More on inception cohorts. [Cited 2014 May 28]. Available from: http://www.med.emory.edu/EMAC/curriculum/prognosis/inception.html

11 Appendices

11.1 Graphs and Tables

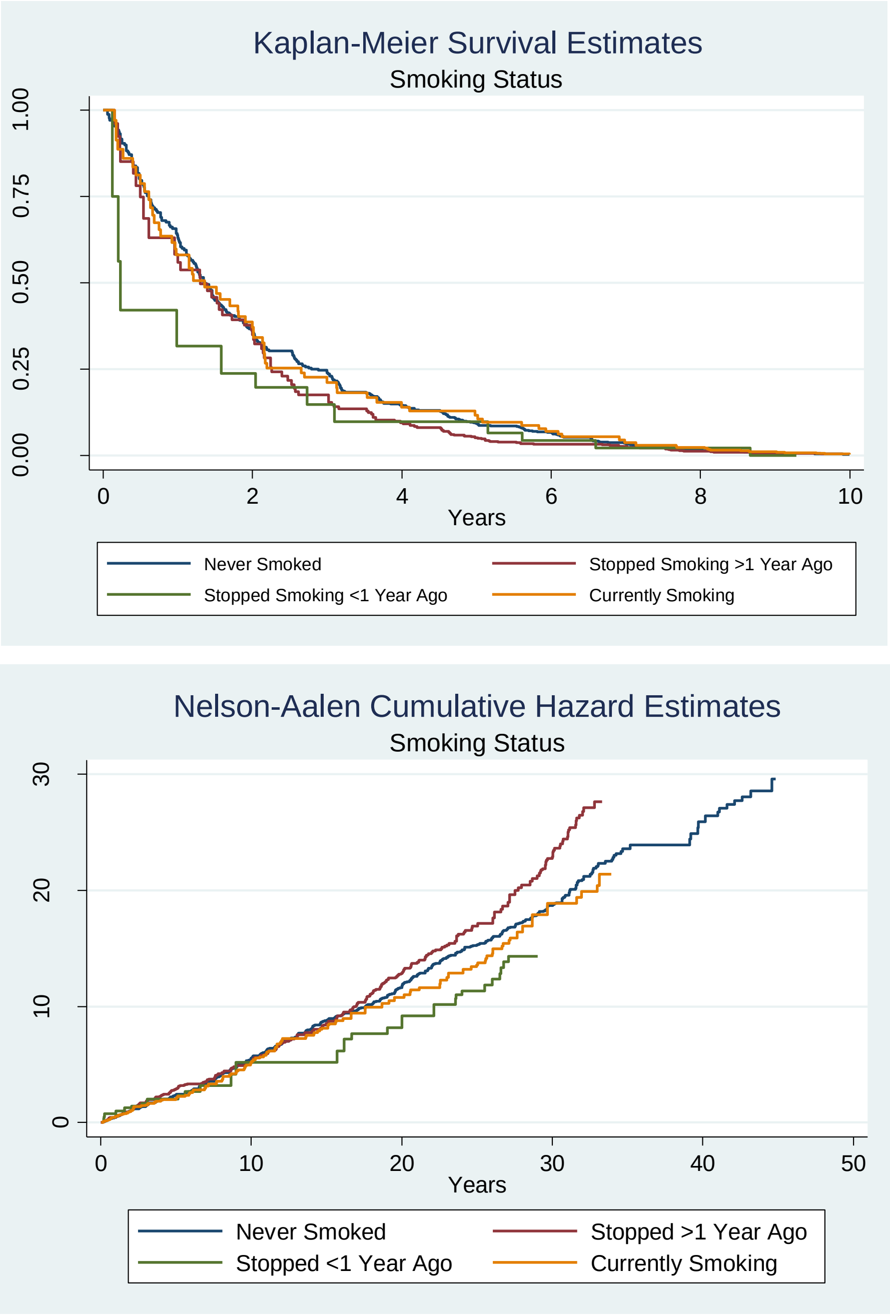


Figure 18 Kaplan-Meier Survival Curves and Nelson-Aalen Cumulative Hazard Estimates according to Smoking Status

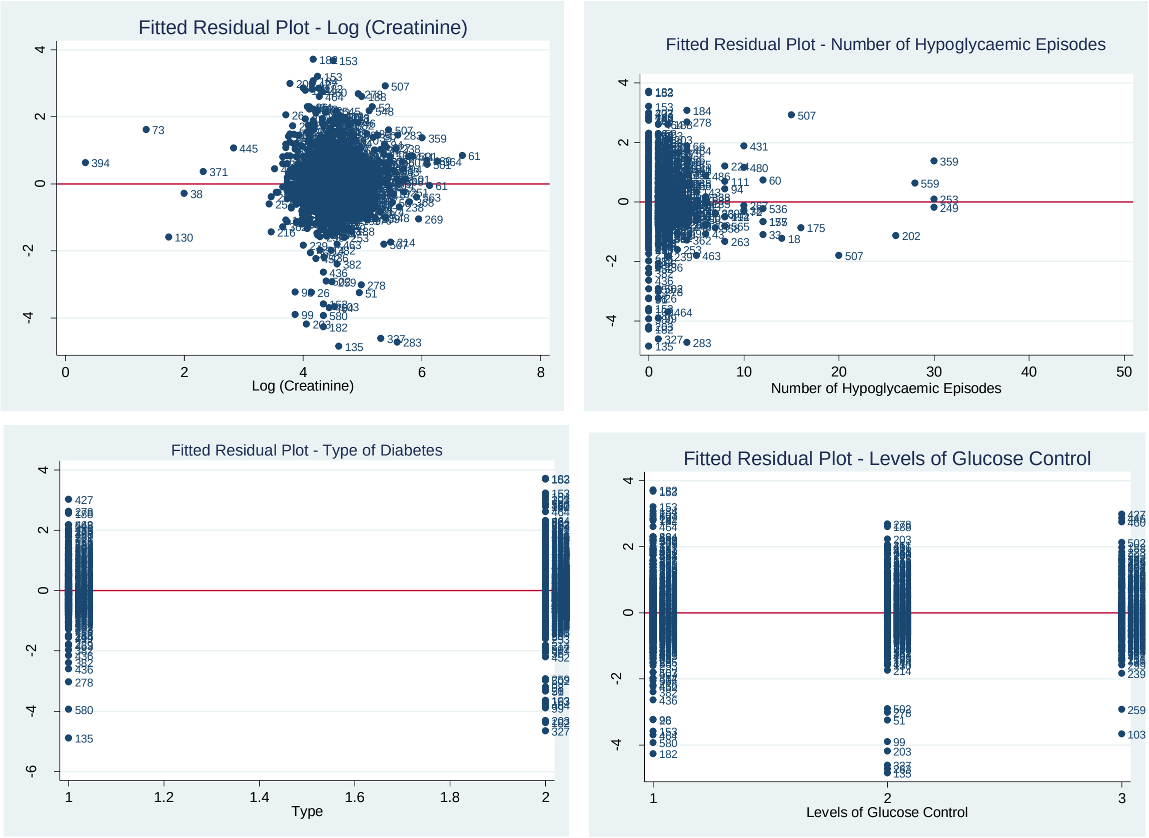


Figure 19 LMM: Fitted Residual Plots according to Log (Creatinine), Number of Hypoglycaemic Episodes, Type of Diabetes and Level of Glucose Control

Table 16 Reclassification of DN Categories – Validation Data

Observed DN Categories	Predicted DN Categories		
	0	**1**	**2**
0	366	45	0
1	42	193	7
2	2	43	53

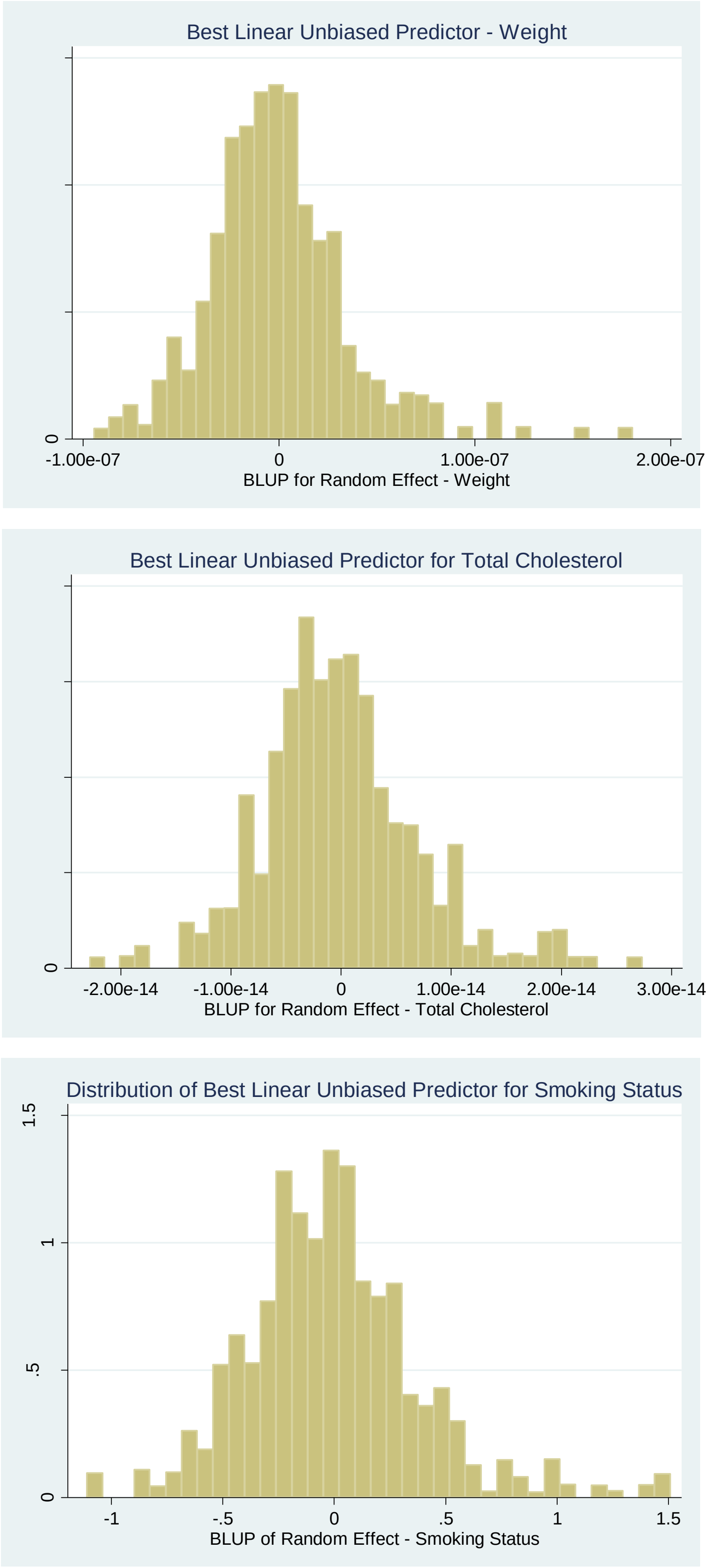


Figure 20 Histograms of the Best Linear Unbiased Predictors for the LMM Random Effects –Weight, Total Cholesterol and Smoking Status

STATA Code

STATA Output