
**UNIVERSITY OF MESSINA**

DEPARTMENT OF ECONOMICS

PhD IN ECONOMICS, MANAGEMENT AND STATISTICS,

CYCLE XXXVIII

---

Doctoral Thesis Title

# Joint Modelling of Multivariate Longitudinal and Time-to-event Data under Bayesian Inference: with Application to Type 2 Diabetes and Hypertension Disease.

**Ph.D. Candidate:**

**Mequanent Wale Mekonen**

**Supervisors:**

Prof. EDOARDO OTRANTO

Prof. ANGELA ALIBRANDI

**Ph.D. Coordinator:**

Prof. DARIO MAIMONE ANSALDO PATTI

A.Y. 2024/25

# Contents

# List of Tables

# List of Figures

# List of Acronyms

| | |
|---|---|
| MCMC | Markov chain Monte Carlo |
| T2D | Type 2 Diabetes |
| BMI | Body Mass Index |
| LDL | low-density lipoprotein |
| HDL | High-Density Lipoprotein |
| DKD | Diabetic Kidney Disease |
| PCA | Principal Component Analysis |
| JAGS | Just Another Gibbs Sampler |
| lMM | Linear Mixed Models |
| PDF | Probability Density Function |
| CGE | cumulative Distribution Function |
| DIC | Deviance Information Criterion |
| SBP | Systolic Blood Pressure |
| DBP | Dystolic Blood Pressure |
| RRS | Residual sum of Squares |
| RMSE | Root Mean Square Error |
| HbA1c | Glycosylated Haemoglobin |
| eGFR | Estimated glomerular filtration rate |
| KDIGO | Kidney Disease: Improving Global Outcomes |
| GGT | Gamma-Glutamyl Transferase |
| GPT | Glutamate Pyruvate Transaminase |
| GOT | Glutamate Oxaloacetate Transaminase |
| MLM | Maximum Likelihood Estimation |
| AIC | Akaike Information Criterion |
| BIC | Bayesian Information Criterion |

# Acknowledgement

First and foremost, I would also like to sincerely thank all my professors for their invaluable support during my PhD journey. I am especially grateful to Professors Edoardo Otranto, Angela Alibrandi, and Zeytu Asfaw. Your guidance, insightful feedback, and constant encouragement have played a crucial role in shaping this thesis.

I would also like to sincerely thank Professor Fabrizio Cesaroni, the former PhD coordinator, whose guidance from the beginning of my program continued throughout my entire PhD journey. His advice and insights have played a significant role in my academic and professional development. I sincerely thank the current PhD coordinator, Prof. Dario Maimone Ansaldo Patti, for his smooth communication and for always being available to answer my questions and provide constructive feedback. I am also grateful to the University of Messina for its generous support and for providing the educational environment and resources necessary to complete this research. I am thankful to my wife, Haymanot Baylie, for her unwavering support. She helped me make important decisions and cared for our child during my absence, enabling me to dedicate myself fully to my studies.

I am also deeply thankful to my friend, Dr. Endeshaw Asea, for his ongoing encouragement and motivation during the most challenging moments of this journey. Lastly, I dedicate this PhD to my courage, determination, and passion for the work I love. I also want to thank everyone I met along the way, including friends and colleagues, who inspired me to continually do my best.

# Introduction

### Overall Aim and Motivation

The overarching aim of this thesis is to propose statistical methods for analysing correlated chronic diseases, specifically type 2 diabetes (T2D), hypertension, and their associated complications. These conditions represent multiple chronic disease outcomes, for which appropriate statistical models are needed to understand the progression of glucose concentration and blood pressure, as well as their relationship with complications. Such methods can enhance our understanding of how multiple disease processes, such as diabetes and hypertension, co-develop with chronic complications and influence each other over time.

The statistical methods explored in this thesis focus on semiparametric mixed-effect models, joint models that integrate longitudinal data with time-to-event processes, incorporating one or multiple longitudinal outcomes, and principal component analysis to show how clinical and biochemical variables are interrelated. These approaches are particularly relevant in settings where subjects are followed over time to monitor disease progression or medical conditions until an event occurs or censoring takes place. Progression is typically assessed through repeated measurements of biomarkers relevant to the disease, and a key clinical interest is often to determine the effect of these biomarkers on the time to an event of interest.

#### Type 2 Diabetes and Hypertension: Chronic Complications

Type 2 diabetes and hypertension frequently coexist, and persons with diabetes have a two-to-four-fold increased risk of hypertension compared with persons without diabetes. Each increases the risk of chronic complications independently, but when they co-occur, the risk increases significantly. The most frequent chronic complications of hypertension and diabetes are cardiovascular disease, chronic kidney disease, stroke, and retinopathy, all of which contribute significantly to mortality. These complications start to arise following the diagnosis of T2D and hypertension and are thought to result primarily from

prolonged exposure to elevated glucose concentrations and blood pressure. Moreover, the clinical and biochemical variables, including body mass index (BMI), low-density lipoprotein (LDL), high-density lipoprotein (HDL), and triglycerides, may be associated with both the trajectories of blood glucose and blood pressure, as well as the onset of chronic complications.

A useful model is needed to understand how correlated chronic disease processes co-develop over time; a review of analytical approaches for our overarching research aim highlights three key statistical methods gaps. First, univariate analysis of correlated longitudinal outcomes ignores the underlying correlation structure and may lead to incorrect inferences. For example, analyzing glucose concentration in individuals with T2D and hypertension using a univariate mixed-effects approach would fail to account for the inherent association between blood pressure and glucose concentration, and vice versa. The multivariate linear mixed model has become the most important and widely used analytical tool for longitudinal data with multiple continuous time series. However, the classical assumptions of multivariate normality for model errors and random effects, as well as a strictly linear relationship between model parameters and the response, may not always be realistic. Therefore, flexible parametric mixed-effects models for skewed longitudinal data with multiple outcomes are essential for obtaining more reliable, less biased statistical inferences.

Second, the time-to-event outcome is often collected alongside one or more longitudinal biomarkers, which are repeatedly measured until the event occurs or censoring occurs. Survival models have typically characterized exposures by a single measure at study baseline or as an average over the study period. Such exposure characterization fails to incorporate the overall characteristics of the longitudinal process. Joint modeling is an approach generally used to model two response processes simultaneously: longitudinal (consisting of repeated measurements) and survival (consisting of time-to-event Processes). However, most joint models introduced in the literature so far have focused on modelling only one longitudinal variable with one time-to-event outcome, herein referred

to as univariate joint modelling.

In fact, if one or more correlated longitudinal biomarkers are repeatedly measured until an event occurs, ignoring this correlation and applying univariate joint modelling may lead to biased estimates. Moreover, much of the existing literature on joint models assumes that longitudinal measurements follow a multivariate normal distribution, which may be unrealistic in practice. To address this, multivariate joint models that accommodate multiple skewed longitudinal outcomes incorporate more information and improve the model's prognostic ability by accounting for correlated longitudinal outcomes and their underlying distributional properties, while remaining theoretically straightforward.

The multivariate joint modeling framework for longitudinal and survival data consists of three components: (i) a model for time-to-event outcomes, (ii) a model for longitudinal marker trajectories, and (iii) a set of shared parameters that link the two processes. Compared with separate models, this framework improves statistical inference by addressing measurement error in time-dependent covariates within survival regression models, enhancing statistical efficiency through the joint use of correlated longitudinal data on non-fatal chronic diseases, and enabling the quantification of associations between longitudinal markers and survival processes, including biochemical variables.

Third, several clinical and biochemical variables have been reported to be associated with microvascular and macrovascular complications in people with T2D. However, the interrelationships among these variables could distort the estimation of their effects on disease progression. Thus, addressing collinearity is essential not only to ensure unbiased statistical inference but also to provide insight into the intercorrelation among biochemical and clinical variables in individuals with type 2 diabetes.

Thus, novel methods that jointly model multiple correlated, skewed longitudinal outcomes with nonlinear trajectories over time, while also addressing the joint analysis of longitudinal and time-to-event processes, accounting for correlation among outcomes, departures from normality, and providing insight into how clinical and biochemical variables are interrelated, remain limited. The central aim of this thesis is to contribute to the

application of such novel methods for correlated chronic disease.

These novel methods can then address clinical objectives such as the following: To what extent are chronic complications occurring for people with diabetes and hypertension? When these chronic complications occurred, and how the trajectory of chronic disease over time, such as diabetes and hypertension, jointly affect the time to event outcomes? To what extent do the progression of two or more chronic diseases correlate with each other over time? How biochemical and clinical variables in people with type 2 diabetes and hypertension are intercorrelated with each other and their effects on the progression of microvascular and macrovascular complications. Since the natural history of certain complications can only be assessed intermittently, the model must account for the characteristics of the study's repeated-measures process or observation scheme.

To address this clinical research objective and apply relatively novel statistical methods, the thesis is organized into three chapters. To illustrate these methods, a retrospective cohort study design was employed, and data were collected from individuals with T2D mellitus and hypertension treated at Felege Hiwot Comprehensive Specialized Hospital in Ethiopia. Additionally, datasets from patients with type 2 diabetes at a polyclinic hospital in Italy were utilized. The classical (frequentist) approach to parameter estimation, which relies on the joint likelihood function, is a well-established method for estimating unknown parameters. However, a significant statistical challenge arises from the computational difficulties associated with high-dimensional random effects. To address this, a Bayesian framework is proposed. Accordingly, a Bayesian approach to statistical inference was adopted in the first two chapters, and the proposed models were assessed through simulation studies.

**Chapter 1. Semiparametric Multivariate Mixed-effects Model for Skewed Longitudinal Data under the Bayesian Approach.**

The objective of the first chapter is to introduce flexible semiparametric multivariate mixed models for analysing multiple correlated chronic disease processes. This approach addresses the complex statistical challenges inherent in longitudinal data, particularly the

impact of deviations from multivariate normality and the effect of nonlinear trajectory patterns on model results. The proposed models are applied to examine the relationship between the trajectories of glucose concentration and blood pressure over time in people with diabetes and hypertension. To assess how the model parameters are estimated and to compare model performance under varying conditions, simulation studies were conducted. Both the application dataset and the simulation results suggest that the flexible parametric multivariate mixed model is effective, yielding minimal bias and a low root-mean-square error in parameter estimation when accounting for skewness and nonlinear trajectory patterns in correlated outcomes.

**Chapter 2. Bayesian Joint Modelling of Bivariate Longitudinal and Time-to-Event Data: With Application of Micro and Macro Vascular Complications in people with Type 2 Diabetes and Hypertension.**

We propose the bivariate joint modeling to address the clinical question of how the trajectories of glucose concentration and blood pressure are related and how their joint progression influences the risk of developing microvascular and macrovascular complications. A simulation study was also conducted to compare parameter-estimation performance under different distributional assumptions for multiple longitudinal outcomes. Both the application and simulation studies show that a multivariate joint model with a multivariate skew normal distribution provides more efficient and relatively accurate parameter estimation.

**Chapter Three. Diabetic Kidney Disease Progression in People with Type 2 Diabetes using Principal Component Analysis and Ordered Logit Model.**

The objective of this chapter is to investigate the intercorrelation of biochemical and clinical variables in individuals with type 2 diabetes and their effects on the progression of kidney disease. Diabetic kidney disease (DKD) is one of the major microvascular complications of diabetes, characterized over time by a decline in kidney function, typically measured by the estimated glomerular filtration rate. To explore these interrelationships and their impact on disease progression, principal component analysis (PCA) combined

with ordered logit models was employed. From the available clinical and biochemical variables, three uncorrelated principal components were extracted. The first component, a linear combination of glycosylated haemoglobin, glycemia, and creatinine, which are positively correlated, showed a strong, statistically significant association with the progression of kidney disease.

**Overall significance**

Advanced statistical methods have broad applicability in both epidemiological and clinical research. Our semiparametric multivariate mixed-effects model, multivariate joint modelling approach, and PCA and ordered logit models can provide physicians with more precise information and stronger evidence on how chronic disease processes and chronic complications co-develop and influence each other over time. This information is crucial for enhancing patients' understanding of disease progression and supporting clinical decision-making. Although this research is motivated by chronic diseases such as hypertension, type 2 diabetes, and their related complications, the fundamental concepts underlying the proposed methods are broadly applicable to other fields, provided that the relevant technical assumptions are met. These statistical methods provide a methodological reference for biostatistics and epidemiology researchers interested in conducting studies within this and related domains.

# Chapter 1

# Semiparametric Multivariate Mixed-effects Model for Skewed Longitudinal Data under the Bayesian Approach.


## Abstract

In medical research, it is common to collect multivariate data by measuring subjects multiple times across different outcomes. They often use univariate mixed-effects models to analyze this data by assuming that both the random effects and errors follow a normal distribution. Additionally, the response variables are considered to be linear functions of the unknown regression parameters. However, the assumption of normality may not always yield reliable results if the data exhibit skewness; outcome variables may have a nonlinear relationship with some covariates, such as time. Furthermore, a univariate mixed-effects model applied to correlated multivariate longitudinal outcomes without accounting for their correlation may yield biased parameter estimates. To address these issues simultaneously, we propose a flexible semiparametric multivariate mixed-effects model that incorporates multiple longitudinal exposures that are significantly correlated, exhibit skewness, and use a nonparametric function to capture nonlinear time effects. The proposed models are illustrated through an application to correlated glucose concentration and blood pressure data to study the association of glucose concentration and blood pressure in individuals with T2D and hypertension. A simulation study is conducted to evaluate the performance of the proposed models. The results of both the application and simulation studies suggest that the semiparametric mixed-effect model under a skew multivariate normal distribution for the random errors performs better than other proposed models, as it accommodates the nonlinear effects of covariates and the asymmetry of longitudinal measurements. In our application, we found a strong association between

the changes in glucose concentration and blood pressure, with the rate of change increasing over time.

**Keywords:** Semiparametric multivariate mixed effect model, skewed multivariate longitudinal data, multivariate skew normal distribution, type 2 diabetes, and hypertension disease.

## 1.1 Introduction

Longitudinal data are often collected in clinical and other follow-up studies, in which a cohort of subjects is monitored, and data are collected at multiple time points to assess how various exposures, processes, or characteristics influence outcomes over time. For example, glucose concentration and blood pressure are repeatedly measured in people with diabetes and hypertension to monitor disease progression over time and to ensure patient safety.

This research was driven by follow-up data obtained from people with type 2 diabetes and hypertension, a significant global health issue that affects both developed and developing countries. Type 2 diabetes and hypertension are chronic diseases that significantly affect populations in both developing and developed countries. In 2021, the global prevalence of diabetes was estimated to be 10.5% (536.6 million) and is projected to reach 12.2% (783.2 million) by 2045. In the African region, the prevalence was 4.5% (23.6 million) in 2021 and is expected to increase to 5.2% (54.9 million) by 2045 (Sun et al., 2022). Type 2 diabetes and hypertension are rapidly rising non-communicable diseases and significant public health challenges in Ethiopia, resulting in disability and premature death due to the long-term effects of untreated diabetes mellitus. The prevalence of hypertension was predicted to be 19.2% (95% CI: 18.4 to 20.0) and 2.8% (95% CI: 2.4 to 3.1) for diabetes in Ethiopia. Substantial variation was observed in the prevalence of these diseases at subnational levels, with the highest prevalence of hypertension observed in Addis Ababa (30.6%) and diabetes in the Somali region (8.7%) (Koye et al., 2022)

The rapid increase of diabetes and hypertension significantly increases the risk of macro-

and microvascular complications, including coronary artery disease, stroke, nephropathy, and retinopathy, leading to rising mortality (Liu et al., 2021) and (Petrie et al., 2018). These conditions represent major public health and socioeconomic challenges. Effective control of key biomarkers, such as blood glucose concentration and blood pressure, can significantly reduce chronic complications and reduce related deaths, although complete prevention remains unattainable (Reaven et al., 2019). These biomarkers provide valuable information about the functioning of the circulatory system and serve as indicators of an individual's cardiovascular health at a specific time point. In clinical practice, individuals diagnosed with hypertension and diabetes undergo routine measurement of glucose concentration and blood pressure at each follow-up visit. These longitudinal assessments serve a critical role in estimating the trajectories of glucose and blood pressure over time.

Moreover, it is essential to quantify treatment efficacy and to inform and optimize individualized patient care strategies. Therefore, studying the trajectory of glucose concentration and blood pressure in individuals with diabetes and hypertension is critical for understanding disease pathogenesis and enhancing patient care. In follow-up studies, longitudinal data exhibit a variety of features over time and across subjects in many real-world settings. Selecting appropriate methods for analysing such longitudinal data is therefore essential to provide an unbiased inference. Linear mixed models (LMMs) are a powerful class of statistical models frequently used in the analysis of longitudinal data because they are more flexible in modelling intra-subject correlations within the same subject (Laird and Ware, 1982).

Although the various extensions of linear mixed-effects models have been applied for modelling the longitudinal data, for example, parametric nonlinear mixed-effects (Wu and Ding, 1999) and (Wu et al., 2004), and sim-parametric nonlinear mixed-effects (Ferede et al., 2024) and (Huang et al., 2011), nevertheless, in all of these studies, at least one of the following issues stands out:

I) Univariate analysis of correlated longitudinal outcomes ignores the underlying correlation structure and may lead to incorrect inferences. For example, analysing glucose

concentration in individuals with type 2 diabetes and hypertension using a univariate mixed effect approach would fail to account for the inherent association between blood pressure and glucose concentration, and vice versa (Aniley et al., 2025; Huang et al., 2015; Ghosh et al., 2007). Even though most analyses use a univariate mixed-effect model, the multivariate linear mixed model (MLMM) has become the most important and commonly used analytical tool for multiple longitudinal data, for example, (Bandyopadhyay et al., 2010) and (Lin and Wang, 2013). All of these studies suggest that multivariate analyses yield better results than univariate analyses, implying that if two or more longitudinal outcomes are correlated, analysing them separately may produce biased estimates of effect sizes.

II) The assumption of a linear relationship between model parameters and the response may not always be realistic. For instance, the lower panel of Figure 1.2 illustrates the relationship between fasting blood sugar and time as well as systolic blood pressure and time among patients with diabetes and hypertension receiving treatment at Felege Hiwot Referral Hospital. The trends observed in both glucose concentration and blood pressure are not linear, indicating that modelling these patterns using a linear mixed effects framework may result in a biased estimator (Yirdaw and Debusho, 2023) and (Aniley et al., 2019).

III) Furthermore, multivariate normal distributions for model errors and random effects are common assumptions in all mixed-effects models. However, this assumption, which may not hold in practice, can mask critical aspects of both between- and within-subject variability. For example, the upper panel of Figure 1.2 displays histograms of repeated glucose concentration and blood pressure measurements for diabetic and hypertensive subjects recruited from the diabetic and hypertensive registration book at Felegehiwot comprehensive specialized hospital; these measurements are notably right-skewed, primarily due to a subset of participants exhibiting exceptionally high glucose and systolic blood pressure levels compared to the rest of the study subjects. Thus, conducting mixed effects under the assumption of multivariate normality in both random effects and ran-

dom errors can result in biased statistical inferences (Verbeke and Lesaffre, 1996) and (Arellano-Valle et al., 2007)

There are several suggestions in the literature for accommodating the asymmetry of longitudinal data, including the multivariate skew-normal, skew-normal/independent (SNI), and multivariate skew-t distributions, to produce robust parameter estimates. The skew-normal distribution was first introduced by (Azzalini, 1985), providing a more flexible approach to modelling asymmetrical data. Since then, a detailed review of skew distributions has been extensively developed (Azzalini and Capitanio, 1999) and (Azzalini and Valle, 1996) and others over the past three decades to model skewed longitudinal data. Depending on the sign of the skewness parameter, the distribution accommodates both positively and negatively skewed data and includes normal distributions as exceptional cases. (Sahu et al., 2003) proposed multivariate skew-normal distributions for Bayesian regression problems; these differ from the skew-elliptical version proposed by (Azzalini, 1985), which is based on a univariate normal distribution. To the best of our knowledge, no study has simultaneously addressed the asymmetric distribution of both random errors and random effects, correlated longitudinal outcomes, and nonlinear time effects within the framework of Bayesian mixed-effects modelling. Thus, we present flexible semiparametric multivariable mixed-effects modelling for skewed longitudinal data with multiple outcomes, introduce Bayesian inference, use a real dataset from people with diabetes and hypertension to illustrate the proposed models, and conduct simulation studies to validate our conclusions.

### 1.1.1 The Multivariate Skew-Normal Distribution

Several extensions of the multivariate normal distribution, such as the multivariate skew-normal and multivariate skew-t distributions (Sahu et al., 2003) and (Arellano-Valle et al., 2007) have been used in the literature to capture asymmetry in longitudinal data. In this study, we proposed the multivariate skew-elliptical distribution introduced by (Sahu et al., 2003), which is implementable for conducting Bayesian inference. A K-dimensional random vector Y follows a k-variate skew normal distribution with $k \times 1$ location vector

$\mu$, $k \times k$ positive definite dispersion matrix $\Sigma$, and $k \times 1$ skewness diagonal matrix $\Delta = \text{diag}(\delta)$ with $\delta = (\delta_1, \delta_2, \ldots, \delta_k)$ being a skewness parameter vector, if its probability density function (PDF) is given by (Sahu et al., 2003):

$$\begin{aligned} f(Y \mid \mu, \Sigma, \Delta) = 2^k \, |\Sigma + \Delta|^{-1/2} \, \phi_k\big[(\Sigma + \Delta)^{-1/2}(Y - \mu)\big] \\ \times \Phi_k\Big[\big(I_k - \Delta(\Sigma + \Delta^2)^{-1}\Delta\big)^{-1/2} \, \Delta(\Sigma + \Delta^2)^{-1}(Y - \mu)\Big] \end{aligned} \tag{1.1}$$

Where $\phi_k$ and $\Phi_k$ represent the pdf and cumulative distribution function (CDF) of the multivariate normal distribution of the random variable Y. $I_k$ denotes the $k \times k$ identity matrix. We denote this distribution in short-hand representation as

$$Y \sim \text{SN}_k(\mu, \Sigma, \Delta). \tag{1.2}$$

The mean and covariance are given by $E(Y) = \mu + \sqrt{\frac{2}{\pi}}\,\delta, \text{Cov}(Y) = \Sigma + \left(1 - \frac{2}{\pi}\right)\Delta^2$. When $\delta = 0$, the $k$-variate skew-normal distribution reduces to the standard $k$-variate normal distribution, and if $\Sigma = \sigma^2 I_k$, the density function factorizes into the product of independent marginal distributions. If $\delta > 0$, the distribution is positively skewed (right-skewed), while if $\delta < 0$, the distribution is negatively skewed (left-skewed).

According to Sahu et al. (2003), if $Y \sim \text{SN}_k(\mu, \Sigma, \Delta)$, it can be formulated through a convenient stochastic representation as.

$$Y = \mu + \Delta|x_0| + \Sigma^{1/2}x_1, \tag{1.3}$$

where $x_0$ and $x_1$ are two independent random vectors distributed as $N_k(0, I_k)$. Let $w = |x_0|$, then $w$ follows a $k$-dimensional standard normal distribution $N_k(0, I_k)$ restricted to the space with $w > 0$. Thus, the hierarchical representation of equation 1.3 can be expressed as:

$$Y \mid w \sim N_k(\mu + \Delta w, \Sigma), \qquad w \sim N_k(0, I_k)\,\mathbf{1}(w > 0) \tag{1.4}$$

## 1.2 Methodology

### 1.2.1 Semiparametric Multivariate Mixed-effects Models

In this section, we introduce a semiparametric multivariate mixed-effects modelling framework for correlated multivariate responses. The model is presented in a general form to highlight its flexibility and potential applicability across a broad range of biomedical and longitudinal data analysis contexts. Let $Y_{ijk} = \left(Y_{i1k}, Y_{i2k}, \ldots, Y_{in_ik}\right)^T$, $k = 1, 2$, be the repeated measurements of the first and second outcomes, respectively, for the $i^{\text{th}}$ subject measured at time $t_{ij}$, $i = 1, 2, \ldots, m$, $j = 0, 1, 2, \ldots, n_i$. $X_{ijk}$ and $Z_{ijk}$ be the design matrices of size $n_i \times p$ and $n_i \times q$ associated withthe vector of fixed effects $\beta_k$ and the vector of random effects $b_{ik}$. Furthermore, $N_{ik}(t_{ik})$ is an unknown smooth function which is used to estimate the effect of measurement time on the $k$th longitudina outcomes, and $e_{ik}$ is the vector of residuals with $e_{i1}$ and $e_{i2}$ for the first and the second outcomes. As suggested by Ferede et al. (2024) and Huang et al. (2011), we consider a semiparametric multivariate mixed effect model within subject variation (random errors) are assumed to follow a multivariate skew normal distribution, and the between-subject variation (random effects) is believed to have a multivariate skew normal distribution and can be expressed as:

$$Y_{ijk} = \beta_k X_{ik} + N_{ik} t_{ijk} + b_{ik} Z_{ik} + e_{ijk} \tag{1.5}$$

$$\text{where } b_i = \left[b_{i11}, b_{i21}, b_{i12}, b_{i22}\right] \sim SN(0, \Delta_{bk},\ \Sigma_b),\ \Sigma_b = \begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \\ v_{41} & v_{42} & v_{43} & v_{44} \end{bmatrix}$$

$$e_{ik} = \begin{bmatrix} e_{i1} \\ e_{i2} \end{bmatrix} \sim SN(0, \Delta_{ek},\ \Sigma_k)\ , \Sigma_k = \begin{bmatrix} \text{var}(\varepsilon_{i1}) & \text{cov}(\varepsilon_{i1}, \varepsilon_{i2}) \\ \text{cov}(\varepsilon_{i2}, \varepsilon_{i1}) & \text{var}(\varepsilon_{i2}) \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}.$$

$\Delta_{bk}$ is a skewness diagonal matrix for random effects with element $\delta_{11}, \delta_{12}, \delta_{21}, \delta_{22}$ and $\Delta_{ek}$ is the skewness diagonal matrix for the random errors, defined as $\Delta_{ejk} = \text{diag}(\delta_{eik1}, \delta_{eik2},$

$\ldots, \delta_{eikn_i})$. In the case where $\delta_{eik} = \delta_{ei1k} = \delta_{ei2k}, = \ldots, = \delta_{en_ik} = \delta_{ek}$, then the matrix simplifies to $\Delta_{ek} = \delta_{ek} I_{n_i}$, indicates that the model captures the dataset's overall skewness. In this study, the parameters $\delta_{e1}$ and $\delta_{e2}$ quantify the skewness of the first and second outcomes, respectively. The random vectors $b_{ik}$ and $e_{ik}$ are assumed to be independent. The semiparametric multivariate mixed-effects model in equation 1.5 offers greater flexibility compared to its parametric counterpart, particularly in capturing nonlinear time effects. It simplifies to a parametric multivariate mixed-effects model when the influence of time on the response is linear. Note that, in the above model, the correlation between the bivariate responses is represented by both the variance-covariance of the random effects ($\Sigma_b$) and the variance-covariance of the random errors ($\Sigma_k$). For instance, in part $\Sigma_b$, the value of $v_{24}$ indicates the association between the random slopes of two longitudinal outcomes. Similarly, in part $\Sigma_k$, the value of $\sigma^2_{12}$ shows how the residual variance of the two longitudinal outcomes is associated.

In the moel equation 1.5, the effect of measurement time $t_{ijk} = \left(t_{i1k}, t_{i2k}, \ldots, t_{in_ik}\right)^T$ the $k$th longitudina outcomes is non-linear and estimated using nonparametric approach. This is achieved by employing a smoothing function $N_{ik}(t_{ik})$, which can be defined as follows:

$$N_{ik}(t_{ik}) = f(N_{ik}(t_{ik})) = U(t_{ik})$$

Where $U(t_{ik})$ represents unknown smoothing functions for the population variations of the longitudinal response $Y_{ik}$ due to time effects $t_{ik}$. A regression spline base, natural cubic basis is used to specify the unknown functions $U(t_{ik})$ because this method is more straightforward in application (Silverman, 1985) and (Yan et al., 2016). The main idea of the regression spline is to approximate $U(t_{ik})$ by using a linear combination of spline basis functions. let $U(t_{ik}) = V_{kl}(t)$ then $V_{lk}(t) = \left(V_{0k}(t), V_{1k}(t), \ldots, V_{(l-1)k}(t)\right)^T$ Mathematically, the specification can be given by

$$\begin{aligned} V_{kl}(t_{ik}) &= \eta_{0k}\, V_{0k}(t_{ik}) + \eta_{1k}\, V_{1k}(t_{ik}) + \cdots + \eta_{(L-1)k}\, V_{(L-1)k}(t_{ik}) \\ &= \sum_{l=0}^{L-1} \eta_{lk}\, V_{lk}(t_{ik}). \end{aligned} \tag{1.6}$$

where $\eta_{lk} = \left(\eta_{0k}, \eta_{1k}, \ldots, \eta_{(l-1)k}\right)^T$ are the unknown vectors of fixed coefficients $\nu_{kl}(t)$. Natural cubic spline bases with the percentile-based knots are used to estimate the effect of time on the longitudinal outcomes. To select the optimal degree of the regression spline and the number of knots, we consider different numbers of knots and compare their deviance information criteria (DICs), choosing the smallest one (Huang et al., 2011). By using equation 1.6, model equation 1.5 can be rewritten as:

$$Y_{ik} = \beta_k X_{ik} + \eta_{lk} V_{lk}(t_i) + b_{ik} Z_{ik} + e_{ik}. \tag{1.7}$$

Let $v_i = (X_{ik}, V_{kl}(t_{ik}))$ and $Z_{ik}$ be the fixed-effect and random-effect design matrices, respectively. Furthermore, let $\gamma_k = (\beta_k^T, \eta_{kl}^T)$ and $b_{ik}^T$ be the associated vectors of fixed-effects and random-effects parameters, respectively. Then, the model in equation 1.7 can be reformulated as.

$$Y_{ik} = \gamma_k v_i + b_{ik} Z_{ik} + e_{ik} \tag{1.8}$$

$$b_{ik} \sim SN(0, \Delta_{bk},\ \Sigma_b,\ \Delta_{bk}), \qquad e_{ik} \sim SN(0, \Delta_{ek},\ \Sigma_k,\ \Delta_{ek}).$$

### 1.2.2 Bayesian Inference for Parameter Estimation

The classical (frequentist) approach of parameter estimation is a well-established method for estimating all unknown parameters using the joint likelihood function. However, this technique can be computationally demanding and may encounter convergence issues, particularly when using the joint likelihood method with the proposed model, which employs a multivariate skew normal distribution for the random errors and random effects (Wu, 2002). The Bayesian inference approach offers a solution by reducing computational complexity and enabling the incorporation of prior knowledge for the unknown parameters (Huang et al., 2022). We present a Bayesian inference method via the Markov chain Monte Carlo (MCMC) procedure to estimate the parameters of the semiparametric multivariate mixed-effect model. A key feature of the Bayesian inference that allows writing the code in Just Another Gibbs Sampler (JAGS) (Plummer et al., 2003), and the model can be formulated in a flexible hierarchical representation. By introducing two random variable

vectors $w_i = \left(w_{i1}, w_{i2}, \ldots, w_{in_i}\right)^{\mathsf{T}}$ of dimension $n_i \times 1$ *and* $g_i = \left(g_{i1}, \ldots, g_{i2(q+1)}\right)^{\mathsf{T}}$ of dimension $2(q+1) \times 1$, we can present the semiparametric multivariate mixed-effect model equation 1.8 hierarchically as suggested by (Sahu et al., 2003).

$$\begin{aligned} Y_i \mid b_i, w_i &\sim N_{2n_i}(\gamma v_i + Z_i b_i + \delta_{ek}(\, w_i + \sqrt{2/\pi}\, I_{2n_i}\,),\ \Sigma_k I_{2n_i}), \\ b_i \mid g_i &\sim N_{2(q+1)}(\delta_{bk}(\, g_i + \sqrt{2/\pi}\, I_{2(q+1)}\,),\ \Sigma_b), \\ w_i &\sim N_{2n_i}(0,\ I_{2n_i})\, I(w_i > 0), \\ g_i &\sim N_{2(q+1)}(0,\ I_{2(q+1)})\, I(g_i > 0). \end{aligned} \tag{1.9}$$

where $Y_i \mid b_i, w_i \sim N_{2n_i}(\gamma v_i + Z_i b_i + \delta_{ek}(\, w_i + \sqrt{2/\pi}\, I_{2n_i}\,), \Sigma_k I_{2n_i})$ is the conditional distribution of the longitudinal responses given the random effects ($b_i$) and latent variable ($w_i$), which follows a multivariate skew normal distribution with mean equal to the sum of the fixed effects ($\gamma v_i$), the random-effects ($Z_i b_i$), and skewness term ($\delta_{ek}(\, w_i + \sqrt{2/\pi})$) with covariance matrix equal to $\Sigma_k$. $v_i$ and $Z_i$ be the covariates design matrices associated with the vector of fixed effects $\beta$ and the vector of random effects $b_i$, respectively. $i = 1, ..., n$, and $n$ is the total number of subjects, $n_i$ is the number of repeated measurements for subject $i$, and $q$ is the number of random slopes and intercepts for each longitudinal outcome. $b_i \mid g_i \sim N_{2(q+1)}(\delta_{bk}(\, g_i + \sqrt{2/\pi}), \Sigma_b)$ is the conditional distribution of the random effect given another latent variable $g_i$, which follows a multivariate skew normal distribution with mean equal to $\delta_{bk}(\, g_i + \sqrt{2/\pi})$ with covariance $\Sigma_b$. $w_i$ and $g_i$ are auxiliary latent variables, each assumed to follow a multivariate normal distribution truncated to non-negative values. These truncated normal variables generate the asymmetric components of both the longitudinal response and random-effects distributions, thereby forming the overall multivariate skew-normal structure. $\delta_{ek}$ and $\delta_{bk}$ are the skewness parameters for random errors and random effects, respectively, and $I_{2n_i}$ and $I_{2(q+1)}$ are identity matrices.

To carry out Bayesian inference, it is necessary to understand and define prior beliefs or distributions over the parameters and observed data to update them. It works by starting with a prior distribution representing initial beliefs about the parameters, then updating it to a posterior distribution after observing data using the likelihood function. Thus,

we need to define prior distributions for all unknown parameters in equation 1.8. The fixed-effect and skewness parameters are assumed to be multivariate normally distributed. Additionally, the variance-covariance matrix for the random effects and random errors is assumed to follow an inverse Wishart distribution, as shown below.

$$\begin{aligned} \gamma &\sim N(\gamma_0,\, \Omega_\Gamma), & \Delta_{ek} &\sim N(\Delta_0,\, \Omega_{\Delta_{ek}}), \\ \Delta_{bk} &\sim N(\Delta_0,\, \Omega_{\Delta_{bk}}), & \Sigma_{ek} &\sim IW(\Omega_1,\, \omega_1), \qquad \Sigma_{bk} \sim IW(\Omega_2,\, \omega_2). \end{aligned} \tag{1.10}$$

Where $\gamma$ is a vector of fixed-effect coefficients, $\gamma_0$ is a prior mean vector for $\gamma$, $\Omega_\Gamma$ is the prior covariance matrix for $\gamma$. $\Delta_0$, $\Omega_{\Delta_{bk}}$, and $\Omega_{\Delta_{ek}}$ represent the prior mean vector and prior covariance matrices associated with the skewness parameters of the random effects ($\Delta_{bk}$) and those of the random errors ($\Delta_{ek}$), respectively. $\Omega_1, \Omega_2$ are the scale or dispersion matrices for the variance covariance of the random errors ($\Sigma_{ek}$) and the variance covariance of the random effects ($\Sigma_{bk}$), respectively, with $\omega_1, \omega_2$ are the respective degrees of freedom.

For convenient implementation, the hyperparameter matrices $\Omega_\Gamma$, $\Omega_{\Delta_{ek}}$, $\Omega_{\Delta_{bk}}$, $\Omega_1$, $\Omega_2$ are assumed to be diagonal. This assumption indicates that these matrices represent independent or uncorrelated components. For instance, if we consider the parameter vector $\gamma$ and the hyperparameter matrix $\Omega_\Gamma$ is diagonal, this suggests that the elements or effect parameters associated with $\gamma$ are independent or uncorrelated. let $\Omega = \left(\gamma,\, \Delta_{ek},\, \Delta_{bk},\, \Sigma_{bk},\, \Sigma_{ek}\right)$ be the collection of unknown population parameters in model equation 1.8. We assume that they are independent of each other in other words $\pi(\Omega) = \pi(\gamma)\,\pi(\Delta_{ek})\,\pi(\Delta_{bk})\,\pi(\Sigma_{ek})\,\pi(\Sigma_{bk})$. let $D_n = \{Y_i, v_i\}$ be the observed data, $f(\cdot)$ be a density function, $f(\cdot/\cdot)$ be a conditional density function and $\pi(\cdot)$ be a prior density function, then the likelihood function is denoted by $L(\Omega \mid D_n) = f(D_n \mid \Omega)$ and its

density function is given by:

$$
\begin{aligned}
f(D_n \mid \Omega) = \prod_{i=1}^{n} \int & f_{2n_i}(Y_i;\, \gamma v_i + Z_i b_i + \delta_{ek}(w_i + \sqrt{2/\pi} \times I_{2n_i}), \Sigma_k \times I_{2n_i}) \\
& \times f_{2(q+1)}(b_i;\, \delta_{bk}(g_i + \sqrt{2/\pi} \times I_{2(q+1)}), \Sigma_b) \\
& \times f_{2n_i}(w_i; 0, I_{2n_i})\, I(w_i > 0) \\
& \times f_{2(q+1)}(g_i; 0, I_{2(q+1)})\, I(g_i > 0)\, db_i
\end{aligned} \tag{1.11}
$$

This means that the likelihood function is the product of each subject's density for the longitudinal outcomes $Y_i$, the random effect $b_i$, and the two latent variables $w_i$ and $g_i$, with integration over ior distributions of the unknown model parameters and the observed data, we can draw samples for statistical inference from the posterior density by combining the likelihood function 1.11 and the priors. The joint posterior density of the unknown parameter is the product of the likelihood and the prior distributions. That is, the joint posterior density of $\Omega$ given the observed data, $D_n$, can be given by
$f(\Omega \mid D_n) \propto f(D_n \mid \Omega) \times \pi(\Omega) =$

$$
\begin{aligned}
f(\Omega \mid D_n) \propto \Big\{ \prod_{i=1}^{n} \int & f_{2n_i}(Y_i;\, \gamma v_i + Z_i b_i + \delta_{ek}(w_i + \sqrt{2/\pi} \times I_{2n_i}), \Sigma_k \times I_{2n_i}) \\
& \times f_{2(q+1)}(b_i;\, \delta_{bk}(g_i + \sqrt{2/\pi} \times I_{2(q+1)}), \Sigma_b) \\
& \times f_{2n_i}(w_i; 0, I_{2n_i})\, I(w_i > 0) \\
& \times f_{2(q+1)}(g_i; 0, I_{2(q+1)})\, I(g_i > 0) db_i \Big\} \times \pi(\Omega)
\end{aligned} \tag{1.12}
$$

The posterior distribution of the equation 1.12 does not have a closed-form solution. As a result, calculating the posterior distribution directly from the observed data is challenging. MCMC procedures have been employed to sample from the posterior distribution using Gibbs sampling.

### 1.2.3 Application to Diabetes and Hypertension Data

## Motivating Glucose Concentration and Blood Pressure Data

We illustrate the proposed method by analysing electronic medical record data from a primary care patient cohort to investigate the trajectory patterns of glucose concentration and blood pressure, along with their associations with demographic factors. The data were collected from patients with diabetes and hypertension at Felegehiwot Comprehensive Specialized Hospital in Ethiopia. This paper includes individuals with diabetes and hypertension who had two or more visits during the 5-year study period, spanning January 2018 to December 2022. After applying the inclusion and exclusion criteria, we included 220 subjects in the study. A total of 1,837 longitudinal measurements were collected from these participants.

Demographic variables (age, sex, and residence) were extracted from their medical chart. The fasting blood glucose in milligrams per deciliter (mg/dL) and systolic blood pressure in millimeters of mercury (mmHg) were measured every 6 months over 5 years and recorded in medical charts. Of the 220 people with type 2 diabetes and hypertension, 132 (60.00%) are male; the mean of the baseline fasting blood glucose and systolic blood pressure are 200.90 (62.27) mg/dL and 142 (14.21) mmHg, respectively, as shown in Table 1.1. The mean age of people with T2D and hypertension was 53.75, with a standard deviation of 11.09 years.

The upper panel of Figure 1.2 shows the histograms of FBS and SBP, respectively. As shown in the graph, the distributions of the longitudinal outcome variables, fasting blood sugar and systolic blood pressure, deviate from normal. Although both outcomes are positively skewed, FBS skewness is much greater than that of SBP. We further examined the normality of the FBS and SBP outcomes using Shapiro–Wilk's test, which revealed that the normality assumption was violated. The bottom panel of Figure 1.2 shows the FBS and SBP trajectories of five randomly selected patients. The randomly selected sample trajectories of FBS and SBP clearly demonstrated fluctuations over time.

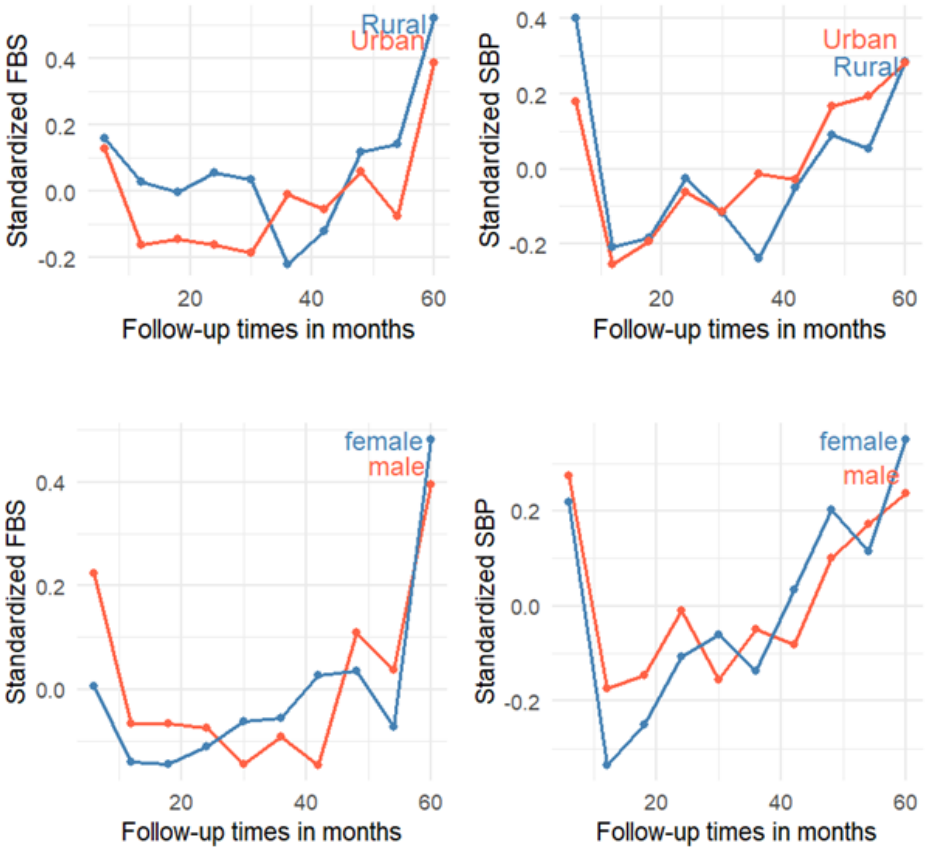



Figure 1.1: The mean trajectory plots of FBS and SBP by place of residence and sex

We fit individual profiles using a univariate linear mixed-effects model for each outcome, assuming that the residual terms and random effects are normally distributed. Figure 1.3 suggests that although estimated subject-specific slopes appear normally distributed, variation in intercepts may not be normally distributed. Figure 1.1 explores the mean trajectory plot of glucose concentration and blood pressure with categorical variables over time. The mean trajectory of glucose concentration in people with T2D and hypertension living in rural areas was higher than in those living in urban areas. In contrast, blood pressure remained higher until the midpoint of the study period (36 months). However, after 36 months, the trajectory reversed. The mean trajectories of glucose concentration and blood pressure differ by sex; females appear to have lower mean trajectories than males, though the trends reverse at times.

Table 1.1: Descriptive statistics for variables at baseline: frequency (proportion) for categorical variables and mean (SD) for quantitative variables

| **Variables** | **Category** | **Proportion / Mean** |
|---|---|---|
| Sex | Male | 132 (60.00%) |
| | Female | 88 (40.00%) |
| Residence | Urban | 143 (65.00%) |
| | Rural | 77 (35.00%) |
| Baseline age | | 53.75 (11.09) |
| Measurement time | | 33 |
| Baseline FBS | | 200.90 (62.27) |
| Baseline SBP | | 142.15 (14.21) |

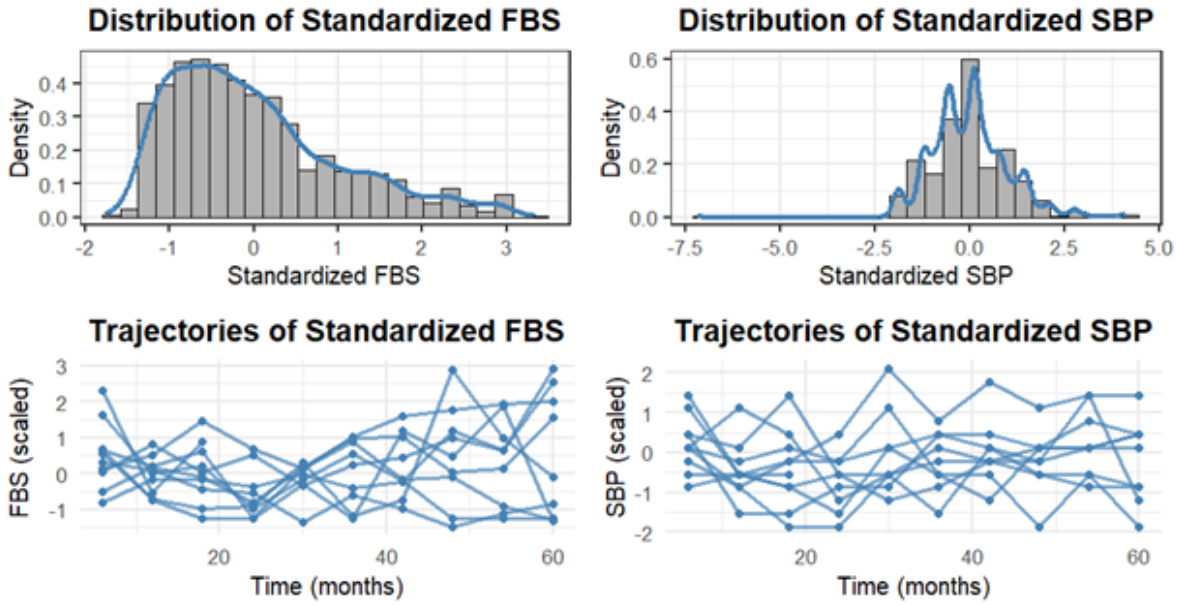



Figure 1.2: The histogram of FBS and SBP (standardized scale) (upper panel) and the trajectory profiles of FBS and SBP (standardized scale) for randomly selected subjects (lower panel) in people with T2D and hypertension.

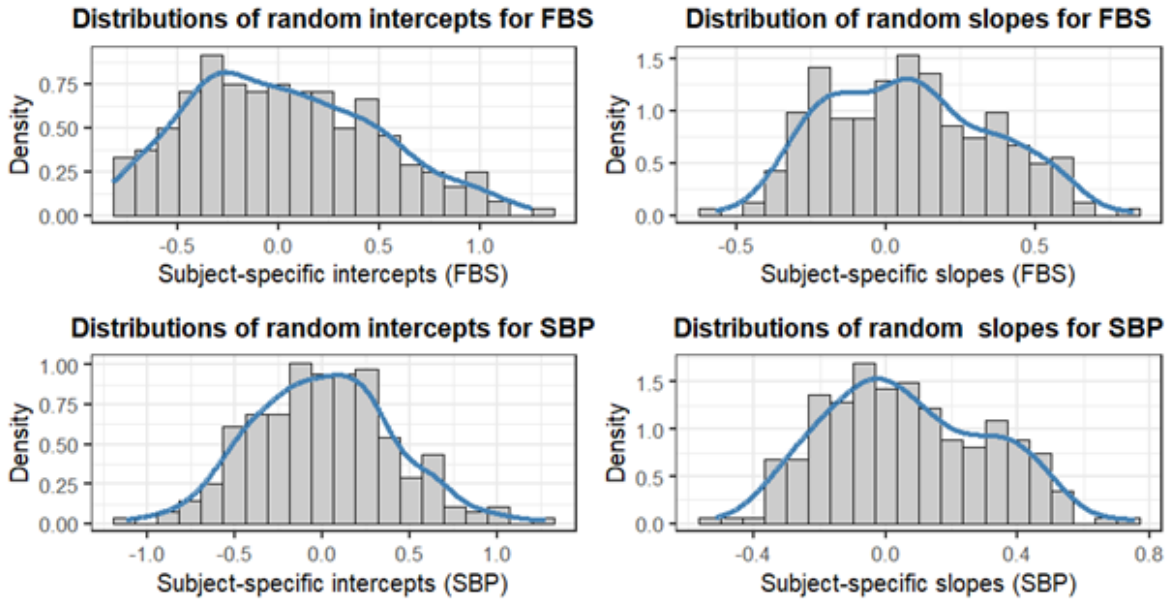



Figure 1.3: Distribution of subject-specific intercepts and slope

### 1.2.4 Model Implementation

We illustrate the proposed methods using the clinical data on diabetes and hypertension described in Section 1.2.3. Sex, age, and residence were included as covariates, and three time-spline basis functions with four random effects were used to model the longitudinal trajectories of glucose concentration and systolic blood pressure. In this model, we assume that the mean trajectories of FBS and SBP differ by sex and residence. We conducted the following scenarios. First, we compare the semiparametric multivariate mixed-effects model with its fully parametric counterpart, which assumes multivariate normal distributions for both the random effects and model errors, to assess how nonlinear time effects influence model results. Second, we investigated the performance of the proposed models under different distributional assumptions of the random effects and random errors. We consider the following fully parametric multivariate mixed-effects model to model the trajectory of glucose concentration and blood pressure:

$$Y_{ijk} = (\beta_{1k} + b_{ik1}) + (\beta_{2k} + b_{ik2})\,\text{time}_{ijk} + \beta_{3k}\,\text{age}_i + \beta_{4k}\,\text{sex}_i + \beta_{5k}\,\text{residence}_i + e_{ijk} \quad (1.13)$$

Where $k = 1$ and 2, denote the repeated measurements of the glucose concentration in mg/dL and systolic blood pressure in mmHg, respectively, for the $i^{\text{th}}$ subject measured at time $t_{ij}$ (expressed in months, $i = 1, 2, \ldots, 220$, $j = 0, 1, 2, \ldots, 10$). $\text{age}_i$ be the age of the $i^{\text{th}}$ subject at baseline. The vector of fixed-effect parameters is $\beta^T = (\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}, \beta_{51}, \beta_{12}, \beta_{22}, \beta_{32}, \beta_{42}, \beta_{52})^T$, representing the intercept and the effects of time, age, sex, and residence on glucose concentration and blood pressure, respectively. The random-effect vector is $b_i = (b_{i11}, b_{i21}, b_{i12}, b_{i22})$, corresponding to the random intercepts and random slopes for the longitudinal outcomes (glucose concentration and blood pressure). $e_{ijk}$ are the random errors with dimension $n_i \times 1$, representing the within-subject residuals for glucose concentration and blood pressure, respectively.

In addition to the fully parametric fixed-effects model, we present a semiparametric (partial linear) multivariate mixed-effects model to analyse the trajectories of glucose concentration and blood pressure.

$$\begin{aligned} Y_{ijk} \;=\; & (\beta_{1k} + b_{ik1}) \;+\; b_{ik2}\,\text{Time}_{ijk} \;+\; \eta_{k0}\,V_{0k}(\text{Time}_{ijk}) \;+\; \eta_{k1}\,V_{1k}(\text{Time}_{ijk}) \\ & +\; \eta_{k2}\,V_{2k}(\text{Time}_{ijk}) \;+\; \beta_{2k}\,\text{age} \;+\; \beta_{3k}\,\text{sex} \;+\; \beta_{4k}\,\text{residance} \;+\; e_{ijk}. \end{aligned} \tag{1.14}$$

$V_{kl}(t) = \big(V_{k0}(\text{Time}_{ijk}),\ V_{k1}(\text{Time}_{ijk}),\ V_{k2}(\text{Time}_{ijk})\big)^T$ is a vector of natural cubic spline bases used in the regression spline method and $\eta_k^T = (\eta_{k0}, \eta_{k1}, \eta_{k2})^T$ are the corresponding fixed effect parameter. To approximate the spline bases, we chose two internal knots located at arbitrary months between 0 and 54, as well as two boundary knots positioned at 0 and 54 months. The locations of the two internal knots were determined from the distribution of observed measurement time points, specifically at the 25th and 75th quantiles. We further explore how skewness in the distributions of model errors and random effects influences the precision of parameter estimates by comparing the following three models.

**Model N:** Model with multivariate normal distribution for both random errors and random effects.

**Model SNE:** Model with multivariate skew-normal distribution for the random errors

and multivariate normal distribution for the random effects.

**Model SNR:** Model with multivariate skew-normal distribution for the random effects and multivariate normal distribution for the random errors.

To perform Bayesian inference for the proposed models, we need to specify the hyper-parameter values for the prior distributions. In the absence of historical data or experiments, we specify weakly informative priors for all model parameters to obtain a proper posterior distribution. Specifically, each component of the fixed effects vectors $\beta$, $\eta$ as well as the skewness parameter $\Delta_{bk}$ and $\Delta_{ek}$ were assumed to follow independent normal distributions with mean zero and variance 100, i.e. $\mathcal{N}(0, 100)$. The prior for the variance–covariance matrices of the random effects, $\Sigma_b$, and random errors, $\Sigma_k$, is assumed to follow an inverse Wishart distribution with $\Omega_{\Delta_{bk}} = \text{diag}(1, 1, 1, 1)$, degrees of freedom $= 5$, and $\Omega_{\Delta_{ek}} = \text{diag}(0.01, 0.01)$, degrees of freedom $= 3$, respectively.

For each model, we ran three MCMC sampling chains, each consisting of 30,000 iterations, with initial values dispersed. We discarded the first 7,000 iterations as a burn-in period and retained every 20th sample thereafter. Thus, we obtain 3450 samples from the targeted posterior distributions and use them to make statistical inferences. We assessed convergence of the generated samples, indicating that the algorithm has reached its equilibrium target distribution, using standard JAGS diagnostics, such as trace plots, density plots, and autocorrelation plots. Figures 1.4 display the trace plot, density plot, and autocorrelation plot of the some parameter used in the illustration. Note that only the plots for some parameters in semiparametric mixed-effect models are shown here; the remaining parameters for this model can be found in the appendix. The trace plots (upper panel) indicate that the lines from the three distinct chains intersect, demonstrating that the algorithm has achieved convergence. It implies that the regression coefficients for the selected parameters converge to their target distributions. Additionally, the posterior density plots in the middle panel provide valuable insights. The density plots from each chain closely overlap and resemble the density estimates from the other chains, indicating that all chains are sampling from the exact posterior distribution. We further

monitor convergence using autocorrelation plots (bottom panel), which show a drop in autocorrelation after lag 3 for all parameters, and for others, it trails off quickly. The autocorrelations decay rapidly to zero within just a few lags, and all chains display consistent behaviour. This indicates independence between successive samples, efficient exploration of the posterior distribution, and substantial evidence that the chains have converged to the target distribution.

In addition to visual inspection, it is essential to employ formal statistical tests to assess the convergence of the samples generated from the MCMC procedure. In this study, we employed the Gelman-Rubin statistic (Gelman and Rubin, 1992) to evaluate whether multiple chains have converged to the same target distribution, a crucial step in ensuring reliable inference. The results are not presented here, but all parameters have R-hat values below 1.005, indicating that their distributions are converging toward the target distribution.

## 1.3 Results and Discussion

### 1.3.1 Comparison of Model Fitting Results

We conducted the following scenarios. First, a fully parametric multivariate mixed-effects model is compared with a partially parametric mixed-effects model that assumes normal distributions for both random effects and model errors to evaluate how the non-linearity of the model parameter influences the modeling results. Second, we examined how the asymmetric distributions of random errors and random effects (Models SNE and SNR) fit the longitudinal data and how their parameter estimates compare with those from the standard multivariate normal distribution (Model N). To improve the efficiency of the estimation algorithm and reduce variability due to measurement errors, we standardized longitudinal data on fasting blood sugar and systolic blood pressure. Additionally, to avoid unstable estimates, the covariates age and measurement time were standardized. Therefore, the results are expressed in standard units rather than the original units of

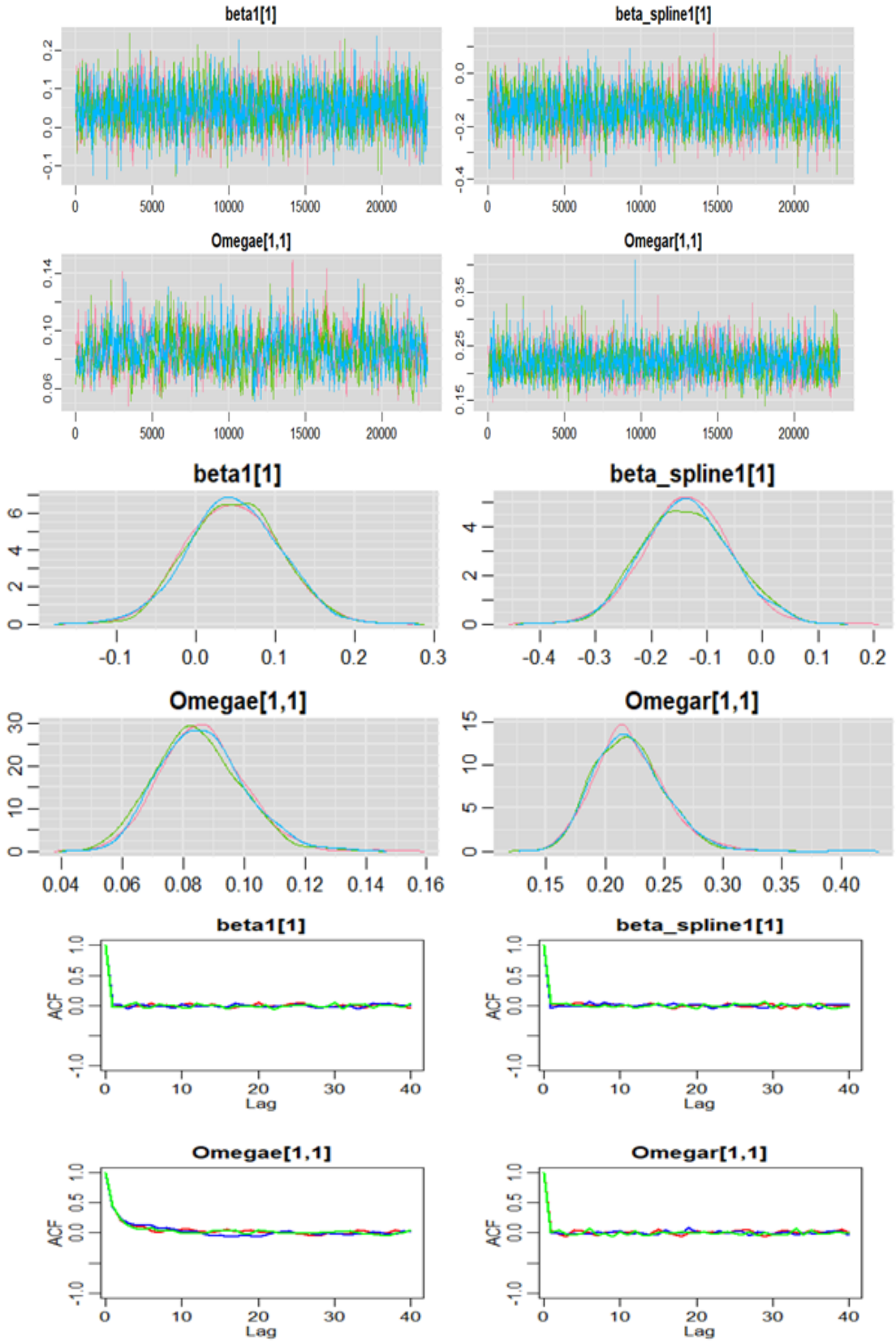



Figure 1.4: Trace plot (upper panel), density plot (middle panel), and autocorrelation function plot (lower panel) for some parameters

measurement.

To select the best-fitting model, the deviance information criterion (DIC) is used, along with the residual sum of squares (RSS) and expected predictive deviance. Like other model selection criteria, we caution that DIC, RSS, and expected predictive deviance are not intended to identify the "correct" model, but rather to compare a collection of alternative model formulations.

We begin by initially comparing the performance of two models: the proposed semiparametric (partial linear) multivariate mixed-effects model and a fully parametric multivariate mixed-effects model that assumes normal distributions for both the random effects and model errors. The posterior mean, standard deviation, and 95% credible interval of the two models are shown in Table 1.2. The results show that the posterior mean of all fixed-effect parameters in a fully parametric multivariate mixed-effects model is slightly lower than in a semiparametric multivariate mixed-effects model. For instance, the estimates from a fully parametric multivariate mixed-effects model, with values $\beta_{11}$ = 0.002 and $\beta_{12}$ = 0.022, increase to 0.067 and 0.163 in a semiparametric multivariate mixed-effects model. In contrast, the estimated within-subject variances and covariances in the fully parametric model are larger than in the semiparametric model.

The DIC, RSS, and expected predictive deviance values for the fully parametric multivariate mixed effects model are greater than those for the semiparametric multivariate mixed effects model, indicating that the semiparametric model fits the data better than the fully parametric model (see Table 1.3 ). This suggests that accounting for the nonlinearity of the time effect, particularly when the data exhibit a nonlinear relationship, as shown in the lower panels of Figure 1.2, improves model fitting and parameter estimation by providing the flexibility of the multivariate mixed-effects model. After selecting the semiparametric multivariate mixed effects model as the most suitable model that fits the data perfectly, we further investigate how the asymmetric distribution of random effects and model errors contributes to the modelling results by fitting models N, SNE, and SNR.

Table 1.2: Comparison of posterior mean and standard deviation (SD) between the fully parametric multivariate mixed effect model and the semiparametric multivariate mixed effect model

| **Parameter** | **Fully parametric** | | | **Semiparametric** | | |
|---|---|---|---|---|---|---|
| | Estimate | SD | 95% CI | Estimate | SD | 95% CI |
| $\beta_{11}$ | 0.002 | 0.060 | 0.001, 0.116 | 0.067 | 0.071 | 0.019, 0.207 |
| $\beta_{31}$ | 0.047 | 0.037 | 0.022, 0.120 | 0.051 | 0.036 | 0.026, 0.126 |
| $\beta_{41}$ | -0.069 | 0.075 | -0.119, 0.077 | -0.068 | 0.075 | -0.119, 0.080 |
| $\beta_{51}$ | 0.141 | 0.078 | 0.088, 0.294 | 0.137 | 0.075 | 0.085, 0.282 |
| $\beta_{12}$ | 0.022 | 0.057 | -0.018, 0.138 | 0.163 | 0.073 | 0.114, 0.302 |
| $\beta_{32}$ | 0.131 | 0.037 | -0.018, 0.204 | 0.135 | 0.037 | 0.108, 0.206 |
| $\beta_{42}$ | -0.065 | 0.076 | -0.116, 0.081 | -0.063 | 0.074 | -0.113, 0.080 |
| $\beta_{52}$ | 0.043 | 0.079 | -0.010, 0.201 | 0.043 | 0.076 | -0.007, 0.193 |
| $\sigma^2_{11}$ | 0.601 | 0.023 | 0.585, 0.647 | 0.579 | 0.022 | 0.564, 0.175 |
| $\sigma^2_{12}$ | 0.206 | 0.017 | 0.193, 0.240 | 0.187 | 0.017 | 0.175, 0.222 |
| $\sigma^2_{22}$ | 0.651 | 0.024 | 0.634, 0.700 | 0.631 | 0.024 | 0.615, 0.678 |

The DIC, EPD, and expected predictive deviance in Model SNE are smaller than those for Model N and Model SNR, as shown in Table 1.3. This suggests that Model SNE provides a better fit to the data than Model N and Model SNR. Furthermore, Figure 1.6 shows the diagnostic plots of observed values vs. fitted values of FBS and SBP based on Model N, Model SNE, and Model SNR to assess the goodness-of-fit of the proposed models. As shown in the figure, Model SNE provides a much better fit to the observed values of FBS and SBP compared to Model SNR and Model N. Thus, according to the DIC, RSS, expected predictive deviance, and the diagnostic plots, we conclude that the semiparametric multivariate mixed effect model, which assumes a skewed multivariate normal distribution for the model errors, is the best fit model for the observed skewed FBS and SBP data. Thus, we further report the detailed findings from Model SNE.

Bayesian inference across the three models shows that almost all population parameters are overestimated in models N and SNR relative to model SNE. The standard deviation of the fixed-effects parameters in model SNE is much smaller than those in the other two models (see Table 1.4). The findings suggest that failing to account for skewness in the mixed-effects model may lead to substantial underestimation or overestimation of the parameter estimates. This indicates that the semiparametric multivariate mixed-effects

model, which uses a multivariate skew-normal distribution for model errors, yields results that are relatively unbiased compared to the model that assumes a multivariate normal distribution.

Future more, the magnitude of the estimated within-subject variances $\sigma^2_{11}$, $\sigma^2_{22}$ and covariance $\sigma^2_{12}$, (see Table 1.4) and between-subject variance-covariance values ($v_{11}$ to $v_{34}$) (see Table 1.5) for model SNE are smaller than those for models N. The difference is expected because high variability and skewness are related to the model's assumption of a skewed distribution of model errors and random effects.

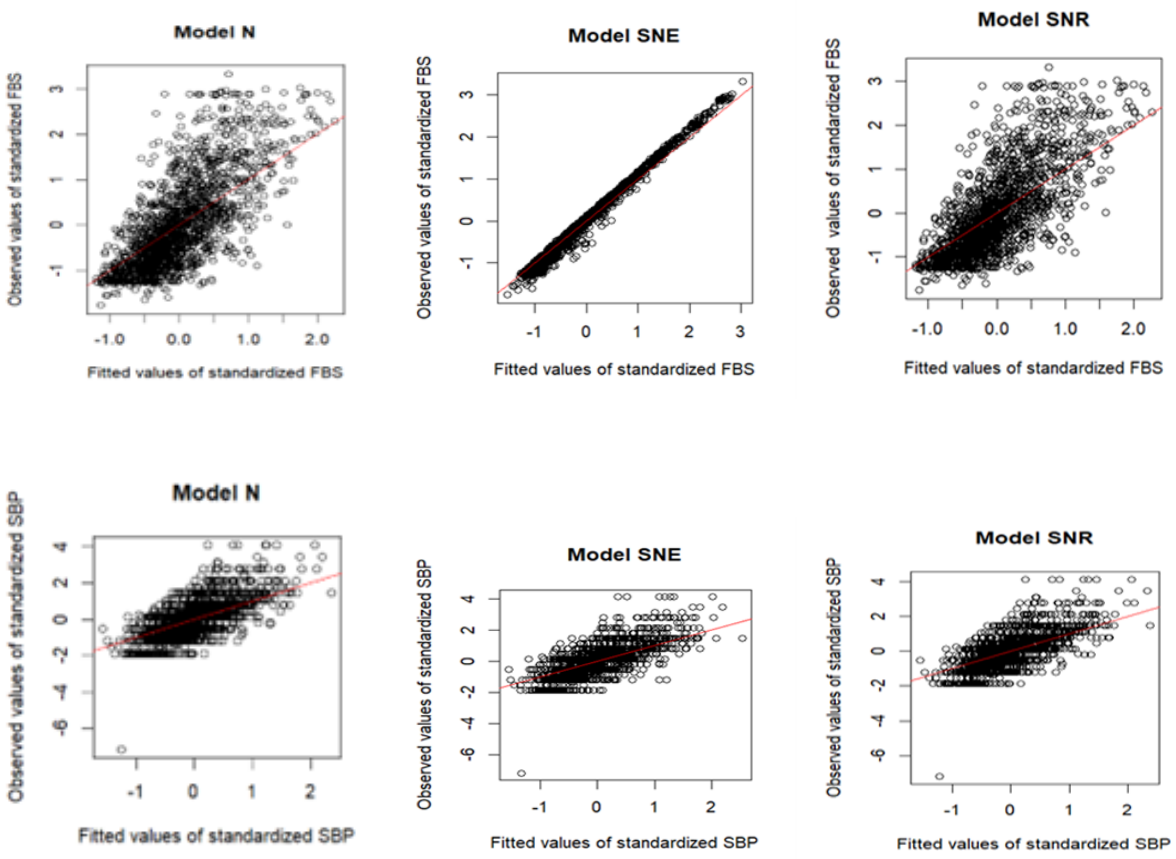



Figure 1.6: The observed values versus fitted values of FBS and SBP based on model N, model SNE, and model SNR.

Table 1.3: Model comparison using expected predictive deviance, RSS, and DIC criteria

| criteria | Semiparametric | Fully parametric | Model N | Model SNE | Model SNR |
|---|---|---|---|---|---|
| Deviance | 8393.025 | 8495.05 | 8393.025 | 4844.393 | 8375.145 |
| RSS | 1897.858 | 1971.358 | 1897.858 | 939.127 | 1883.926 |
| DIC | 9422.5 | 9484.8 | 9422.5 | 4667.2 | 9377.9 |

**Results of analysis based on Model SNE**

The posterior mean (PM), the corresponding standard deviation (SD), and the 95% credible interval (LCL and UCL) for the fixed-effect parameters and skewness across the three models are presented in Table 1.4. The estimated skewness parameters for the random errors and random effects explain the asymmetry in the longitudinal outcomes. The posterior mean estimates of the skewness parameters accounted for in the random errors $\delta_1$ and $\delta_2$ are 0.99 and 0.32, with their corresponding 95% confidence intervals (0.98, 1.00) and (0.28, 0.43), respectively. This indicates positive Skewness, providing evidence of right skewness in our bivariate, correlated longitudinal data. In addition, the skewness of fasting blood glucose is greater than that of systolic blood pressure, which implies that the glucose concentration data is more skewed than the systolic blood pressure data. Thus, incorporating a skewness parameter in the model is necessary to provide unbiased parameter estimates.

The progression of fasting blood glucose levels in people with type 2 diabetes and hypertension living in rural areas was faster than in those living in urban areas. For instance, the residence coefficient ($\beta_{41} = 0.05$, 95% CI (0.01, 0.17)) indicates that the standardized mean FBS value for people with diabetes and hypertension in rural areas is 0.05 mg/dl higher than for those in urban areas, when the effects of other covariates remain constant.

The variances associated with both the random intercept and random slope were significant, underscoring substantial inter-individual variability in baseline disease severity

Table 1.4: Summary of estimated posterior mean (PM), standard deviation (SD), and 95% credible intervals for fixed effects, skewness, based on Models N, SNE, and SNR.

| Response | Model | Stat | $\beta_{11/12}$ | $\beta_{21/22}$ | $\beta_{31/32}$ | $\beta_{41/42}$ | $\eta_{01/02}$ | $\eta_{11/12}$ | $\eta_{21/22}$ | $\delta_{1/2}$ | $\delta_{11/12}$ | $\delta_{21/22}$ | $\sigma^2_{12}$ | $\sigma^2_{11/22}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **FBS** | Model N | PM | 0.06 | 0.05 | -0.06 | 0.13 | -0.18 | -0.02 | 0.48 | – | – | – | 0.18 | 0.57 |
| | | SD | 0.07 | 0.03 | 0.07 | 0.07 | 0.09 | 0.15 | 0.09 | – | – | – | 0.01 | 0.02 |
| | | LCL | 0.01 | 0.02 | -0.11 | 0.08 | -0.24 | -0.12 | 0.41 | – | – | – | 0.17 | 0.56 |
| | | UCL | 0.20 | 0.12 | 0.08 | 0.28 | 0.05 | 0.28 | 0.67 | – | – | – | 0.22 | 0.62 |
| | Model SNE | PM | 0.04 | 0.04 | -0.01 | 0.05 | -0.13 | -0.08 | 0.43 | 0.99 | – | – | 0.03 | 0.08 |
| | | SD | 0.05 | 0.02 | 0.06 | 0.06 | 0.07 | 0.13 | 0.08 | 0.01 | – | – | 0.02 | 0.01 |
| | | LCL | 0.01 | 0.02 | -0.05 | 0.01 | -0.18 | -0.16 | 0.37 | 0.98 | – | – | 0.01 | 0.07 |
| | | UCL | 0.15 | 0.10 | 0.10 | 0.17 | 0.01 | 0.16 | 0.60 | 1.00 | – | – | 0.07 | 0.11 |
| | Model SNR | PM | 0.07 | 0.05 | -0.06 | 0.13 | -0.17 | -0.02 | 0.48 | – | 0.28 | -0.01 | 0.18 | 0.57 |
| | | SD | 0.07 | 0.03 | 0.07 | 0.07 | 0.09 | 0.15 | 0.10 | – | 0.24 | 0.14 | 0.01 | 0.02 |
| | | LCL | 0.02 | 0.02 | -0.11 | 0.08 | -0.23 | -0.12 | 0.41 | – | 0.10 | -0.11 | 0.17 | 0.56 |
| | | UCL | 0.21 | 0.12 | 0.07 | 0.28 | 0.01 | 0.28 | 0.68 | – | 0.62 | 0.29 | 0.22 | 0.62 |
| **SBP** | Model N | PM | 0.16 | 0.13 | -0.06 | 0.04 | 0.04 | -0.33 | 0.44 | – | – | – | – | 0.63 |
| | | SD | 0.07 | 0.03 | 0.07 | 0.07 | 0.09 | 0.15 | 0.09 | – | – | – | – | 0.02 |
| | | LCL | 0.11 | 0.11 | -0.11 | -0.01 | -0.02 | -0.43 | 0.37 | – | – | – | – | 0.58 |
| | | UCL | 0.30 | 0.20 | 0.08 | 0.18 | 0.23 | -0.02 | 0.62 | – | – | – | – | 0.67 |
| | Model SNE | PM | 0.15 | 0.13 | -0.04 | 0.01 | -0.05 | -0.36 | 0.41 | 0.32 | – | – | – | 0.58 |
| | | SD | 0.07 | 0.03 | 0.07 | 0.07 | 0.09 | 0.15 | 0.09 | 0.05 | – | – | – | 0.02 |
| | | LCL | 0.10 | 0.10 | -0.09 | -0.03 | -0.01 | -0.46 | 0.35 | 0.28 | – | – | – | 0.56 |
| | | UCL | 0.29 | 0.20 | 0.10 | 0.16 | 0.23 | -0.05 | 0.59 | 0.43 | – | – | – | 0.63 |
| | Model SNR | PM | 0.16 | 0.12 | -0.05 | 0.03 | 0.03 | -0.34 | 0.43 | – | 0.36 | -0.04 | – | 0.62 |
| | | SD | 0.07 | 0.03 | 0.07 | 0.08 | 0.09 | 0.15 | 0.09 | – | 0.22 | 0.13 | – | 0.02 |
| | | LCL | 0.11 | 0.10 | -0.10 | -0.01 | -0.02 | -0.44 | 0.37 | – | 0.30 | -0.13 | – | 0.61 |
| | | UCL | 0.31 | 0.19 | 0.08 | 0.18 | -0.02 | -0.44 | 0.37 | – | 0.63 | 0.19 | – | 0.67 |

and progression among people with type 2 diabetes and hypertension, as shown in Table 1.5. This result confirms the need to incorporate random effects into the model to account for the correlation among repeated measurements within subjects appropriately. Additionally, a positive random intercept for a given subject suggests that the $i^{\text{th}}$ subject has higher baseline glucose and blood pressure levels than the population average. In contrast, a positive random slope implies that the $i^{\text{th}}$ subject experiences a more rapid increase in glucose concentration and blood pressure over time relative to the average trajectory.

The estimated overall posterior covariance between glucose concentration and blood pressure progression is significantly positive, despite varying degrees of association. Moreover, the random intercept and random slopes are positively associated with both glucose concentration and blood pressure; specifically, the covariance values of $v_{12} = 0.10$ and $v_{34} = 0.06$, in model SNE expressed in standardized units, indicate the covariance between the random intercept and slope for glucose concentration and blood pressure, respectively. This shows that a higher glucose concentration at the time of diabetes diagnosis may lead

to faster progression of glucose concentration. In the same way, people whose blood pressure was higher at the first measurement show a quicker rise in blood pressure. This suggests that patients with worse disease severity at the beginning tend to have a faster disease progression rate and vice versa.

Table 1.5: A summary of the estimated posterior mean of the variance covariance matrix of the random effects, the corresponding standard deviation (SD), and 95% credible interval for model N, model SNE, and model SNR.

| **Model** | **Stat** | $v_{11}$ | $v_{22}$ | $v_{33}$ | $v_{44}$ | $v_{12}$ | $v_{13}$ | $v_{14}$ | $v_{23}$ | $v_{24}$ | $v_{34}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| N | PM | 0.29 | 0.15 | 0.23 | 0.13 | 0.11 | 0.12 | 0.11 | 0.06 | 0.08 | 0.07 |
| | SD | 0.03 | 0.02 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 |
| | LCL | 0.27 | 0.13 | 0.21 | 0.11 | 0.10 | 0.10 | 0.09 | 0.05 | 0.06 | 0.06 |
| | UCL | 0.37 | 0.20 | 0.30 | 0.17 | 0.16 | 0.18 | 0.15 | 0.10 | 0.12 | 0.11 |
| SNE | PM | 0.22 | 0.11 | 0.22 | 0.12 | 0.10 | 0.09 | 0.09 | 0.05 | 0.06 | 0.06 |
| | SD | 0.02 | 0.01 | 0.03 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 | 0.01 |
| | LCL | 0.19 | 0.10 | 0.20 | 0.10 | 0.08 | 0.07 | 0.08 | 0.04 | 0.05 | 0.05 |
| | UCL | 0.26 | 0.14 | 0.29 | 0.16 | 0.10 | 0.13 | 0.13 | 0.07 | 0.08 | 0.10 |
| SNR | PM | 0.24 | 0.15 | 0.17 | 0.12 | 0.11 | 0.11 | 0.10 | 0.06 | 0.08 | 0.07 |
| | SD | 0.05 | 0.02 | 0.04 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 |
| | LCL | 0.20 | 0.13 | 0.14 | 0.11 | 0.09 | 0.10 | 0.09 | 0.05 | 0.06 | 0.06 |
| | UCL | 0.35 | 0.20 | 0.26 | 0.17 | 0.16 | 0.18 | 0.15 | 0.10 | 0.12 | 0.11 |

### 1.3.2 Simulation Study

Simulation studies were conducted to evaluate the performance of the proposed semi-parametric multivariate mixed-effects model under different distributional assumptions of both random errors and random effects. We want to investigate how well the parameters are estimated across the proposed models. The simulation studies were designed as follows. The bivariate correlated longitudinal outcomes $Y_{ijk}$ were simulated using the

following semiparametric multivariate mixed effects model:

$$Y_{ijk} = (\beta_{1k} + b_{ik1}) + b_{i2k}\,\text{Time}_{ij} + \eta_{k0}\,V_{0k}(\text{Time}_{ijk}) + \eta_{1k}\,V_{1k}(\text{Time}_{ijk}) + \eta_{2k}\,V_{2k}(\text{Time}_{ijk}) + \beta_{2k}\,\text{age}_i + \beta_{3k}\,\text{sex}_i + \beta_{4k}\,\text{residence}_i + e_{ijk}, \tag{1.15}$$

Where $Y_{ij1}$ and $Y_{ij2}$ are the respective glucose concentration and blood pressure for $i^th$ subject at measured time $t_{ij}$, $\beta^T = \big(\beta_{1k},\, \beta_{2k},\, \beta_{3k},\, \beta_{4k}\big)^T$ are intercept, the effect of age, sex, and residence on glucose concentration and blood pressure. $\eta_k^T = \big(\eta_{k0},\, \eta_{k1},\, \eta_{k2}\big)^T$ are the nonlinear time effect on glucose concentration and blood pressure. $b_i = \big(b_{i11},\, b_{i21},\, b_{i12},\, b_{i22}\big)$ are the random intercept and random slope for glucose concentration and blood pressure. $V_{0k}(\text{Time}_{ijk}),\quad V_{1k}(\text{Time}_{ijk}),\quad V_{2k}(\text{Time}_{ijk})$ is a vector of natural cubic spline bases used to estimate the nonlinear effects of measurement time on glucose concentration and blood pressure through the regression spline method. To generate these spline bases, we consider 10 equally spaced visiting time points between 0 and 54 ($t_{ij} = 0, 6, 12, \ldots, 54$) where the longitudinal measurements are taken from each subject. The locations of these knots were determined based on the quantiles of the distribution of observed measurement time points. The model includes three covariates: age, sex, and residence. Age, a continuous variable, is simulated from a normal distribution with a mean of 54 and a standard deviation of 11. In contrast, sex and residence, categorical variables, are simulated from Bernoulli distributions with probabilities of 0.32 and 0.4, respectively. The random effects vector $b_i = (b_{i11}, b_{i21}, b_{i12}, b_{i22})$ is simulated from a multivariate normal distribution with mean zero and a variance–covariance matrix obtained from the real data in model SNE.

To introduce skewness into the longitudinal data, random errors were generated using a gamma distribution. Specifically, the random errors for fasting blood sugar and systolic blood pressure were generated using a gamma distribution with both the shape and scale parameters set to 1. We set the true values of model parameters as those obtained from real data analysis for Model SNE, which are present in Table 1.4. The parameter vectors are defined as follows: $\beta_1^{\mathsf{T}} = (\beta_{11},\, \beta_{21},\, \beta_{31},\, \beta_{41})^{\mathsf{T}} = (0.048,\, 0.043,\, -0.013,\, 0.057)^{\mathsf{T}}, \beta_2^{\mathsf{T}} =$

$(\beta_{12},\ \beta_{22},\ \beta_{32},\ \beta_{42})^{\mathsf{T}} = (0.155,\ 0.132,\ -0.042,\ 0.016)^{\mathsf{T}}, \eta_1^{\mathsf{T}} = (\eta_{01},\ \eta_{11},\ \eta_{21})^{\mathsf{T}} = (-0.136,$
$-0.085,\ 0.435)^{\mathsf{T}}, \eta_2^{\mathsf{T}} = (\eta_{02},\ \eta_{12},\ \eta_{22})^{\mathsf{T}} = (0.052,\ -0.360,\ 0.419)^{\mathsf{T}}$.

In accordance with the study design described above, we generate a sample of 400 subjects and repeat the simulation 100 times (i.e., create 100 simulated datasets). We then fit models N, SNE, and SNR to each simulated data set. The Bayesian framework for statistical inference, the specification of prior distributions, and the tools for convergence diagnostics were similar to those in the model implementation section 1.2.4. We used 7000 iterations after discarding the initial 2000 iterations to draw an inference. To compare the performance of the estimated parameters across the three models, we computed biases,$B(\theta) = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i - \theta_T$, Root Mean Square Error$(\theta) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\theta}_i - \theta_T\right)^2}$, and the deviance information criterion (DIC), where $\hat{\theta}_i$ is the estimate of the true parameter $\theta_T$ from the $i^{\text{th}}$ simulation and $\theta_T$ the true parameter value, which are obtained from the application dataset, and m is the total number of replications. Table 1.6 summarizes the simulation results, including the true parameter value (TP), bias, RMSE, and DIC value. Among the three models, the model SNE has the lowest DIC (10418.97), indicating the best model fit. In contrast, the multivariate skew-normal distribution model for the random effects (model SNR) has a DIC of 18228.97, and the multivariate normal distribution (model N) has a DIC of 18333.44.

Since one of the main focuses of this simulation study is to compare parameter estimation across models N, SNE, and SNR, we assessed the impact of distributional assumptions on estimates of the fixed effects parameters. The multivariate semiparametric mixed-effect model, which employs a multivariate skew-normal distribution for the model error, estimates the fixed-effect parameters with slightly smaller bias and root-mean-square errors than the other models. Thus, our simulation results suggest that although unbiased estimation is still possible under normality, failure to appropriately account for the true features of the random effects and random errors leads to biased inference. Figure 1.7 displays a bar plot illustrating the bias induced by misusing the model distribution assumption for some parameters in the mixed-effects model. Among all three models, model

SNE has the relatively smallest RMSE.

Table 1.6: Summary of true parameter (TP) values, bias, and RMSE for Models N, Model SNE, and Model SNR.

| **Parameter** | **TRV** | **Model N** | | **Model SNE** | | **Model SNR** | |
|---|---|---|---|---|---|---|---|
| | | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| $\beta_{11}$ | 0.048 | -0.272 | 0.278 | -0.279 | 0.283 | -0.277 | 0.283 |
| $\beta_{21}$ | 0.043 | 0.003 | 0.021 | 0.000 | 0.018 | 0.001 | 0.021 |
| $\beta_{31}$ | -0.013 | 0.001 | 0.044 | -0.006 | 0.040 | 0.003 | 0.045 |
| $\beta_{41}$ | 0.057 | 0.000 | 0.049 | 0.003 | 0.033 | -0.001 | 0.046 |
| $\eta_{01}$ | -0.136 | -0.011 | 0.052 | -0.003 | 0.042 | 0.009 | 0.056 |
| $\eta_{11}$ | -0.085 | -0.021 | 0.093 | -0.004 | 0.069 | 0.000 | 0.108 |
| $\eta_{21}$ | 0.435 | -0.005 | 0.049 | -0.006 | 0.045 | 0.003 | 0.051 |
| $\beta_{12}$ | 0.155 | -0.127 | 0.141 | -0.130 | 0.143 | -0.116 | 0.131 |
| $\beta_{22}$ | 0.132 | 0.004 | 0.029 | 0.002 | 0.025 | 0.000 | 0.025 |
| $\beta_{32}$ | -0.042 | 0.006 | 0.052 | 0.002 | 0.054 | -0.001 | 0.053 |
| $\beta_{42}$ | 0.016 | 0.000 | 0.057 | 0.006 | 0.051 | 0.004 | 0.047 |
| $\eta_{02}$ | 0.052 | -0.005 | 0.063 | -0.004 | 0.056 | 0.004 | 0.056 |
| $\eta_{12}$ | -0.360 | -0.007 | 0.114 | -0.010 | 0.101 | -0.010 | 0.102 |
| $\eta_{22}$ | 0.419 | 0.002 | 0.051 | -0.006 | 0.052 | -0.003 | 0.060 |
| **DIC value** | – | 18333.44 | | 10418.97 | | 18228.97 | |

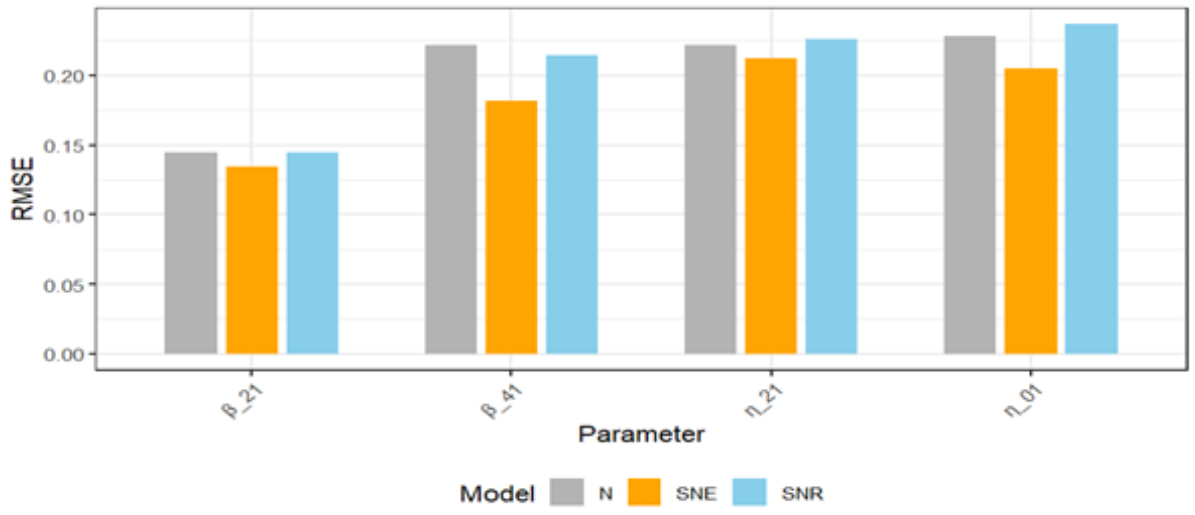



Figure 1.7: Bar plots of the estimated root mean square error for the semiparametric multivariate mixed effect model with different distributions of random errors and random effects

### 1.3.3 Discussion

A flexible semiparametric multivariate mixed-effect model was used to analyse correlated multivariate longitudinal outcomes that exhibit skewness and nonlinear trajectory pat-

terns over time, with Bayesian parameter estimation. We illustrate the proposed model to examine the relationship between longitudinal patterns of glucose concentration and blood pressure, and to assess the effects of covariates that influence these patterns among individuals with type 2 diabetes and hypertension. Incorporating the asymmetric feature of data into the mixed-effect model by modifying the commonly assumed distributions has received much attention over the past few years (Huang et al., 2022) and (Xu et al., 2021). Our study results also support this conclusion: the skewness parameters in the semiparametric multivariate mixed-effect models are significantly positive, indicating positive skewness (right skew) in both glucose concentration and blood pressure data. Moreover, the semiparametric multivariate mixed-effects model better fits the data than the fully parametric model. This finding is consistent with those reported by (Aniley et al., 2025) and (Huang et al., 2011).

The results from the various mixed-effect models indicate that including a skew-normal distribution and a non-linear time effect improve the efficiency and precision of parameter estimation. The simulation studies also confirmed the results of the real-data analysis, showing that Model SNE achieves greater efficiency and accuracy in parameter estimation when the covariate exhibits a non-linear relationship and when significant skewness is present in the response variables. This result is similar to those reported in (Huang et al., 2011, 2022; Ferede et al., 2024).

The study also compares the parameter estimates obtained from the semiparametric mixed-effects model, which assumes that random effects follow a multivariate skew distribution (model SNR), with those from the multivariate normal distribution (model N). We found that the modeling results based on the multivariate skew-normal distribution for random effects are similar to those from a multivariate normal distribution. This finding is consistent with (Huang and Dagne, 2010) and (Huang et al., 2011), despite using a data set on HIV dynamic response, and it features a mixed-effects joint model with a skew-normal distribution.

From a clinical perspective, we found a strong, direct relationship between the trajec-

tories of FBS and SBP, with the rate of change increasing over time. This finding is consistent with several earlier studies showing an association between blood sugar and blood pressure in patients with T2D (Suharto and Nurseskasatmata, 2020) and (Lv et al., 2018). Even if patients diagnosed with hypertension and diabetes typically start treatment immediately to lower both glucose and blood pressure, our results contradicted this expectation, showing a rapid increase in both glucose levels and blood pressure over time. Results from this study will inform healthcare policy and improve service delivery, making care for patients with the two correlated chronic diseases more effective and efficient.

Age is directly associated with glucose concentration and blood pressure; as diabetic and hypertensive individuals are aging, they tend to have elevated blood pressure and glucose levels over time. Even though the extent of the relationships varies across studies, our findings are similar to those of a previous study in (Jaffa et al., 2016; Godana et al., 2023; Feleke et al., 2021). The study suggests that older individuals with type 2 diabetes should effectively manage their glucose and blood pressure levels, and healthcare professionals should consider their age when planning treatment.

The trajectory of fasting blood glucose and blood pressure levels in individuals with type 2 diabetes and hypertension living in rural areas is faster than that of those residing in urban areas. This finding is consistent with the research by (Shita and Isayu, 2022; Andargie and Zeru, 2018), which indicates that individuals living in rural areas experience a faster progression of glucose concentration and blood pressure. The possible explanation for this difference may be the lower awareness of treatment adherence among individuals with type 2 diabetes and hypertension who live in rural areas. Therefore, healthcare professionals should provide clear information and provide focused attention to individuals from these areas to help them manage their glucose levels and blood pressure effectively.

Most previous analyses of the relationship between glucose and blood pressure trajectories focused on regression models and correlation coefficients from measurements taken at single time points. These methods do not reveal how the two patterns change and relate over time. Our model addresses these limitations by using follow-up measurements

of blood sugar and blood pressure over time in people with T2D and hypertension, applying a flexible mixed-effects model. Additionally, the model accounts for the nonlinear trajectory of the time effect and the skewness of the longitudinal measurements, offering greater flexibility in understanding the association between blood glucose and blood pressure trajectories and enhancing the model's fit to the data.

The proposed method may be of significant importance for chronic disease research that involves longitudinal follow-up data, characterized by nonlinear effects of measurement time and a departure from symmetric distributions. This is because accurate statistical inference about disease progression is essential for robust conclusions and reliable clinical decisions. Although the semiparametric mixed-effects model provides flexibility in modelling a nonlinear covariate effect on correlated, skewed longitudinal outcomes, it has a limitation in interpreting the nonlinear impact of covariates on these outcomes. Moreover, the statistical inference for the random effect was conducted using a parametric approach, which assumed either a multivariate normal distribution or a multivariate skew normal distribution due to computational intensity; however, it is also possible to estimate this non-parametrically using a nonparametric Dirichlet process. Our models may be extendable to more advanced mixed-effect models. For instance, (I) statistical inference for nonparametric random effects could be integrated to enhance the capture of within-subject correlation (Li et al., 2010) while the longitudinal outcomes exhibit skewness. (II) Alternative and more flexible distributional assumptions could be explored for both random errors and random effects; for example, multivariate skew-t distributions (Chen and Huang, 2015), multivariate normal/independent distributions (Bandyopadhyay et al., 2010), and fully nonparametric distributions. Although these extensions are beyond the present scope, our ongoing investigation warrants their inclusion in our future research.

### 1.3.4 Conclusion

We have proposed a semiparametric Bayesian modelling approach with flexible distributions for multivariate longitudinal data to examine the progression of glucose concentra-

tion and blood pressure and identify their determinant in people with type 2 diabetes and hypertension. Accounting for skewness in the model is essential when data exhibits noticeable skewness, as estimated parameters can differ significantly between models with skewed distributions and those with normal distributions. Our simulation studies also suggested that a semiparametric Bayesian modelling approach with flexible distributions provides relatively unbiased parameter estimates and has lower RMSE than the normal distribution. We recommend carefully considering the specifications of functional forms for longitudinal biomarkers, the distributional assumptions of model errors, and the correlations among multiple longitudinal outcomes when modeling complex longitudinal data.

# Bibliography


Andargie, A. A. and Zeru, M. A. (2018). A longitudinal data analysis on risk factors for developing type-2 diabetes mellitus at the university of gondar comprehensive specialized hospital, gondar, ethiopia. *Journal of Public Health and Epidemiology*, 10(6):171–182.

Aniley, T. T., Debusho, L. K., and Diriba, T. A. (2025). Bivariate semiparametric mixed models to longitudinally measured systolic and diastolic blood pressure of adult diabetic patients. *Annals of Data Science*, pages 1–28.

Aniley, T. T., Debusho, L. K., Nigusie, Z. M., Yimer, W. K., and Yimer, B. B. (2019). A semi-parametric mixed models for longitudinally measured fasting blood sugar level of adult diabetic patients. *BMC medical research methodology*, 19(1):13.

Arellano-Valle, R. B., Bolfarine, H., and Lachos, V. H. (2007). Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, 34(6):663–682.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, pages 171–178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602.

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

Bandyopadhyay, D., Lachos, V. H., Abanto-Valle, C. A., and Ghosh, P. (2010). Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. *Statistics in Medicine*, 29(25):2643–2655.

Chen, J. and Huang, Y. (2015). A bayesian mixture of semiparametric mixed-effects joint models for skewed-longitudinal and time-to-event data. *Statistics in medicine*, 34(20):2820–2843.

Feleke, B. E., Feleke, T. E., Kassahun, M. B., Adane, W. G., Fentahun, N., Girma, A., Alebachew, A., Misgan, E., Desyibelew, H. D., Bayih, M. T., et al. (2021). Glycemic control of diabetes mellitus patients in referral hospitals of amhara region, ethiopia: A cross-sectional study. *BioMed research international*, 2021(1):6691819.

Ferede, M. M., Dagne, G. A., Mwalili, S. M., Bilchut, W. H., Engida, H. A., and Karanja, S. M. (2024). Flexible bayesian semiparametric mixed-effects model for skewed longitudinal data. *BMC Medical Research Methodology*, 24(1):56.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Ghosh, P., Branco, M. D., and Chakraborty, H. (2007). Bivariate random effect model using skew-normal distribution with application to hiv-rna. *Statistics in medicine*, 26(6):1255–1267.

Godana, A. A., Molla, B. T., and Abatihun, D. (2023). Bayesian longitudinal modeling of blood pressure measurements of hypertensive patients at wachemo university nigist elleni mohamed memorial teaching and referral hospital hosanna, southern ethiopia. *Heliyon*, 9(12).

Huang, Y., Chen, J., Xu, L., and Tang, N.-S. (2022). Bayesian joint modeling of multivariate longitudinal and survival data with an application to diabetes study. *Frontiers in big Data*, 5:812725.

Huang, Y., Chen, R., and Dagne, G. (2011). Simultaneous bayesian inference for linear, nonlinear and semiparametric mixed-effects models with skew-normality and measurement errors in covariates. *The International Journal of Biostatistics*, 7(1):000010220215574679 1292.

Huang, Y., Chen, R., Dagne, G., Zhu, Y., and Chen, H. (2015). Bayesian bivariate linear mixed-effects models with skew-normal/independent distributions, with application to aids clinical studies. *Journal of Biopharmaceutical Statistics*, 25(3):373–396.

Huang, Y. and Dagne, G. (2010). Skew-normal bayesian nonlinear mixed-effects models with application to aids studies. *Statistics in Medicine*, 29(23):2384–2398.

Jaffa, M. A., Gebregziabher, M., Luttrell, D. K., Luttrell, L. M., and Jaffa, A. A. (2016). Multivariate generalized linear mixed models with random intercepts to analyze cardiovascular risk markers in type-1 diabetic patients. *Journal of applied statistics*, 43(8):1447–1464.

Koye, D. N., Melaku, Y. A., Gelaw, Y. A., Zeleke, B. M., Adane, A. A., Tegegn, H. G., Gebreyohannes, E. A., Erku, D. A., Tesfay, F. H., Gesesew, H. A., et al. (2022). Mapping national, regional and local prevalence of hypertension and diabetes in ethiopia using geospatial analysis. *BMJ open*, 12(12):e065318.

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

Li, Y., Lin, X., and Müller, P. (2010). Bayesian inference in semiparametric mixed models for longitudinal data. *Biometrics*, 66(1):70–78.

Lin, T.-I. and Wang, W.-L. (2013). Multivariate skew-normal at linear mixed models for multi-outcome longitudinal data. *Statistical Modelling*, 13(3):199–221.

Liu, Y., Li, J., Dou, Y., and Ma, H. (2021). Impacts of type 2 diabetes mellitus and hypertension on the incidence of cardiovascular diseases and stroke in china real-world setting: a retrospective cohort study. *BMJ open*, 11(11):e053698.

Lv, Y., Yao, Y., Ye, J., Guo, X., Dou, J., Shen, L., Zhang, A., Xue, Z., Yu, Y., and Jin, L. (2018). Association of blood pressure with fasting blood glucose levels in northeast china: a cross-sectional study. *Scientific reports*, 8(1):7917.

Petrie, J. R., Guzik, T. J., and Touyz, R. M. (2018). Diabetes, hypertension, and cardiovascular disease: clinical insights and vascular mechanisms. *Canadian Journal of Cardiology*, 34(5):575–584.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.

Reaven, P. D., Emanuele, N. V., Wiitala, W. L., Bahn, G. D., Reda, D. J., McCarren, M., Duckworth, W. C., and Hayward, R. A. (2019). Intensive glucose control in patients with type 2 diabetes—15-year follow-up. *New England Journal of Medicine*, 380(23):2215–2224.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics*, 31(2):129–150.

Shita, N. G. and Isayu, A. S. (2022). Predictors of blood glucose change and microvascular complications of type 2 diabetes mellitus patients in felege hiwot and debre markos referral hospital, north west ethiopia. *BMC Endocrine Disorders*, 22(1):136.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):1–21.

Suharto, I. P. S. and Nurseskasatmata, S. E. (2020). Blood glucose influence on cholesterol and blood pressure of patients with type ii diabetes mellitus. *STRADA: Jurnal Ilmiah Kesehatan*, 9(2):629–634.

Sun, H., Saeedi, P., Karuranga, S., Pinkepank, M., Ogurtsova, K., Duncan, B. B., Stein, C., Basit, A., Chan, J. C., Mbanya, J. C., et al. (2022). Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes research and clinical practice*, 183:109119.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433):217–221.

Wu, H. and Ding, A. A. (1999). Population hiv-1 dynamics in vivo: applicable models and inferential tools for virological data from aids clinical trials. *Biometrics*, 55(2):410–418.

Wu, H., Zhao, C., and Liang, H. (2004). Comparison of linear, nonlinear and semi-parametric mixed-effects models for estimating hiv dynamic parameters. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 46(2):233–245.

Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to aids studies. *Journal of the American Statistical association*, 97(460):955–964.

Xu, L., Huang, Y., Chen, H., Mbah, A., and Cheng, F. (2021). Joint modeling analysis of multivariate skewed-longitudinal and time-to-event data with application to primary biliary cirrhosis study.

Yan, C., Chen, R., and Huang, Y. (2016). Mixed-effects models with skewed distributions for time-varying decay rate in hiv dynamics. *Communications in Statistics-Simulation and Computation*, 45(2):737–757.

Yirdaw, B. E. and Debusho, L. K. (2023). Semiparametric modelling of diabetic retinopathy among people with type ii diabetes mellitus. *BMC medical research methodology*, 23(1):7.

# Chapter 2

# Bayesian Joint Modeling of Bivariate Longitudinal and Time-to-Event Data: With Application of Micro and Macro Vascular Complications in People with Type 2 Diabetes and Hypertension


## Abstract

Many medical and epidemiological studies record longitudinal measurements until a time-to-event outcome occurs. When there is an association between the time-to-event and the longitudinal outcomes, separately modelling them may lead to biased estimates. Joint modelling of longitudinal and survival data is an effective method for analysing their relationship. In most cases, univariate joint modelling of longitudinal and time-to-event outcomes is an effective method to evaluate their association. However, this model-based analysis can yield biased estimates when multiple longitudinal outcomes are highly correlated and do not follow a multivariate normal distribution. To overcome the problem, we develop a bivariate joint model with a skewed multivariate normal distribution, providing a flexible approach for non-symmetric and correlated longitudinal outcomes. The proposed model specification consists of two sub-models linked by shared random effects. This involved the Cox proportional hazards model for time-to-event data and the multivariate mixed-effects model for correlated longitudinal data following bivariate skew-normal distributions. We estimate the parameters using a Bayesian framework and implement Markov chain Monte Carlo methods in R with JAGS. We assess the performance of the proposed models via simulations and apply the methodology to a data set to assess the association between longitudinal blood sugar and blood pressure measures and time to chronic complications. Our studies suggest a strong, significant positive rela-

tionship between patterns of blood glucose levels and systolic blood pressure. Over time, increases in systolic blood pressure raised the risk of microvascular and macrovascular complications. In bivariate joint modeling, employing a skewed multivariate normal distribution instead of a standard multivariate normal distribution improves model fit and yields more accurate parameter estimates for the longitudinal biomarker.

**key words**, Bayesian inference, bivariate joint model, longitudinal and time to event data, type 2 diabetes and hypertension, micro and macro vascular complications.

## 2.1 Introduction

Joint modeling of longitudinal and survival outcomes involves simultaneously analyzing longitudinal and survival data, accounting for the relationship between repeated measurements and time-to-event outcomes. It consists of two sub-models: a longitudinal sub-model for longitudinal outcomes and a survival sub-model for time-to-event outcomes (Faucett and Thomas, 1996) and (Guo and Carlin, 2004). Joint modeling of longitudinal and survival outcomes is an active area of research in biostatistics and epidemiology. It improves inference for time-to-event outcomes by accounting for repeated measurements of endogenous time-varying covariates. It also measures the rate of change in endogenous time-varying covariates, accounting for differences between subjects and changes over time within the same subject, while exploring the relationship between repeated measurements and the time-to-event outcome (Guo and Carlin, 2004) and (Tsiatis and Davidian, 2004).

Follow-up studies often involve repeated measurements of multiple responses from the same subject over time, together with one or more time-to-event outcomes. For example, in studies of diabetes and hypertension, researchers and physicians typically collect two types of outcomes: longitudinal measurements of glucose concentration and blood pressure at each follow-up and the time to onset of microvascular or macrovascular complications. Glucose concentration is commonly used as a longitudinal marker of diabetes severity, while blood pressure serves as a marker of hypertension severity. In addition,

investigators are often interested in determining when an event will occur, how many individuals experience it, and how repeated measurements influence the outcome.

Most of the joint models in the existing statistical literature focus on (I) models with a single longitudinal outcome linked to time-to-event data; see, for example, (Baghfalaki et al., 2013) and (Sweeting and Thompson, 2011). However, patient assessments often include two or more longitudinal outcomes that could affect time-to-event results. Extending the univariate joint model to the multivariate joint model that considers possible correlated longitudinal outcomes enables us to incorporate more information about the time-to-event outcome. This enhances prognostic accuracy and provides researchers with a deeper understanding of the complex dynamics of disease progression, ultimately leading to more accurate effect estimates (Huang et al., 2022).

II), The commonly assumed distribution for model error in the longitudinal outcome submodel is the normal distribution due to mathematical tractability and computational convenience. However, the assumption of a multivariate normal distribution may not be realistic and could lead to inaccurate statistical inferences (Ferede et al., 2022) and (Hickey et al., 2018). As a result, a normal distribution for model errors may not capture true variability within and between individuals, and it may not be strong enough to handle deviations from symmetry.

III), Many studies in practice aim to gather data on multiple longitudinal responses, potentially showing significant correlations. For instance, there may be a correlation between FBS and SBP; failing to account for this correlation could lead to biased results and reduce the efficiency of effect estimation. The classical joint model has been formulated in the univariate data framework; ignoring the correlation between longitudinal exposures can lead to bias (Huang et al., 2022).

To address the identified limitation, we employed a Bayesian bivariate joint model, using an appropriate distribution for bivariate longitudinal data while optimizing inference of the time-to-event process. A fully Bayesian framework was utilized for statistical inference due to its computational efficiency and its ability to facilitate inference from

complex joint models of this type. The novelty of this study lies in its integration of three pivotal elements: (I) modelling longitudinal data with inherent feature distributions; (II) within-subject correlations arising from repeated measurements within each subject; and (III) bivariate longitudinal outcomes that exhibit correlation. The main objective of this study is to introduce a bivariate joint modelling method to examine the dynamic association between measuring multiple biomarkers over time and their joint effect on the risk of micro- and macrovascular complications in people with type 2 diabetes and hypertension.

## 2.2 literature Review

T2D and hypertension are frequently coexisting, and persons with diabetes have a two-to-four-fold increased risk of hypertension compared with persons without diabetes (Mancia, 2005). Each increases the risk of microvascular and macrovascular complications independently, but when they co-occur, the risk increases significantly, especially for women (Ceriello et al., 2023). Coexistence of hypertension and diabetes is associated with increased microvascular and macrovascular complications like cardiovascular disease, chronic kidney disease, stroke, and retinopathy (Viazzi et al., 2019) and (Wan et al., 2020). An et al. (2015) identified these complications as the primary cause of death in diabetic and hypertension patients, with 76.4% of diabetic patients reporting at least one complication (Hu et al., 2015).

Multiple studies have found that poor glycaemic control, as measured by glycosylated haemoglobin (HbA1c), fasting blood sugar, and systolic and diastolic blood pressure, increases the risk of chronic complications. For example, (Caruso et al., 2021; Ceriello et al., 2023; Sartore et al., 2023; Wan et al., 2016) demonstrated that the development of chronic complications in people with type 2 diabetes is highly associated with variability of HbA1c, FBS, and SBP. Furthermore, variability in HbA1c and blood pressure influences the development of chronic kidney disease, with different effects on albuminuria and glomerular filtration rate (Ceriello et al., 2017).

To understand the nature of disease progression and identify risk factors for micro- and macrovascular complications based on clinical and biochemical variables, an appropriate statistical model is needed. Most the previous studies have employed classical regression approaches, including logistic regression, linear regression, and Cox proportional hazards models to predict the time to onset of microvascular and macrovascular complications (Ceriello et al., 2022; Gao et al., 2022; Dorajoo et al., 2017). While these models are relatively straightforward to implement, they have notable limitations. When assessing the effects of risk factors on complications, they often rely on single measurements per risk factor, typically obtained during a single clinical visit, thereby disregarding repeated measurements collected over time.

Risk factors such as FBS, HbA1, diastolic blood pressure (DBP), and SBP are inherently dynamic and are measured repeatedly over time. Including the dynamic nature of repeated measurements could enhance the accuracy of risk estimates for micro- and macrovascular complications. Since repeated measurements accumulate substantial information that affects the hazard of the event (the risk of chronic complications), an appropriate statistical model is needed to incorporate this information into the risk of chronic complications for patients with diabetes and hypertension.

The longitudinal data can be essential predictors or surrogates of a time-to-event, such as disease diagnosis, organ transplantation, chronic complications, or death. Most studies analyze longitudinal and survival data separately. The survival regression model, especially the Cox regression model for time-to-event data, and the linear mixed-effects model for longitudinal data are active areas of research for analyzing outcomes separately. The limitations of separate modelling of longitudinal and survival outcomes have been extensively discussed in the literature: methods that analyse them separately yield biased effect-size estimates, especially when they are correlated (Ibrahim et al., 2010).

Joint modelling is a novel and robust method for resolving these deficiencies. Joint modelling, which simultaneously models both longitudinal and survival processes, can effectively incorporate longitudinal information into the survival model (Tsiatis and Da-

vidian, 2004). Joint modelling is more flexible than separate analyses and may better reflect biological plausibility because repeated measurements are a strong biomarker for time-to-event outcomes. Over the past two decades, there has been marked growth in both methodological advances and the practical application of joint models for longitudinal and time-to-event outcomes in the literature. Most of the early methodological work was motivated by problems arising in ADIS and cancer research, for example (Baghfalaki et al., 2013) and (Guo and Carlin, 2004). Currently, joint modeling methods have been used in other areas of clinical research, including studies on cardiovascular disease and kidney transplantation (Jun et al., 2017).

In this study, we examined individuals with type 2 diabetes and hypertension. Throughout the follow-up period, various types of data on individuals with diabetes and hypertension, including their overall health status, were collected, including periodic measurements of blood glucose levels and systolic and diastolic blood pressure. Additionally, the occurrence of diabetes and hypertension-related complications, such as cardiovascular disease, chronic kidney disease, stroke, and retinopathy, was recorded. Repeated measurements, such as FBS, SBP, and DBP, serve as important endogenous covariates or surrogates for time-to-event outcomes, such as the onset of diabetic complications. As a result, data for patients with diabetes and hypertension include two distinct types of outcomes: longitudinal outcomes, like FBS and SBP measured at multiple time points, and time-to-event outcomes, such as the duration until the occurrence of chronic complications.

Therefore, using a novel Bayesian multivariate joint modeling approach, we aim to examine the relationship between the trajectories of blood pressure and glycemic control over time and the risk of chronic complications among people with T2D and hypertension. After the other covariate is considered, our central hypothesis is that a rise in FBS and SBP would be strongly linked to all-cause of chronic complications.

## 2.3 Methodology

### 2.3.1 Joint Modeling for Bivariate Longitudinal and Time-to-Event Data

Joint modeling of longitudinal and survival data is a method that simultaneously analyzes repeated measurements and time-to-event outcomes. The bivariate joint model, which incorporates multiple longitudinal exposures, extends the univariate joint model. It consists of two basic components: the longitudinal component and the time-to-event (survival) component. The study employs a bivariate linear mixed-effects model to analyse fast blood glucose and systolic blood pressure (longitudinal sub-model) and a Cox proportional hazards model to analyse time to Microvascular and macrovascular complications (survival sub-model). The Bivariate joint model links these two data types through subject-specific random effects to enhance statistical inference and association between bivariate longitudinal and time-to-event data. We proceed by addressing the longitudinal sub-model, followed by the time-to-event sub-model. Additionally, we explore how to associate these two sub-models using subject-specific random effects and random slopes.

### 2.3.2 The longitudinal Outcomes Sub-model

The longitudinal data set consists of follow-up measurements of two outcomes for each study unit over a specified study period. These follow-up measurements for the same study unit are generally interdependent. Consequently, each study unit in the population is expected to exhibit a unique pattern of outcomes over time. The mixed-effects model addresses this between-subject variability by estimating subject-specific random effects in addition to the fixed parameters that are consistent across individuals (Laird and Ware, 1982). Suppose there are n subjects and let $Y_{ijk}$ be a measurement of the $K^{\text{th}}$ longitudinal outcome of the $i^{\text{th}}$ subject measured at time $j = 1, 2, 3, \ldots, n_i, \quad k = 1, 2, \ldots, k, \quad i = 1, 2, \ldots, n$. Thus, the bivariate linear mixed-effects model for the longitudinal sub-model

can be expressed as follows.

$$Y_{ijk} = \mu_{ij} + e_{ijk}, \tag{2.1}$$

$$\mu_{ij} = \beta_k X_{ik} + b_{ik} Z_{ik}, \tag{2.2}$$

$$Y_{ijk} = \beta_k X_{ik} + b_{ik} Z_{ik} + e_{ijk}. \tag{2.3}$$

Where $X_{ik}$ and $Z_{ik}$ represent the design matrices containing the predictors for the fixed-effects regression coefficients $\beta_k$ and for the random-effects regression coefficients $b_i$, respectively.

$$\text{where } b_i = \left[b_{i11}, b_{i21}, b_{i12}, b_{i22}\right] \sim SN(0,\ \Sigma_b),\ \Sigma_b = \begin{bmatrix} v_{11} & v_{12} & v_{13} & v_{14} \\ v_{21} & v_{22} & v_{23} & v_{24} \\ v_{31} & v_{32} & v_{33} & v_{34} \\ v_{41} & v_{42} & v_{43} & v_{44} \end{bmatrix}$$

$$e_{ijk} = \begin{bmatrix} e_{i1} \\ e_{i2} \end{bmatrix} \sim SN(0, \Delta_{ek},\ \Sigma_k)\ ,\Sigma_k = \begin{bmatrix} \text{var}(\varepsilon_{i1}) & \text{cov}(\varepsilon_{i1}, \varepsilon_{i2}) \\ \text{cov}(\varepsilon_{i2}, \varepsilon_{i1}) & \text{var}(\varepsilon_{i2}) \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix}.$$

The random error term $e_{ijk}$ and the random effects $b_{ik}$ are assumed to be independent of each other. As presented graphically in section 2.3.5, the longitudinal outcome of fast blood glucose and systolic blood pressure follows an asymmetric (skew) distribution. Hence, in this study, we assume that the model errors $e_{ijk}$ follow a multivariate skew-normal distribution with mean vector 0 and covariance matrix $\Sigma_k$ and the skewness parameter $\Delta ek$. For more detailed discussions on the explanation of the multivariate skewed normal distribution and its potential applications. let us consider a $k$-dimensional random vector $Y$ that follows a $k$-variate skew-normal distribution with location vector $\mu$, dispersion matrix $\Sigma \in \mathbb{R}^{k\times k}$, and skewness matrix $\Delta = \text{diag}(\delta)$, with $\delta = (\delta_1, \delta_2, \ldots, \delta_k)^\top$

being the skewness parameter vector, if its pdf is given by:

$$f(Y \mid \mu, \Sigma, \Delta) = 2^k \left|\Sigma + \Delta\right|^{-1/2} \phi_k\left[(\Sigma + \Delta)^{-1/2}(Y - \mu)\right] \times \Phi_k\left[\left(I_k - \Delta(\Sigma + \Delta^2)^{-1}\Delta\right)^{-1/2} \Delta(\Sigma + \Delta^2)^{-1}(Y - \mu)\right]. \tag{2.4}$$

where $\phi_k$ and $\Phi_k$ represent the PDF of Y and the CDF of Y, respectively. $I_k$ denotes the $k \times k$ identity matrix. We formally denote the defined PDF of Y in short as

$$Y \sim \mathrm{SN}_k(\mu, \Sigma, \Delta),$$

with expectation and covariance given by $E(Y) = \mu + \sqrt{\frac{2}{\pi}}\,\delta, \mathrm{cov}(Y) = \Sigma + \left(1 - \frac{2}{\pi}\right)\Delta^2$. According to (Sahu et al., 2003), if $Y \sim \mathrm{SN}_k(\mu, \Sigma, \Delta)$, then stochastic representation of $Y$ :

$$Y = \mu + \Delta|x_0| + \Sigma^{1/2}x_1, \tag{2.5}$$

where $x_0$ and $x_1$ are independent random vectors with distribution $N_k(0, I_k)$. Let $w = |x_0|$; then $w$ follows a $k$-dimensional standard normal distribution $N_k(0, I_k)$ restricted to the space $w > 0$. Thus, a two-level hierarchical representation of the model in equation 2.3 is given by:

$$Y \mid w \sim N_k(\mu + \Delta w, \Sigma), \qquad w \sim N_k(0, I_k)\, I(w > 0). \tag{2.6}$$

Note that when $\Delta = 0$, the hierarchical representation of equation 2.6 above reduces to the calsical multivariate normal distribution $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

### 2.3.3 The time to Event Outcome Sub-model

Survival analysis is a widely used statistical method in medical research for modeling and examining the time to the occurrence of a particular event or outcome. It is also called reliability analysis in engineering and duration analysis in economics or sociology. The survival submodel is used to estimate the time to micro- and macrovascular complications.

Cox proportional hazards regression is a popular technique for examining how exposure affects the duration time until an event occurs (Cox, 1972). Let $T_i^*$ be the true event time and $C_i$ the censoring time. Then, the observed event time for subject $i$ is defined as $T_i = \min(T_i^*, C_i)$ , and let $\tau_i$ denote the event indicator for the $i^{\text{th}}$ subject, given by

$$\tau_i = \begin{cases} 1, & \text{if the event occurs during the study period,} \\ 0, & \text{if the event does not occur or the subject is lost to follow-up.} \end{cases}$$

For the survival (time-to-event) outcome, we propose the following hazard function

$$\lambda_i(t \mid G_i(t), X_i) = \lambda_0(t) \exp\Big(M\big(G_i(t)\big) + \alpha^\top X_i\Big). \tag{2.7}$$

where $G_i(t) = \mu_{ij},\ 0 < j < t$ is the history of the longitudinal process up to time $t$, $\lambda_0(t)$ is the baseline hazard function, $\alpha$ is the vector of coefficient parameters corresponding to the covariates $X$. The function $M(\cdot)$ is a known function of $G_i(t)$ and specifies the features of the longitudinal outcome process that are included as predictors of the time-to-event outcome. Different forms of association between the longitudinal process $M(\cdot)$ and the hazard function $\lambda(t)$ may be specified, such as:

$$M(G_i(t)) = \begin{cases} \gamma\, \mu_i(t), & \text{hazard associated with the underlying level of the biomarker at time } t \\ \gamma\, \dfrac{d\mu_i(t)}{dt}, & \text{hazard associated with the slope of the longitudinal profile at time } t \\ \gamma \displaystyle\int_0^t \mu_i(s)\, ds, & \text{hazard associated with the accumulated longitudinal process up to time } t \\ \gamma^\top b_i, & \text{hazard associated with the random effects} \end{cases}$$

In all cases, $\gamma$ denotes the vector of coefficient parameters representing the effect of the longitudinal trajectory on the event time.

Each type of association can be applied in joint models depending on the specific research

objective; in this study, we employ a shared random-effect association. The distribution of $T_i$, the time to diagnosis of micro and macro vascular complications for subject i, depends on the random-effects of individual-subject specific longitudinal processes $b_{ik}$ which are computed from equation 2.3 and other baseline measurement values of clinical and biochemical variables. The two outcome processes, the longitudinal and time-to-event outcomes, are associated through the random effect $b_{ik}$. It is also expected that the other baseline covariates are risk predictors for the event time. In particular, the conditional hazard rate of $T_i$ at time $t_i$ is represented as.

$$\lambda(t_i \mid b_{ik}, X_i) = \lambda_0(t_i) \exp\big(\boldsymbol{\gamma}^\top b_i + \alpha^\top X_i\big) \tag{2.8}$$

where $\lambda_0(t_i)$ is the baseline hazard value, $\gamma$ is the vector of coefficient parameters corresponding to the random effects $b_{ik}$, and $\alpha$ is the vector of coefficient parameters corresponding to the covariates $X_i$. The parameter vector $\boldsymbol{\gamma}$ links the two sub-models and indicates the effect of the longitudinal process on the time-to-event outcome. Let $\tau_i = (\tau_{i1}, \ldots, \tau_{in_i})^\top$ be an indicator of whether the event occurred by follow-up time $t_{ij}$ or not. Specifically,

$$\tau_{ij} = \begin{cases} 1, & \text{if the subject develops a chronic complication by time } t_{ij}, \\ 0, & \text{if the subject does not develop a chronic complication by time } t_{ij}. \end{cases}$$

From Equation 2.8, the probability of the event time is

$$\begin{aligned} P(T \leq t_{ij}) = p_{ij} &= P(\tau_{ij} = 1 \mid \tau_{ij^*} = 0,\ 0 < j^* < j) \\ &= 1 - S(T), \end{aligned} \tag{2.9}$$

where $S(T)$ denotes the survival probability at time $t_{ij}$, that is, the probability that an individual remains free from chronic complications up to $t_{ij}$, thereby quantifying the

risk-free duration. Formally,

$$
\begin{aligned}
S(T) = P(T_i = t_{ij} \mid T_i > t_{i(j-1)}) &= 1 - \frac{P(T_i \geq t_{ij})}{P(T_i \geq t_{i(j-1)})} \\
&= 1 - \exp\left( - \int_{t_{i(j-1)}}^{t_{ij}} \lambda_0(t)\, dt \, \exp\big(\Upsilon^\top b_{ik} + \alpha^\top X_i\big) \right) \\
&= 1 - \exp\Big( - \exp(\upsilon_{0j} + \Upsilon^\top b_{ik} + \alpha^\top X_i) \Big), \qquad (2.10)
\end{aligned}
$$

where $\upsilon_{0j} = \log\left( \int_{t_{i(j-1)}}^{t_{ij}} \lambda_0(t)\, dt \right), \qquad j = 1, 2, \ldots, n_i$. Compared to parameter estimation based directly on the baseline hazard function $\lambda_0(t)$, the current observation system only requires estimation of the finite parameters $\upsilon_{0j}$, rather than the entire unknown baseline hazard. The contribution to the likelihood from the time-to-event model for the $i^{\text{th}}$ subject is

$$
f(\tau_i \mid b_{ik}, X_i) = \prod_{j=1}^{n_i} f(\tau_{ij} \mid \tau_{ij^*},\ 0 < j^* < j;\ b_{ik}, X_i) \qquad (2.11)
$$

where

$$
f(\tau_{ij} \mid \tau_{ij^*},\ 0 < j^* < j;\ b_{ik}, X_i) = p_{ij}^{\tau_{ij}}\, (1 - p_{ij})^{1 - \tau_{ij}},
$$

with $\tau_{ij} = 0$ indicating that the subject remains event-free and $\tau_{ij} = 1$ indicating that the event has occurred.

### 2.3.4 Bayesian Inference

Joint statistical inference across all parameters in both the longitudinal and survival models is essential to accurately capture and account for the underlying relationships between the longitudinal and time-to-event processes. The classical (Frequentist) approach of parameter estimation is a well-established method for estimating all unknown parameters using the joint likelihood function. However, this technique can be computationally demanding and may encounter convergence issues, particularly when using the joint likelihood method with the proposed model under a multivariate skew-normal distribution for random errors (Wu, 2002) and (Wu et al., 2010). The Bayesian inference approach offers a solution by reducing computational complexity and enabling the incorporation

of prior knowledge for the unknown parameters. Therefore, in this chapter, we used a fully Bayesian method with MCMC techniques for the bivariate linear mixed-effect sub-model (equation 2.3) and the survival sub-model (equation 2.8) to estimate all parameters simultaneously.

Recall that we assumed a multivariate skew-normal distribution for the random errors $e_{ijk}$ in the bivarite linear mixed effects model. Let $Y_{ijk}$ denote the measurement of the $k^{\text{th}}$ longitudinal outcome for the $i^{\text{th}}$ subject at time $j$, where $j = 1, 2, 3, \ldots, n_i$, $k = 1, 2$, and $i = 1, 2, \ldots, n$. To implement the MCMC procedure for the joint model, we introduce the $n_i \times 1$ random variable vector $w_i$, which is derived from the stochastic representation of the multivariate skew-normal distribution as shown in equation 2.6. Thus, we hierarchically reformulate the longitudinal and survival submodels as follows.

$$\begin{aligned} Y_i \mid b_i, w_i &\sim N_{2n_i}\big(\beta X_i + b_i Z_i + \Delta_k(w_i + \sqrt{2/\pi}\ \times I_{2n_i}), \Sigma_{ek} \times I_{2n_i}\big), \\ b_i &\sim N_{2(q+1)}(0, \Sigma_{bK}), \\ w_i &\sim N_{2n_i}(0, I_{n_i})\, I(w_i > 0), \\ T_i &\sim F(t_i \mid b_i, X_i, \lambda_0) = \int f(\tau_i \mid b_i, X_i), \end{aligned} \tag{2.12}$$

where $Y_i \mid b_i, w_i \sim N_{2n_i}\big(\beta X_i + b_i Z_i + \Delta_k(w_i + \sqrt{2/\pi}\ \times I_{2n_i}), \Sigma_k \times I_{2n_i}\big)$ is the conditional distribution of the longitudinal responses given the random effects $(b_i)$ and latent variable $(w_i)$, which follows a multivariate skew normal distribution with mean equal to the sum of the fixed effects $(\beta X_i)$, the random-effects $(Z_i b_i)$, and skewness term $(\Delta_k(w_i + \sqrt{2/\pi}))$ with covariance matrix equal to $\Sigma_k$. $X_i$ and $Z_i$ be covariates of the design matrices associated with the vector of fixed effects $\beta$ and the vector of random effects $b_i$, respectively. $i = 1, ..., n$, and $n$ is the total number of subjects, $n_i$ is the number of repeated measurements for subject $i$, and $q$ is the number of random slopes and intercepts for each longitudinal outcome. $b_i \sim N_{2(q+1)}(0, \Sigma_b)$ is the random effect, which follows a multivariate skew normal distribution with mean equal to zero and covariance $\Sigma_b$. $w_i$ is an auxiliary latent variable, which follows a multivariate normal distribution truncated to be greater than zero. These truncated normal variables generate the asymmetric com-

ponents of longitudinal response distributions, thereby forming the overall multivariate skew-normal structure. $\Delta_k$ is the skewness parameter for random errors and $I_{2n_i}$ is identity matrices. The event time for subject $i$, denoted as $T_i \sim F(t_i \mid b_i, X_i, \lambda_0)$, is described by CDF ($F(t_i \mid b_i, X_i, \lambda_0)$) that depends on the subject-specific random effects $b_i$, the covariate vector $X_i$, and the baseline hazard parameter $\lambda_0$. $X_i$ is the covariate that affects the event time. The right-hand expression indicates that this CDF is obtained by integrating the conditional event-time density $f(\tau_i \mid b_i, X_i)$ over time.

To complete the Bayesian specification, the observed data and the prior distributions for all unknown parameters should be defined. let $\Omega = \left(\beta, \Upsilon, \alpha, \Sigma_b, \Sigma_k, \Delta_k\right)$ be the collection of unknown population parameters in the bivariate linear mixed effect model 2.3 and the Cox regression sub model 2.8. All fixed-effect parameters, such as beta, alpha, gamma, and the skewness parameter, are assumed to follow a multivariate normal distribution. On the other hand, the random error term and the random effect variances and covariances follow an inverse Wishart distribution.

$$\begin{aligned} &\beta \sim N(\beta_0, \Psi_\beta), \quad \Upsilon \sim N(\Upsilon_0, \Psi_\Upsilon), \quad \alpha \sim N(\alpha_0, \Psi_\alpha), \\ &\delta_k \sim N(\delta_{k0}, \Psi_1), \quad \Sigma_{eK} \sim IW(\Psi_2, \omega_1), \quad \Sigma_{bk} \sim IW(\Psi_3, \omega_2). \end{aligned} \tag{2.13}$$

Where $\beta$ is a vector of fixed-effect coefficients for the bivariate correlated longitudinal outcomes, $\alpha$ is a vector of fixed-effect coefficients for the time-to-event outcomes. $\gamma$ is the shared random effect that relates the longitudinal and time-to-event outcomes. $\beta_0$, $\gamma_0$ and $\alpha_0$ are prior mean vectors for $\beta$, $\gamma$ and $\alpha$, respectively. $\Psi_\beta$, $\Psi_\gamma$ and $\Psi_\alpha$ are the prior covariance matrices $\beta$, $\gamma$ and $\alpha$, respectively. $\delta_{k0}$ and $\Psi_1$ are the prior mean vector and the prior covariance matrix, respectively, for the skewness parameter of the random error ($\delta_k$). $\Psi_2, \Psi_3$ are the scale or dispersion matrices for the variance covariance of the random errors ($\Sigma_{ek}$) and the variance covariance of the random effects ($\Sigma_{bk}$), respectively, and $\omega_1, \omega_2$ are the respective degrees of freedom.

For the convenience of implementation, we assume that the hyperparameter matrices $\Psi_\beta$, $\Psi_\Upsilon$, $\Psi_\alpha$, $\Psi_1$, $\Psi_2$, and $\Psi_3$ are diagonal. Furthermore, we assume that all parameters in the

parameter space ($\Omega$) are independent of each other, that is $\pi(\Omega) = \pi(\beta)\,\pi(\Upsilon)\,\pi(\Sigma_b)\,\pi(\Sigma_k)$ $\pi(\delta_k)$. Once we have defined the prior distributions for the unknown parameters, we can conduct Bayesian inferences for those parameters using their posterior distributions. Let $D_n = \{Y, X, T, \tau\}$ denotes the observed dataset for patients with diabetes and hypertension, encompassing longitudinal outcomes such as fasting blood sugar and systolic blood pressure, covariates from both sub-models, as well as recorded event times and corresponding event indicators. Furthermore, let $f(\cdot)$ be a density function, $f(\cdot/\cdot)$ be a conditional density function, $F(\cdot/\cdot)$ be a conditional cumulative density function, and $\pi(\cdot)$ be a prior density function. Then, the likelihood function is denoted by $L(\Omega \mid D_n) = f(D_n \mid \Omega)$, and its density function is given.

$$
\begin{aligned}
f(D_n \mid \Omega) = \prod_{i=1}^{n} \int & f_{2n_i}\Big(Y_i;\ \beta X_i + Z_i b_i + \Delta_k(w_i + \sqrt{2/\pi}\, I_{2n_i}), \Sigma_k I_{2n_i}\Big) \\
& \times f_{2(q+1)}(b_i;\ 0, \Sigma_b) \\
& \times f_{2n_i}(w_i; 0, I_{2n_i})\, I(w_i > 0) \\
& \times f(\tau_i \mid b_i, X_i)\ \ db_i
\end{aligned} \tag{2.14}
$$

After defining the prior distributions of the unknown model parameters and the observed data, we can draw samples for statistical inference from the posterior density by combining the likelihood function 2.14 and the priors. The joint posterior density of the unknown parameter is the product of the likelihood and the prior distributions. That is, the joint posterior density of $\Omega$ given the observed data, $D_n$, and the prior distribution given by $f(\Omega \mid D_n) \propto f(D_n \mid \Omega) \times \pi(\Omega)$

$$
\begin{aligned}
f(D_n \mid \Omega) = \Big\{ \prod_{i=1}^{n} \int & f_{2n_i}\Big(Y_i;\ \beta X_i + Z_i b_i + \Delta_k(w_i + \sqrt{2/\pi}\, I_{2n_i}), \Sigma_k I_{2n_i}\Big) \\
& \times f_{2(q+1)}(b_i;\ 0, \Sigma_b) \\
& \times f_{2n_i}(w_i; 0, I_{2n_i})\, I(w_i > 0) \\
& \times f(\tau_i \mid b_i, X_i)\ \ db_i \Big\}, \times \pi(\Omega)
\end{aligned}
$$

Bayesian estimates through direct utilization of the posterior distribution in equation 2.15 are not practically achievable. Nevertheless, the `BUGS` syntax provides tools to conduct Bayesian analyses of complex statistical models using MCMC techniques. We obtained samples from the full conditional posterior distributions of the parameters through the Metropolis–Hastings algorithm combined with the Gibbs sampler, as specified in equation 2.15. This allowed us to estimate the posterior means and standard deviations of the parameters. The MCMC procedure was implemented in `R` using `JAGS` via the `R2jags` package.

### 2.3.5 Application of Type 2 Diabetes and Hypertension Data

This research is motivated by longitudinal and time-to-event data obtained from patients with diabetes and hypertension. We are particularly interested in associating longitudinal measurements of fasting blood sugar fluctuations and systolic blood pressure with the time to onset of microvascular and macrovascular complications. Data were collected from individuals with type 2 diabetes and hypertension who received treatment at Felege Hiwot Comprehensive Specialized Hospital in northern Ethiopia. This study is a retrospective cohort analysis of subjects over five years, specifically from January 2018 to December 2022. The cohort includes all individuals diagnosed with type 2 diabetes and hypertension who started diabetes and hypertension treatment in 2018, had at least two measurements of fasting plasma glucose and systolic blood pressure, and did not have any prior history of micro- or macrovascular complications (such as retinopathy, nephropathy, neuropathy, heart disease, or stroke) at the start of treatment. The study excludes patients with type 1 or gestational diabetes, as well as those who have had only one clinical visit.

To illustrate the proposed methods, we selected 220 subjects according to the specified inclusion and exclusion criteria. From each study subject, we extract the clinical, demographic, and biochemical data from their medical chart. We took baseline clinical variables (body mass index, age, sex, 0 = male, 1 = female, residence, 0 = urban, 1 = rural, marital status, 1 = single, 2 = married) and biochemical variables (HDL, LDL, total cholesterol, and triglycerides) from their medical histories. The onset of complications,

the censoring time, and the repeated measurements of FBS and sSBP were extracted from the patients' medical records. Micro and macro vascular complications, including retinopathy, nephropathy, heart disease, stroke, and neuropathy, were assessed by physicians at each follow-up visit.

The following table and figure represent the details of the motivating data set. Table 2.1 presents the baseline demographic, clinical, and biochemical characteristics of the study subjects by event status (Non-event, free from micro- and macrovascular complications; Event, developing micro- and macrovascular complications at the end of the study period). During the five-year follow-up period, 73 (33.20%) individuals experienced micro- or macro-complications, while 147 (66.80%) individuals were either under follow-up at the end of the study or lost to follow-up during the study.

Of the 220 people with T2D and hypertension, 132 (60.00%) are male; the mean baseline FBS and SBP are 200.90 (62.27) mg/dL and 142 (14.21) mmHg, respectively. The mean age of people with T2D and hypertension was 53.75, with a standard deviation of 11.09 years. The descriptive table generally shows that people who developed micro- or macro-complications were older and had higher body mass index, lower HDL, higher triglycerides, higher total cholesterol, and higher LDL. They were also more likely to be taking medication to lower blood glucose levels and to be from urban areas. Figure 2.1 A and B present the histograms of repeated measurements of FBS and SBP levels, respectively. As shown in the graph, the distributions of the longitudinal outcome variables, FBS and SBP, deviate from normality (positively skewed). We further examined the normality of the outcome variables, FBS and SBP, using Kolmogorov-Smirnov and Shapiro-Wilk tests. Both show p-values less than 0.001, providing strong evidence that the data are not normally distributed. Therefore, a normality assumption is not quite realistic for this dataset. Figure 2.2 C and D show the fasting blood glucose and systolic blood pressure trajectories of five randomly selected patients. The randomly selected sample trajectories of FBS and SBP clearly demonstrated fluctuations over time. The cumulative incidence function and the survival function show the progression of micro-

and macrovascular complications over time. It indicates that the cumulative incidence of micro- and macro-vascular complications among individuals with type 2 diabetes and hypertension is 28 individuals at 24 months, 44 individuals at 32 months, 60 individuals at 48 months, and 73 individuals at 60 months. As illustrated in the figure, the cumulative incidence of micro- and macro-vascular complications is rising sharply each month. We are mainly focused on the cumulative incidence plots presented in figure 2.3 E, which may be influenced by the trajectories of fasting plasma glucose and systolic blood pressure, figure 2.1 A and B. Furthermore, disregarding the correlation between the two longitudinal outcomes and assuming normally distributed error terms may yield an inaccurate estimate.

Table 2.1: Descriptive statistics for variables at baseline, frequencies (proportions) for categorical variables, and mean (SD) for continuous variables (unstandardized).

| **Variables** | **Total** | **Non-Event** | **Event** |
|---|---|---|---|
| Number of patients | 220 | 147 (66.80%) | 73 (33.20%) |
| Gender: Male | 132 (60.00%) | 87 (39.54%) | 45 (20.45%) |
| Gender: Female | 88 (40.00%) | 60 (27.27%) | 28 (12.72%) |
| Marital: Single | 72 (32.72%) | 52 (23.63%) | 20 (9.09%) |
| Marital: Married | 148 (67.27%) | 95 (43.18%) | 50 (24.09%) |
| Residence: Urban | 143 (65.00%) | 95 (43.18%) | 48 (21.81%) |
| Residence: Rural | 77 (35.00%) | 52 (23.63%) | 25 (11.36%) |
| Type: Others | 76 (35.45%) | 49 (22.27%) | 27 (12.27%) |
| Type: Metformin | 144 (65.45%) | 98 (44.54%) | 46 (20.90%) |
| Family history: Yes | 99 (45.00%) | 67 (30.45%) | 32 (14.54%) |
| Family history: No | 121 (55.00%) | 80 (36.36%) | 41 (18.63%) |
| Age | 53.75 (11.09) | 52.31 (11.08) | 56.61 (10.61) |
| BMI | 22.46 (3.78) | 20.64 (2.85) | 26.11 (2.67) |
| HDL | 45.36 (15.25) | 49.81 (15.63) | 36.39 (9.50) |
| LDL | 113.12 (37.94) | 97.12 (32.04) | 145.00 (26.91) |
| Triglycerides | 191.15 (59.53) | 166.13 (51.60) | 241.52 (39.10) |
| Total cholesterol | 190.27 (63.38) | 165.76 (57.48) | 239.64 (42.67) |
| FBS | 200.90 (62.27) | 205.14 (66.33) | 192.35 (79.15) |
| DBP | 92.00 (5.40) | 91.00 (5.45) | 93.43 (7.15) |
| SBP | 142.15 (14.21) | 146.19 (13.19) | 133.91 (12.58) |

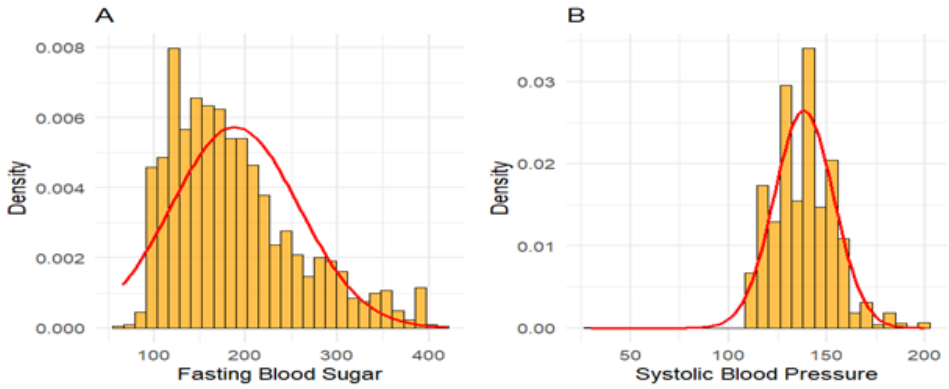



Figure 2.1: Histogram of glucose concentration and blood pressure

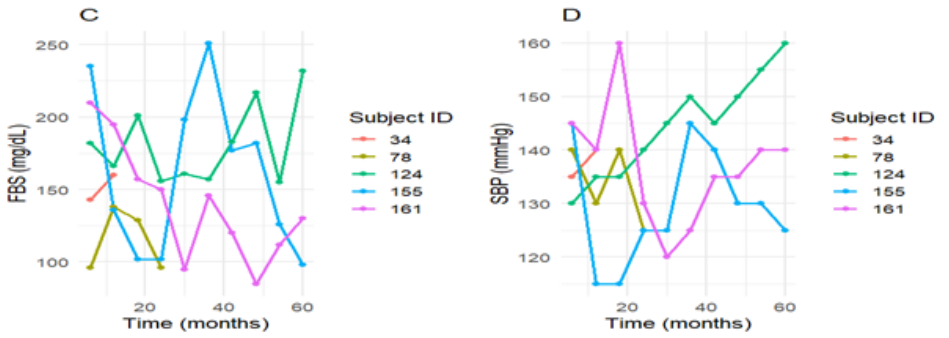



Figure 2.2: Trajectory of glucose concentration and blood pressure for randomly selected subjects.

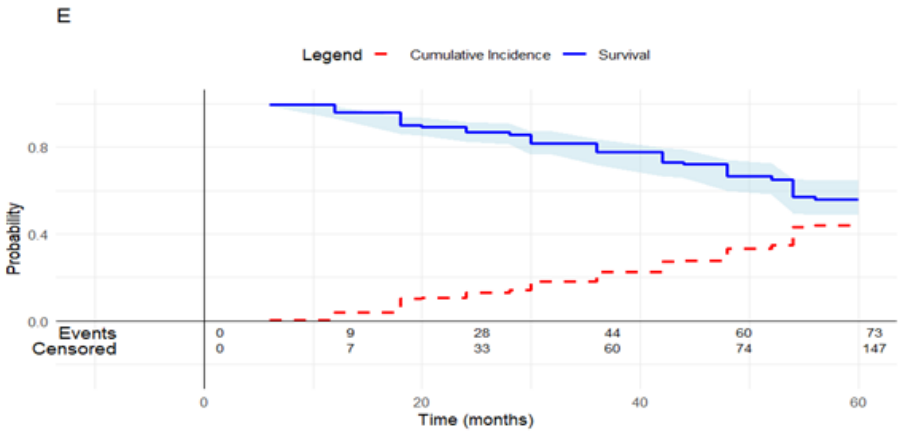



Figure 2.3: Kaplan–Meier and cumulative incidence curve

### 2.3.6 Implementation of the model

We present a bivariate joint modeling approach to analyze a diabetes and hypertension dataset, incorporating longitudinal fasting blood sugar $Y_{ij1}$ and systolic blood pressure,

$Y_{ij2}$ considered as correlated bivariatelongitdunal outcomes, where $i = 1, 2, \ldots, 220$ subjects and $j = 1, 2, \ldots, 10$ with the time to chronic complications as the event outcome. The bivariate linear mixed-effect model incorporates covariates such as measurement time (time), residence (0 = urban, 1 = rural), and sex (0 = male, 1 = female) for both fasting blood sugar and systolic blood pressure. According to model equation 2.3, the bivariate linear mixed-effects model is specified as follows:

$$Y_{ijk} = (\beta_{1k} + b_{ik1}) + (\beta_{2k} + b_{ik2})\,\text{Time}_{ij} + \beta_{3k}\,\text{age}_i + \beta_{4k}\,\text{sex}_i + \beta_{5k}\,\text{residence}_i + e_{ijk}, \quad (2.15)$$

Here, $k = 1, 2$ represents the two longitudinal outcomes: glucose concentration in mg/dL and systolic blood pressure in mmHg, respectively, for the $i^{\text{th}}$ subject measured at time $t_{ij}$ (expressed in months), where $i = 1, 2, \ldots, 220$, and $j = 0, 1, 2, \ldots, 10$. The fixed-effects parameter vector is given by $\beta^\top = (\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}, \beta_{51}, \beta_{12}, \beta_{22}, \beta_{32}, \beta_{42}, \beta_{52})^\top$, where the elements correspond to the intercept, the effects of time, age, sex, and residence on glucose concentration ($k = 1$) and systolic blood pressure ($k = 2$), respectively. The random-effects vector is defined as $b_i = (b_{i11}, b_{i21}, b_{i12}, b_{i22})$, represents the random intercepts and slopes for glucose concentration and blood pressure. $e_{ijk}$ denotes the random error, with $e_{ij} = (e_{ij1}, e_{ij2})^\top$ being a vector of within-subject residuals of dimension $n_i \times 1$, corresponding to glucose concentration and systolic blood pressure, respectively.

Micro- and macrovascular complications are the events of interest in the survival submodel. A study subject may either be under follow-up at the end of the study period or drop out during the study period, which is considered right-censoring ($C_i$). On the other hand, during the study period, the subject may develop chronic complications at any time ($T_i$). Then, for each study subject, the observed event time ($T_i$) would be a $\min(T_i, C_i)$. This observed time may depend on baseline age, body mass index, HDL, LDL, and triglyceride. The classical survival model can include all baseline measurements as covariates, but it cannot include longitudinal measurements. As a result, we needed to use a joint model that incorporates subject-specific longitudinal processes as time-varying covariates in the survival model. Despite the various parameterization approaches

available to link the longitudinal measurement to the time-to-event data (Andrinopoulou et al., 2017; Hickey et al., 2018; Mwanyekange et al., 2018), the study employed a shared random effect (Wulfsohn and Tsiatis, 1997). This shared random effect method links the longitudinal submodel, specifically the subject-specific intercept and the effects of fasting blood sugar and systolic blood pressure, to the survival submodel, which models the time until micro- and macro-vascular complications occur. The study included baseline age, body mass index, HDL, LDL, and triglyceride as predictors of the time to chronic complications. As shown from equation 2.8 of the Cox regression model, we can formulate the model as follows:

$$\begin{aligned}\lambda(t_i \mid X_i, b_i) = \lambda_0(t)\, \exp\Big(&\gamma_1 b_{i11} + \gamma_2 b_{i21} + \gamma_3 b_{i12} + \gamma_4 b_{i22} \\ &+ \alpha_1\, \text{age}_i + \alpha_2\, \text{BMI}_i + \alpha_3\, \text{HDL}_i \\ &+ \alpha_4\, \text{LDL}_i + \alpha_5\, \text{Triglyceride}_i\Big). \end{aligned} \tag{2.16}$$

where $\lambda(t_i \mid X_i, b_i)$ is the hazard rate or instantaneous rate at time $t$, given covariates $X_i$ and $b_i$, $\lambda_0(t)$ is the baseline hazard rate when all covariates are zero, $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ represents the effects of the random intercepts and random slopes of fasting blood glucose and systolic blood pressure for the time to onset of micro- and macro-vascular complications, and $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ are the effects of age, BMI, HDL, LDL, and triglycerides, respectively, the time to the onset of micro- and macro-vascular complications.

As shown in Figure 2.1, the longitudinal outcome variables, fasting blood sugar and systolic blood pressure, follow a multivariate skew-normal distribution. Thus, assuming a multivariate normal distribution of the random errors may lead to incorrect parameter estimates (Chen and Luo, 2018) and (Huang et al., 2022). To obtain relatively unbiased parameter estimates, we used a multivariate skew-normal distribution, an extension of the multivariate normal distribution that accommodates skewness. We explore how skewness in the distribution of model errors affects parameter-estimation precision by comparing two models.

**Model Normal:** Bivariate joint model with multivariate normal distribution for the random error.

**Model Skew-Normal:** Bivariate joint model with multivariate skew-normal distribution for the random error.

To carry out the Bayesian inference process, we must assign specific values to the hyper-parameters in the prior distribution as specified in the equation in 2.13. Vague priors with large variances were specified for all parameters in both the longitudinal and survival models. Specifically, each component of the fixed effects vectors $\beta, \gamma$ and $\alpha$ as well as the skewness parameter $\Delta_{ek}$ were assumed to follow independent normal distributions with mean zero and variance 100, i.e. $N(0, 100)$. The prior for the variance–covariance matrices of the random effects, $\Sigma_b$, and random errors, $\Sigma_k$, is assumed to follow an inverse Wishart distribution with $\Omega_{\Delta_{bk}} = \text{diag}(0.01, 0.01, 0.01, 0.01)$, degrees of freedom $= 5$, and $\Omega_{\Delta_{ek}} = \text{diag}(0.01, 0.01)$, degrees of freedom $= 3$, respectively.

The MCMC procedure was carried out in R, interfacing with JAGS via the R2jags package. It is essential to verify the algorithm's convergence to ensure accurate inference of the population parameter before interpreting the model parameter. The convergence of the samples generated from the MCMC procedure applied to the diabetes and hypertension data was assessed through graphical diagnostics, such as density and trace plots, as well as formal statistical tests, using the Gelman-Rubin statistic (Gelman and Rubin, 1992). For each model, we ran three MCMC sampling chains, each consisting of 40,000 iterations, with initial values set to the default (a randomly generated value). We discarded the first 10,000 iterations as a burn-in period and retained every 20th sample thereafter. Thus, we obtain 4500 samples from the targeted posterior distributions and use them to make statistical inferences. Figure 2.4 displays trace plots for some selected regression coefficients. The trace plots indicate that the lines from the three distinct chains converge or intersect, demonstrating that the algorithm has achieved convergence. It implies that the regression coefficients for the selected parameters converge to their target distributions.

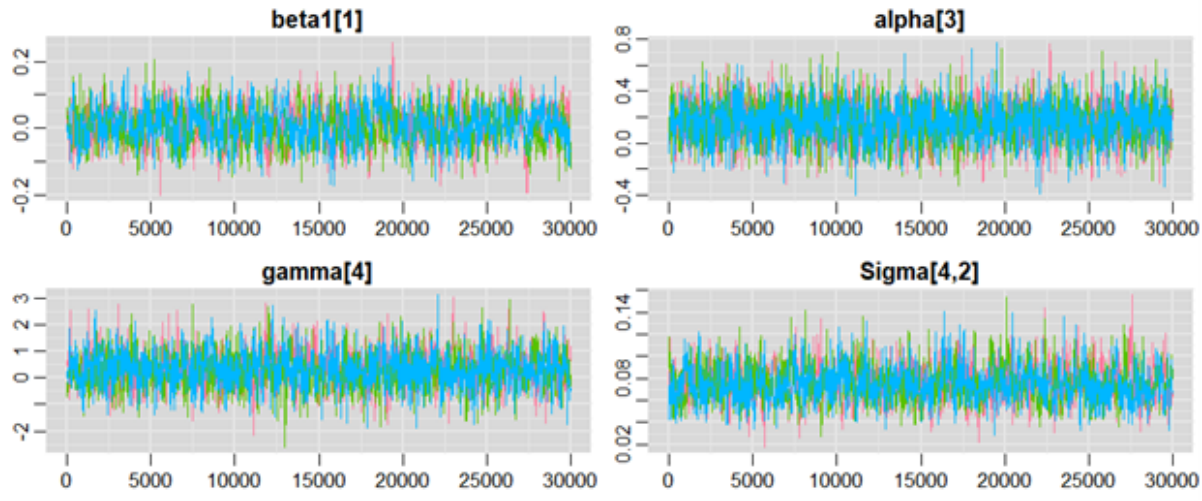



Figure 2.4: Trace plots

Additionally, the posterior density plots in Figure 2.5 provide valuable insights. The density plots from each chain closely overlap and resemble the density estimates from the other chains, indicating that all chains are sampling from the same posterior distribution.

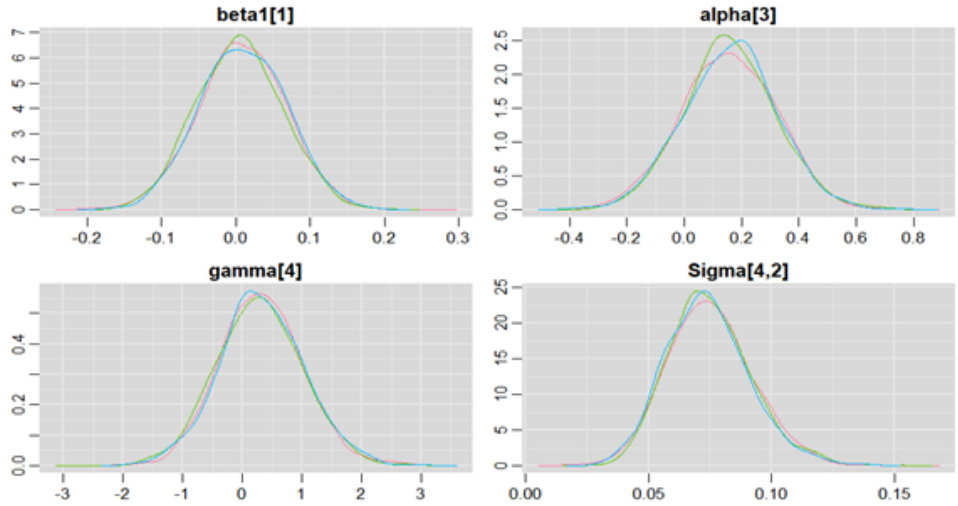



Figure 2.5: Density plots

In addition to visual inspection, it is essential to use formal statistical tests to assess the convergence of the samples generated by the MCMC procedure. In this study, we used the Gelman-Rubin statistic to evaluate whether multiple chains have converged to the same target distribution, which is essential for ensuring reliable inference. Table 2.2 displays the R hat values for each parameter. As the chains converge on the target distributions, the R hat value approaches 1, indicating that the chains have converged and are sampling

from the target distributions. Conversely, values larger than 1 indicate non-convergence. Although it's well known that values below approximately 1.05 are generally considered acceptable, it's essential to recognize this as a broad guideline. As shown in Table 2.2, the R-hat values of all parameters are below 1.05, indicating that the posterior distributions of the parameters have converged to the target distribution.

Table 2.2: Results of the Gelman–Rubin (R-hat) test of convergence.

| Parameter | $\beta_{11}$ | $\beta_{21}$ | $\beta_{31}$ | $\beta_{41}$ | $\beta_{51}$ | $\beta_{12}$ | $\beta_{22}$ |
|---|---|---|---|---|---|---|---|
| R-hat | 1.035 | 1.031 | 1.017 | 1.005 | 1.007 | 1.009 | 1.035 |
| Parameter | $\beta_{32}$ | $\beta_{42}$ | $\beta_{52}$ | $\beta_{52}$ | $\alpha 1$ | $\alpha 2$ | $\alpha 3$ |
| R-hat | 1.009 | 1.080 | 1.006 | 1.002 | 1.001 | 1.0021 | 1.0061 |
| Parameter | $\alpha 4$ | $\alpha 5$ | $\gamma 1$ | $\gamma 2$ | $\gamma 3$ | $\gamma 4$ | |
| R-hat | 1.002 | 1.002 | 1.016 | 1.011 | 1.001 | 1.003 | |
| Parameter | $v_{11}$ | $v_{22}$ | $v_{33}$ | $v_{44}$ | $v_{12}$ | $v_{13}$ | $v_{14}$ |
| R-hat | 1.003 | 1.004 | 1.002 | 1.003 | 1.004 | 1.010 | 1.003 |
| Parameter | $v_{32}$ | $v_{34}$ | $\delta_1$ | $\delta_2$ | $\sigma^2_{11}$ | $\sigma^2_{12}$ | $\sigma^2_{22}$ |
| R-hat | 1.005 | 1.005 | 1.002 | 1.002 | 1.004 | 1.003 | 1.003 |

## 2.4 Results and Discussion

### 2.4.1 Results and Model Comparison

Table 2.4 reports the posterior means, standard deviations (SD), and 95% credible intervals for all parameters from the multivariate linear mixed-effects model and the Cox proportional hazards model, accounting for the distribution of the model error. To improve the efficiency of the estimation algorithm and reduce variability from measurement errors, we standardized the longitudinal data for fasting blood sugar and systolic blood pressure. Additionally, to avoid unstable estimates, the quantitative covariates were standardized. Therefore, the results are expressed in standard units rather than the original

units of measurement.

Firstly, in the multivariate linear mixed effect sub-model, the values of the parameters ($\beta$) in the under model normal are slightly higher than those in the Model Skew-Normal. There are some differences in the magnitude of the coefficient ($\alpha$) in the survival sub-model between the two models. At the same time, there is a significant difference in the estimates of the association parameter ($\gamma$). This is especially clear for the random intercept and slope coefficients for fast blood glucose. These results indicate significant differences in estimates of the association parameter for longitudinal and time-to-event data under different distributional assumptions for the model errors.

To select the model that best fits the data, the Bayesian selection criterion, the Deviance Information Criterion (DIC), as proposed by (Spiegelhalter et al., 2002), is applied. As with other model selection criteria, it is essential to note that DIC is not intended to identify the "correct" model, but rather to facilitate comparisons between models that fit the data reasonably well. Table 2.4 provides a summary of the DIC values for model Normal and model skew normal. The skewed normal model has a smaller DIC value than the normal model, indicating a better fit of the data. Therefore, the results from the selected joint model, the skewed normal model, will be used for further interpretation and discussion.

The model-skewed normal in Table 2.4 estimates that the skewness parameters for fasting blood sugar (1.282) and systolic blood pressure (0.422) are significantly positive. This implies that the distribution of fasting blood glucose and systolic blood pressure data is significantly positively skewed. Thus, it may suggest that accounting for the longitudinal asymmetry in the joint model provides a better fit to the data than the classical multivariate normal distribution. Based on the model-skewed normal distribution, the estimated results of the multivariate linear mixed-effect submodel in Table 2.4 show that measurement time and age are positively associated with the trajectories of fasting blood sugar and systolic blood pressure. The progression of fasting blood glucose and blood pressure levels in people with type 2 diabetes and hypertension living in rural areas is

higher than in those living in urban areas. Moreover, female diabetic and hypertensive patients exhibited higher fasting blood sugar levels compared to their male counterparts.

In the multivariate linear mixed-effects model, the covariance between the random effects and random intercept reflects the variability in both the random effects and intercept, as well as the relationship between the longitudinal trajectories of fasting blood glucose and systolic blood pressure, as shown in Table 2.3. The association values for the random intercept and slope of FBS and SBP indicate a positive correlation, with values of 0.075 and 0.040, respectively, in individuals with T2D and hypertension. A multivariate linear mixed-effects model shows that the progression of fasting blood glucose is associated with systolic blood pressure, with specific coefficients of 0.075 and 0.040. This positive relationship implies a strong link between changes in fasting blood glucose and systolic blood pressure, as the 95% credible interval does not include zero. Keep in mind that the magnitude does not reflect the strength of the relationship, since both longitudinal measurements were standardized.

Table 2.3: Summary of the estimated posterior mean (PM) of variance–covariance matrix parameters for random errors and random effects, standard deviation (SD), and 95% credible intervals (CI) for each model

| **Parameter** | **Normal Model** | | | **Skew-Normal Model** | | |
|---|---|---|---|---|---|---|
| | **PM** | **SD** | **95% CI** | **PM** | **SD** | **95% CI** |
| $\sigma^2_{11}$ | 0.603 | 0.023 | 0.559, 0.649 | 0.340 | 0.015 | 0.086, 0.143 |
| $\sigma^2_{22}$ | 0.652 | 0.025 | 0.606, 0.702 | 0.592 | 0.027 | 0.549, 0.653 |
| $\sigma^2_{12}$ | 0.206 | 0.017 | 0.173, 0.241 | 0.201 | 0.024 | 0.001, 0.094 |
| $v_{11}$ | 0.288 | 0.040 | 0.218, 0.372 | 0.170 | 0.030 | 0.174, 0.289 |
| $v_{22}$ | 0.144 | 0.023 | 0.104, 0.194 | 0.077 | 0.016 | 0.074, 0.137 |
| $v_{33}$ | 0.239 | 0.022 | 0.067, 0.122 | 0.219 | 0.017 | 0.063, 0.129 |
| $v_{44}$ | 0.126 | 0.036 | 0.178, 0.315 | 0.110 | 0.033 | 0.172, 0.303 |
| $v_{12}$ | 0.108 | 0.022 | 0.093, 0.155 | 0.071 | 0.013 | 0.061, 0.099 |
| $v_{13}$ | 0.121 | 0.020 | 0.091, 0.169 | 0.075 | 0.019 | 0.084, 0.159 |
| $v_{14}$ | 0.107 | 0.022 | 0.091, 0.152 | 0.079 | 0.015 | 0.068, 0.111 |
| $v_{23}$ | 0.058 | 0.021 | 0.044, 0.101 | 0.035 | 0.014 | 0.026, 0.064 |
| $v_{34}$ | 0.073 | 0.019 | 0.059, 0.112 | 0.060 | 0.018 | 0.048, 0.097 |
| $v_{24}$ | 0.075 | 0.019 | 0.037, 0.086 | 0.040 | 0.019 | 0.034, 0.109 |

The estimated association parameters alpha and gamma in the results of the survival submodel indicate how longitudinal measurements, along with clinical and biochemical

variables, influence the time to the development of chronic complications. The effect of blood pressure on the timing of these complications is estimated by $\gamma_3$ and $\gamma_4$, which reflect a positive relationship between the rate of change in systolic blood pressure and the risk of microvascular and macrovascular complications. When considering the hazard ratio, a 1-standardized-unit increase in the SBP trajectory, or a 1-unit rise in blood pressure over 6 months (on a standard scale), corresponds to a 1.682-fold increase ($\exp(0.520) = 1.682$) in the risk of microvascular and macrovascular complications. This effect size can also be interpreted as a 1-unit increase in trajectory systolic blood pressure being associated with a 66% higher risk of these complications.

We also found that baseline body mass index, age, and HDL are positively associated with the risk of microvascular and macrovascular complications. However, random intercepts and slopes for fasting blood sugar, triglycerides, and LDL are negatively related to the risk of chronic complications. Figure 2.6 illustrates the hazard ratio (HR) in the survival sub-model, which is based on the skew normal model, longitudinal effects of association parameters, and baseline covariates related to the risk of chronic complications.

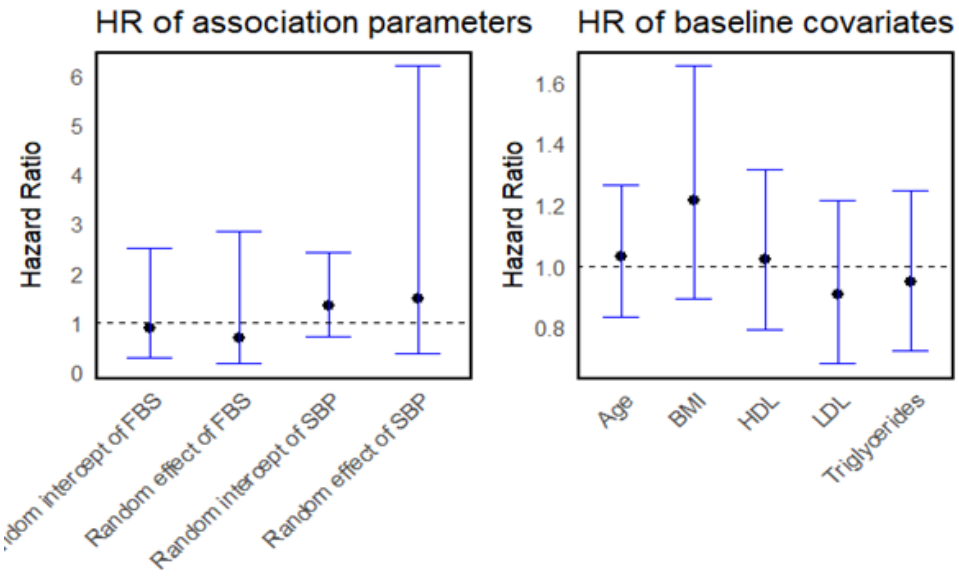



Figure 2.6: Hazard ratio of survival sub-model based on model skew normal; longitudinal effects of association parameter and baseline covariates in risk of chronic complications

Table 2.4: Posterior mean, standard deviation (SD), and 95% credible intervals (CI) for parameters under model normal and model skew-normal models.

| **Parameter** | **Model Noraml** | | | **Model skew normal** | | |
|---|---|---|---|---|---|---|
| | Posterior mean | SD | 95% CI | Posterior mean | SD | 95% CI |
| **Bivariate linear mixed-effects model parameters estimates** | | | | | | |
| $\beta_{11}$ | 0.008 | 0.058 | (-0.104, 0.121) | 0.022 | 0.046 | (-0.073, 0.112) |
| $\beta_{21}$ | 0.071 | 0.033 | (0.006, 0.138) | 0.041 | 0.027 | (-0.012, 0.095) |
| $\beta_{31}$ | 0.045 | 0.037 | (-0.026, 0.117) | 0.028 | 0.031 | (-0.021, 0.116) |
| $\beta_{41}$ | -0.078 | 0.075 | (-0.224, 0.072) | 0.011 | 0.056 | (-0.100, 0.116) |
| $\beta_{51}$ | 0.120 | 0.079 | (-0.032, 0.275) | 0.018 | 0.064 | (-0.062, 0.131) |
| $\beta_{12}$ | 0.030 | 0.057 | (-0.083, 0.145) | 0.028 | 0.056 | (-0.095, 0.121) |
| $\beta_{22}$ | 0.067 | 0.032 | (0.006, 0.131) | 0.055 | 0.031 | (0.003, 0.125) |
| $\beta_{32}$ | 0.112 | 0.037 | (0.043, 0.185) | 0.107 | 0.035 | (0.042, 0.182) |
| $\beta_{42}$ | -0.078 | 0.076 | (-0.226, 0.066) | -0.042 | 0.076 | (-0.197, 0.095) |
| $\beta_{52}$ | 0.045 | 0.077 | (-0.107, 0.191) | 0.012 | 0.076 | (-0.115, 0.176) |
| $\delta_1$ | | | | 1.282 | 0.003 | (0.995, 1.000) |
| $\delta_2$ | | | | 0.422 | 0.062 | (0.207, 0.445) |
| **Survival model parameters estimates** | | | | | | |
| $\gamma_1$ | 0.239 | 0.425 | (-0.603, 1.083) | -0.620 | 0.540 | (-1.217, 0.912) |
| $\gamma_2$ | -0.510 | 0.609 | (-1.781, 0.622) | -0.027 | 0.740 | (-1.834, 1.047) |
| $\gamma_3$ | 0.283 | 0.320 | (-0.362, 0.917) | 0.234 | 0.314 | (-0.348, 0.886) |
| $\gamma_4$ | 0.312 | 0.709 | (-1.054, 1.731) | 0.520 | 0.717 | (-0.978, 1.882) |
| $\alpha_1$ | 0.038 | 0.109 | (-0.178, 0.246) | 0.026 | 0.107 | (-0.181, 0.236) |
| $\alpha_2$ | 0.159 | 0.162 | (-0.162, 0.474) | 0.220 | 0.161 | (-0.112, 0.502) |
| $\alpha_3$ | 0.042 | 0.130 | (-0.215, 0.299) | 0.007 | 0.129 | (-0.232, 0.272) |
| $\alpha_4$ | -0.121 | 0.146 | (-0.402, 0.171) | -0.081 | 0.147 | (-0.383, 0.193) |
| $\alpha_5$ | -0.066 | 0.140 | (-0.347, 0.212) | -0.029 | 0.139 | (-0.145, 0.100) |
| DIC | | 9105.747 | | | 3738.161 | |

## 2.5 Simulation Studies

We conduct the following simulation studies to compare the proposed multivariate joint model across different distributions of random errors and to examine how deviations from longitudinal biomarker symmetry influence modeling results. The design of the simulated dataset is similar to the diabetes and hypertension data used in Section 2.3.5. We generate two longitudinal outcomes, one time-to-event outcome, and covariates as follows. The longitudinal measurement time points are twice a year, with a maximum follow-up of 5 years, and the visit time intervals are identical for all subjects. The bivariate, correlated longitudinal outcomes were generated using the multivariate linear mixed-effects submodel described in equation 2.3. We set the true values of the model parameters to those obtained from real data analysis of the model skew normal, as presented in Table 2.4 in the following way. $\beta_1 = (\beta_{11}, \beta_{21}, \beta_{31}, \beta_{41}, \beta_{51}) = (0.022, 0.041, 0.028, 0.011, 0.018)$, $\beta_2 = (\beta_{12}, \beta_{22}, \beta_{32}, \beta_{42}, \beta_{52}) = (0.028, 0.055, 0.107, -0.042, 0.0105)$, $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4) = (-0.620, -0.020, 0.234, 0.520)$, and $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.026, 0.220, 0.007, -0.081, -0.029)$. Random intercept random effects vector ($b_i = (b_{11}, b_{21}, b_{12}, b_{22})$) was simulated from a multivariate normal distribution with mean vector of zero and variance covariance matrix $\boldsymbol{\Sigma}_b$ where their values are $\text{diag}(1, 1, 1, 1)$. To introduce skewness into the longitudinal data, random errors were generated using a gamma distribution. Specifically, the random errors for fasting blood sugar and systolic blood pressure were generated using a gamma distribution with both the shape and scale parameters set to 1. Similar to the real data analyses, the time-to-event data were simulated using the proportional hazard model described in equation 2.8. For each subject, survival time was simulated from an exponential distribution with a constant baseline hazard rate of 0.1. An exponential distribution with a mean equal to 0.1 is used to generate the censoring time.Thus, the observed survival time of the $i^{\text{th}}$ subject was computed as $T_i = \min(T_i^*, C_i)$. The event

status for the $i^{\text{th}}$ subject was denoted by

$$\tau_i = \begin{cases} 1, & \text{if } T_i \leq C_i, \\ 0, & \text{otherwise.} \end{cases}$$

The covariates included in both sub-models are mimicked from the real dataset and simulated based on variable types. For example, the qualitative covariates residence (urban as the reference) and sex (male as the reference) were simulated from Bernoulli distributions with success probabilities of 0.32 and 0.40, respectively. The continuous covariates were simulated from a normal distribution with mean and standard deviation chosen according to the application dataset.

Due to the extensive computation, we simulate only 50 data sets, each containing 150 subjects, to make the study feasible. However, it still takes about 1 hour for each replicate. Then each dataset is fitted with the model normal and skew-normal models to determine which model fits the data perfectly and to identify which model provides a relatively unbiased parameter estimate. The Bayesian inference in Section 2.3.4 is used to estimate the parameters of each model, and the specification of prior distributions and the convergence diagnostics tools were similar to the model implementation in Section 2.3.6. All the MCMC samplers were implemented using JAGS and the R package R2jags. To compare the performance of parameter estimates under the two models, we computed biases, bias,$B(\theta) = \frac{1}{M}\sum_{i=1}^{M}\hat{\theta}_i - \theta_T$, RMSE$(\theta) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\hat{\theta}_i - \theta_T\right)^2}$, and the deviance information criterion (DIC), where $\hat{\theta}_i$ is the estimate of the true parameter $\theta_T$ from the $i^{\text{th}}$ simulat and $\theta_T$ is the true parameter value, which are obtained from the application dataset and m is the total number of replications. Table 2.5 summarizes the simulation results for the two models. It includes the true parameter (TP) values, bias, and root-mean-square error for fixed-effect parameters. A positive bias value indicates that the estimated value overestimates the true parameter value, while a negative value indicates an underestimate. For instance, the estimated biases for $\beta_{42}$ and $\beta_{52}$ are positive, indicating that these parameters overestimate the effects of sex and residence on systolic blood

pressure. On the other hand, the smallest values of bias and mean squared error indicate that the model provides a better estimate than the others.

For most parameters, particularly $\alpha$, the bias and RMSE in the joint model with a skewed normal error distribution were smaller than those in the corresponding model assuming normality. Model comparison using the DIC further confirmed that the multivariate joint model with a skewed multivariate normal distribution yielded a lower DIC, indicating a better fit than the standard multivariate normal approach. Thus, we conclude that the skew-normal model outperforms the standard model, producing estimates with lower bias and mean squared error. Overall, the simulation studies indicate that assuming a symmetric normal distribution can lead to inaccurate and inefficient inferences when data exhibit non-normal features.

Table 2.5: Summary of true parameter (TP) values, bias, and RMSE for the model normal and the model skew-normal model.

| **Parameter** | **TP** | **model normal** | | **model Skew-normal** | |
|---|---|---|---|---|---|
| | | Bias | RMSE | Bias | RMSE |
| $\beta_{11}$ | 0.022 | -1.490 | 1.494 | -1.490 | 1.494 |
| $\beta_{21}$ | 0.041 | -0.006 | 0.087 | -0.001 | 0.086 |
| $\beta_{31}$ | 0.028 | -0.020 | 0.089 | -0.014 | 0.087 |
| $\beta_{41}$ | 0.011 | -0.031 | 0.180 | -0.039 | 0.186 |
| $\beta_{51}$ | 0.018 | -0.030 | 0.214 | -0.020 | 0.218 |
| $\beta_{12}$ | 0.028 | -1.242 | 1.429 | -1.286 | 1.293 |
| $\beta_{22}$ | 0.055 | -0.009 | 0.094 | -0.001 | 0.104 |
| $\beta_{32}$ | 0.107 | -0.023 | 0.093 | -0.026 | 0.100 |
| $\beta_{42}$ | -0.042 | 0.043 | 0.195 | 0.021 | 0.197 |
| $\beta_{52}$ | 0.012 | 0.024 | 0.236 | 0.020 | 0.236 |
| $\gamma_1$ | -0.620 | 0.001 | 0.298 | 0.000 | 0.296 |
| $\gamma_2$ | -0.027 | -0.039 | 0.257 | -0.034 | 0.265 |
| $\gamma_3$ | 0.234 | 0.066 | 0.342 | 0.012 | 0.252 |
| $\gamma_4$ | 0.520 | -0.831 | 0.820 | -0.068 | 0.314 |
| $\alpha_1$ | 0.026 | 0.070 | 0.274 | 0.069 | 0.276 |
| $\alpha_2$ | 0.220 | 0.062 | 0.282 | 0.062 | 0.273 |
| $\alpha_3$ | 0.007 | 0.033 | 0.275 | 0.030 | 0.266 |
| $\alpha_4$ | -0.081 | -0.046 | 0.224 | -0.038 | 0.224 |
| $\alpha_5$ | -0.029 | 0.015 | 0.234 | 0.014 | 0.232 |
| DIC | 7369 | | | 6600 | |

## 2.6 Discussion

As more studies are conducted, it is crucial to take measures over time to evaluate a patient's health status in the context of specific events using a joint model approach. However, most existing joint modelling analyses typically focus on a univariate frame-

work, involving a single longitudinal outcome and a single time-to-event, and assume that the model errors follow a normal distribution. Motivated by hypertension and diabetes disease, this study extends the univariate joint models to multivariate joint modelling with multivariate skew normal distribution to take into consideration skewness in the longitudinal outcomes' glucose concentration and blood pressure, and piecewise constant functions to flexible model the hazard functions. The two sub-models were connected through the sharing of random effects to assess the relationship between longitudinal and time-to-event processes. We applied the Bayesian estimation technique to estimate all parameters in the joint models simultaneously.

Our results demonstrated the use of multivariate joint modelling to examine how patterns of glucose concentration and blood pressure trajectories were associated with the risk of chronic disease complications. We also considered standard but essential data features that may affect the discovery of the true disease progression, including correlations between longitudinal measurements and their distributions.

Two types of multivariate joint models were used based on the distributional assumptions of the longitudinal outcomes: one assuming a multivariate normal distribution and the other a multivariate skewed normal distribution for the random errors. In the model skew normal, the estimates of the skewness parameters $\delta_1$ and $\delta_2$ are positive, implying that the distributions of fast blood glucose concentration and blood pressure are not multivariate normal. The DIC value is also used to compare the models, revealing that the multivariate joint model with a skewed multivariate normal distribution provided a better fit than the standard multivariate normal approach. Therefore, a multivariate joint model with a multivariate skew normal distribution offers more efficient and relatively accurate parameter estimation, making it a better alternative to the standard (symmetric) distribution-based model, which is widely assumed in statistical research. The simulation study also confirmed the application data set scenario: a bivariate joint model with a multivariate skewed normal distribution of the random errors performs better than a bivariate joint model with a multivariate normal distribution. Although the dataset and

methods differ (Xu et al., 2021; Huang et al., 2022; Baghfalaki et al., 2013), they further supported these findings by demonstrating that incorporating a skewness parameter or using an asymmetric multivariate normal distribution in the longitudinal sub-model produced better results than the classical multivariate normal distribution, which aligns with my conclusions.

We then used a multivariate joint model with skewed longitudinal outcomes to examine how patterns in glucose concentration and systolic blood pressure are associated with the risk of microvascular and macrovascular complications. The model not only captured the association between glucose concentration and systolic blood pressure with the risk of micro and macrovascular complications but also identified predictors of glucose concentration, blood pressure, and chronic complications.

When we consider the clinical implications, the overall incidence of microvascular complications in people with T2D and hypertension was 33.20% over the five-year study period. These findings are similar to previous research in Bahir Dar, Ethiopia (32.4% over seven years); (Shita and Muluneh, 2021), Gondar, Ethiopia (28% during a follow-up of 6 years) (Wolde et al., 2018); and Wollega, Ethiopia (31.2%)(Korsa et al., 2019). There is a strong association between changes in glucose levels and blood pressure in individuals with type 2 diabetes (T2D) and hypertension. Elevated trends in glucose and blood pressure each carry significant risks for long-term complications. When these two patterns are linked, the overall risk rises considerably (Viazzi et al., 2019) and (Wan et al., 2020). Therefore, this evidence should guide health professionals and clinical decision-makers to effectively manage both blood pressure and glucose levels simultaneously.

Among the covariates we included in the proportional hazards model, age, BMI, and HDL are directly associated with the risk of microvascular and macrovascular complications. However, triglycerides and LDL show an inverse relationship with the risk of these complications, but their effects are not significant. Despite using different statistical methods and different datasets, studies (Jelinek et al., 2017) and (Liu et al., 2022) report similar findings: age and lipid levels (triglycerides, LDL, and HDL) are significantly associated

with chronic complications such as diabetic retinopathy and diabetic kidney disease. Our findings suggest that normalizing body mass index and HDL levels could be an effective strategy for preventing micro- and macrovascular complications.

Our results have clinical implications for interpreting the trajectory of blood pressure and blood glucose in relation to the risk of micro- and macrovascular complications in individuals T2D and hypertension. Results from the joint modeling approach suggest that, as blood pressure patterns increase over time, individuals with T2D and hypertension face a higher risk of developing retinopathy, chronic kidney disease, heart disease, or stroke. Additionally, the timing of these complications is indirectly related to the fasting blood sugar trajectory, which differs from previous expectations; therefore, further research may be necessary. The relationships among these trends were less apparent when using other classical regression models, such as survival, linear, nonlinear, and logistic regression. This evidence highlights a unique advantage of this modeling framework: its ability to provide an overall view of follow-up measurements for one or more biomarkers in relation to the time until the onset of micro- and macrovascular complications.

For SBP, our findings are similar to those seen in previous studies. Li et al. (2023) noted a 1.70-fold increased risk of heart failure for each one standardized unit increase in SBP, which is similar to the 1.682-fold rise in the risk of micro- and macrovascular complications found here. Ceriello et al. (2023) also observed that patients with high variability in both body weight and blood pressure had the highest risk of cardiovascular disease (HR = 1.81; 95% confidence interval = 1.61-2.05) compared with patients with low variability in both body weight and total cholesterol. Dorajoo et al. (2017) investigates the association between variability in HbA1c or systolic blood pressure (SBP) and retinopathy in Asians with T2D using a multivariate regression model; reported SBP was significantly associated with diabetic retinopathy (odds ratio 1.03, 95% confidence interval 1.01-1.05). From a clinical perspective, these estimates were broadly in line, suggesting that optimizing blood pressure control in people with diabetes and hypertension may reduce the risk of micro- and macrovascular complications. Maintaining optimal systolic blood pressure

is crucial for reducing the risk of chronic complications in the management of blood sugar and blood pressure.

The methodology presented in this chapter has potential for future research extensions. This chapter focuses on a single-type survival event only. In cases involving multiple event types, the proposed bivariate joint model can be extended to handle competing-risks survival data (Ferede et al., 2022; Li et al., 2023; Hu et al., 2009). For the longitudinal outcomes, FBS and SBP, we used a fully parametric (linear) mixed-effects submodel. However, in some applications, the true relationship between the longitudinal outcomes and time effects may be nonlinear. For example, in our data involving individuals with T2D and hypertension, the trajectories of FBS and SBP over time are nonlinear. Consequently, a fully parametric modelling approach may lack the flexibility needed to capture these complex longitudinal patterns accurately. Therefore, our approach can be expanded to a more flexible semi-parametric multivariate mixed-effect model by including non-parametric smoothing functions of time and addressing competing risk survival times, which could be a focus of future work.

## 2.7 Conclusion

We use a multivariate joint modelling approach with shared random intercepts and slopes to analyse multiple longitudinal markers with skewed distributions and time-to-event outcomes simultaneously. Bayesian inference using R and JAGS was used to estimate the model parameters of joint models for multiple longitudinal processes and a time-to-event outcome. We illustrate the method using electronic medical record data from a primary care patient cohort to investigate the associations between glucose concentration, blood pressure, and the risk of micro- and macrovascular complications. Our results demonstrated that both glucose concentration and blood pressure measurements are severely skewed; therefore, the multivariate linear mixed-effects model with skew-normal random errors yielded relatively unbiased parameter estimates compared to the classical assumption of multivariate normality.

We conducted simulation studies to evaluate the performance of the proposed model, and the results indicate that the estimated parameter under the skew multivariate normal distribution is a relatively unbiased estimate compared to symmetric distributions. The joint model framework provides a platform for exploring various features of longitudinal measures of disease risk, extending beyond the traditional approach that relies solely on baseline measures in cohort studies. Although the motivation for this study arose from a diabetes and hypertension data study, the methodology for joint models of multiple longitudinal processes and time-to-event outcomes applies to many clinical and epidemiologic studies in which the association between longitudinal measures and time-to-event outcomes is often of interest. This study has some limitations. First, it was assumed that chronic complications were caused by hypertension and diabetes mellitus, which may have led to an overestimation of their prevalence among patients with hypertension and diabetes. Second, data on specific potentially significant predictors, such as the type of intervention, were not available, which may have influenced the outcome variables.

# Bibliography

An, Y., Zhang, P., Wang, J., Gong, Q., Gregg, E. W., Yang, W., Li, H., Zhang, B., Shuai, Y., Chen, Y., et al. (2015). Cardiovascular and all-cause mortality over a 23-year period among chinese with newly diagnosed diabetes in the da qing igt and diabetes study. *Diabetes care*, 38(7):1365–1371.

Andrinopoulou, E.-R., Rizopoulos, D., Takkenberg, J. J., and Lesaffre, E. (2017). Combined dynamic predictions using joint models of two longitudinal outcomes and competing risk data. *Statistical methods in medical research*, 26(4):1787–1801.

Baghfalaki, T., Ganjali, M., and Berridge, D. (2013). Robust joint modeling of longitudinal measurements and time to event data using normal/independent distributions: a bayesian approach. *Biometrical Journal*, 55(6):844–865.

Caruso, P., Scappaticcio, L., Maiorino, M. I., Esposito, K., and Giugliano, D. (2021). Up and down waves of glycemic control and lower-extremity amputation in diabetes. *Cardiovascular diabetology*, 20(1):135.

Ceriello, A., De Cosmo, S., Rossi, M. C., Lucisano, G., Genovese, S., Pontremoli, R., Fioretto, P., Giorda, C., Pacilli, A., Viazzi, F., et al. (2017). Variability in hba1c, blood pressure, lipid parameters and serum uric acid, and risk of development of chronic kidney disease in type 2 diabetes. *Diabetes, Obesity and Metabolism*, 19(11):1570–1578.

Ceriello, A., Lucisano, G., Prattichizzo, F., La Grotta, R., Franzén, S., Gudbjörnsdottir, S., Eliasson, B., and Nicolucci, A. (2023). Risk factor variability and cardiovascular risk among patients with diabetes: a nationwide observational study. *European Journal of Preventive Cardiology*, 30(8):719–727.

Ceriello, A., Lucisano, G., Prattichizzo, F., La Grotta, R., Franzén, S., Svensson, A.-M., Eliasson, B., and Nicolucci, A. (2022). Hba1c variability predicts cardiovascular

complications in type 2 diabetes regardless of being at glycemic target. *Cardiovascular diabetology*, 21(1):13.

Chen, G. and Luo, S. (2018). Bayesian hierarchical joint modeling using skew-normal/independent distributions. *Communications in Statistics-Simulation and Computation*, 47(5):1420–1438.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dorajoo, S. R., Ng, J. S. L., Goh, J. H. F., Lim, S. C., Yap, C. W., Chan, A., and Lee, J. Y. C. (2017). Hba1c variability in type 2 diabetes is associated with the occurrence of new-onset albuminuria within three years. *Diabetes research and clinical practice*, 128:32–39.

Faucett, C. L. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in medicine*, 15(15):1663–1685.

Ferede, M. M., Mwalili, S., Dagne, G., Karanja, S., Hailu, W., El-Morshedy, M., and Al-Bossly, A. (2022). A semiparametric bayesian joint modelling of skewed longitudinal and competing risks failure time data: With application to chronic kidney disease. *Mathematics*, 10(24):4816.

Gao, M., Zhong, Z., Yue, Y., and Liu, F. (2022). Correlation between glycaemic variability and prognosis in diabetic patients with ckd. *Endokrynologia Polska*, 73(6):947–953.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The american statistician*, 58(1):16–24.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). Joint

models of longitudinal and time-to-event data with more than one event time outcome: a review. *The international journal of biostatistics*, 14(1):20170047.

Hu, H., Sawhney, M., Shi, L., Duan, S., Yu, Y., Wu, Z., Qiu, G., and Dong, H. (2015). A systematic review of the direct economic burden of type 2 diabetes in china. *Diabetes Therapy*, 6(1):7–16.

Hu, W., Li, G., and Li, N. (2009). A bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in medicine*, 28(11):1601–1619.

Huang, Y., Chen, J., Xu, L., and Tang, N.-S. (2022). Bayesian joint modeling of multivariate longitudinal and survival data with an application to diabetes study. *Frontiers in big Data*, 5:812725.

Ibrahim, J. G., Chu, H., and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of clinical oncology*, 28(16):2796–2801.

Jelinek, H. F., Osman, W. M., Khandoker, A. H., Khalaf, K., Lee, S., Almahmeed, W., and Alsafar, H. S. (2017). Clinical profiles, comorbidities and complications of type 2 diabetes mellitus in patients from united arab emirates. *BMJ Open Diabetes Research and Care*, 5(1):e000427.

Jun, J. E., Lee, S.-E., Lee, Y.-B., Ahn, J. Y., Kim, G., Jin, S.-M., Hur, K. Y., Lee, M.-K., and Kim, J. H. (2017). Glycated albumin and its variability as an indicator of cardiovascular autonomic neuropathy development in type 2 diabetic patients. *Cardiovascular Diabetology*, 16(1):127.

Korsa, A. T., Genemo, E. S., Bayisa, H. G., and Dedefo, M. G. (2019). Diabetes mellitus complications and associated factors among adult diabetic patients in selected hospitals of west ethiopia. *The Open Cardiovascular Medicine Journal*, 13(1).

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

Li, S., Nuyujukian, D. S., McClelland, R. L., Reaven, P. D., Zhou, J., Zhou, H., and Li, G. (2023). A joint model of the individual mean and within-subject variability of a longitudinal outcome with a competing risks time-to-event outcome. *arXiv preprint arXiv:2301.06584*.

Liu, W., Du, J., Ge, X., Jiang, X., Peng, W., Zhao, N., Shen, L., Xia, L., Hu, F., and Huang, S. (2022). The analysis of risk factors for diabetic kidney disease progression: a single-centre and cross-sectional experiment in shanghai. *BMJ open*, 12(6):e060238.

Mancia, G. (2005). The association of hypertension and diabetes: prevalence, cardiovascular risk and protection by blood pressure reduction. *Acta Diabetologica*, 42(Suppl 1):s17–s25.

Mwanyekange, J., Mwalili, S., and Ngesa, O. (2018). Bayesian inference in a joint model for longitudinal and time to event data with gompertz baseline hazards.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003). A new class of multivariate skew distributions with applications to bayesian regression models. *Canadian Journal of Statistics*, 31(2):129–150.

Sartore, G., Ragazzi, E., Caprino, R., and Lapolla, A. (2023). Long-term hba1c variability and macro-/micro-vascular complications in type 2 diabetes mellitus: a meta-analysis update. *Acta diabetologica*, 60(6):721–738.

Shita, N. G. and Muluneh, E. K. (2021). Predictors of blood glucose change and vascular complication of type 2 diabetes mellitus patients in felege hiwot referral hospital, north west ethiopia. *Scientific Reports*, 11(1):12974.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Sweeting, M. J. and Thompson, S. G. (2011). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, 53(5):750–763.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834.

Viazzi, F., Bonino, B., Mirijello, A., Fioretto, P., Giorda, C., Ceriello, A., Guida, P., Russo, G. T., De Cosmo, S., Pontremoli, R., et al. (2019). Long-term blood pressure variability and development of chronic kidney disease in type 2 diabetes. *Journal of hypertension*, 37(4):805–813.

Wan, E. Y. F., Fung, C. S. C., Fong, D. Y. T., and Lam, C. L. K. (2016). Association of variability in hemoglobin a1c with cardiovascular diseases and mortality in chinese patients with type 2 diabetes mellitus—a retrospective population-based cohort study. *Journal of Diabetes and its Complications*, 30(7):1240–1247.

Wan, E. Y. F., Yu, E. Y. T., Chin, W. Y., Fong, D. Y. T., Choi, E. P. H., and Lam, C. L. K. (2020). Association of visit-to-visit variability of systolic blood pressure with cardiovascular disease, chronic kidney disease and mortality in patients with hypertension. *Journal of Hypertension*, 38(5):943–953.

Wolde, H. F., Atsedeweyen, A., Jember, A., Awoke, T., Mequanent, M., Tsegaye, A. T., and Alemu, S. (2018). Predictors of vascular complications among type 2 diabetes mellitus patients at university of gondar referral hospital: a retrospective follow-up study. *BMC endocrine disorders*, 18(1):52.

Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to aids studies. *Journal of the American Statistical association*, 97(460):955–964.

Wu, L., Liu, W., and Hu, X. (2010). Joint inference on hiv viral dynamics and immune suppression in presence of measurement errors. *Biometrics*, 66(2):327–335.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339.

Xu, L., Huang, Y., Chen, H., Mbah, A., and Cheng, F. (2021). Joint modeling analysis of multivariate skewed-longitudinal and time-to-event data with application to primary biliary cirrhosis study.

# Chapter 3

# Progression of Diabetic Kidney Disease in People with Type 2 Diabetes using Principal Component Analysis and Ordered Logit Model


## Abstract

Diabetic kidney disease is one of the main microvascular complications caused by diabetes. Several clinical and biochemical variables are reported to be associated with diabetic kidney disease in people with type 2 diabetes. However, the interrelationships among these variables could distort estimates of their effects on disease progression. The objective of the study is to determine how the biochemical and clinical variables in people with type 2 diabetes are intercorrelated and how they affect the progression of kidney disease. Principal component analysis, combined with ordered logit models, is used to explore the interrelationships between biochemical and clinical variables and their effect on the progression of kidney disease. This retrospective cross-sectional study retrieved data from diabetic individuals in a polyclinic hospital at the University of Messina, Italy. The study identified three uncorrelated principal components. The first component, a linear combination of positively correlated glycosylated haemoglobin, glycemia, and creatinine, has a strongly significant effect on kidney disease progression. Principal component two is a linear combination of positively correlated total cholesterol and low-density lipoprotein. In contrast, Principal Component three is a linear combination of negatively correlated high-density lipoprotein and triglycerides. The cumulative odds, adjacent-category, and continuation-ratio models further revealed that age, sex, body mass index, and metformin treatment significantly affect the progression of kidney disease. However, these effects are not proportional across the stages of kidney disease progression. Flexible ordered logit models, including partial, cumulative odds, adjacent-category, and

continuation-ratio models, were used to address proportionality issues and improve the accuracy of effect estimation. The study's findings conclude that the partial, cumulative odds, adjacent-category, and continuation-ratio models are robust techniques for estimating effects, particularly when predictors have different effects at different stages of disease progression.

Keywords—Diabetic kidney disease, ordered logit model, principal component analysis, type 2 diabetes

## 3.1 Introduction

T2D is a chronic disease caused by the body's inability to produce enough insulin. According to (Siddiqui et al., 2022), the estimated prevalence of diabetes worldwide in 2019 was 9.3% (463 million people), increasing to 10.2% (578 million) by 2030 and 10.9% (700 million) by 2045. The rapid increase in the diabetes population indicates an increasing incidence of microvascular and macrovascular complications, which affect small blood vessels and larger blood vessels, respectively. Diabetic kidney disease (DKD) is one of the main microvascular complications caused by diabetes. Over time, diabetes causes the kidneys to lose function, typically manifested by a decline in the estimated glomerular filtration rate (eGFR) (De Boer and Steffes, 2007). DKD affects 40% (Alicic et al., 2017) and 51.2% (Liu et al., 2023) of people with T2D and is the leading cause of end-stage kidney disease, hospitalization, and death.

Numerous medical researchers utilize electronic record data that includes demographic, biochemical, and clinical variables to investigate the risk factors for the development and progression of kidney function in people with T2D. We can classify these risk factors into modifiable and non-modifiable categories. Studies suggest that identifying and controlling modifiable risk factors can slow the development and progression of kidney disease (Gregorich et al., 2021) and (Tziomalos and Athyros, 2015). Researchers have used a variety of methods to identify predictors for the development and progression of DKD in people with T2D. For instance, studies Bender and Grouven (1998); He et al. (2024);

Siddiqui et al. (2022) used binary logistic regression to classify eGFR into two categories. Their findings demonstrated that clinical and biochemical variables were associated with the progression of kidney function in individuals with type 2 diabetes.

Researchers also used multivariate linear mixed models that aimed to predict the progression of kidney function (Ye et al., 2022) and (Ceriello et al., 2017). Their research output reveals that several modifiable factors, such as glucose, blood pressure, and lipid profile, have a significant impact on the development and progression of kidney disease in people with type 2 diabetes.

Furthermore, studies utilised a survival regression model to identify risk factors for kidney function in people with T2D. For example, a nationwide retrospective cohort study was conducted in Italy using data from the Italian Association of Clinical Diabetologists. They used a Cox regression model to assess risk factors for the incidence of kidney disease and progression of kidney function in T2D (Russo et al., 2016) and (Liu et al., 2023). The result of their investigation revealed that variability in biochemical variables such as HbA1c, blood pressure, and lipid parameters was a significant risk factor for developing DKD in people with T2D. Specifically, (Russo et al., 2016) suggested that the risk of kidney disease is related to cardiovascular disease risk factors; that is, high triglycerides and LDL and HDL concentrations were independent risk factors for the development of DKD.

Although significant research has been conducted, diabetic kidney disease in individuals with T2D continues to be an essential concern for medical researchers. However, relatively little work has been done on simultaneously accounting for the biases induced by the following three issues. Firstly, to determine the risk factor for kidney disease in type 2 diabetes, researchers have so far used a binary logistic regression model, categorizing kidney function by eGFR (DKD vs. non-DKD). In fact, there are five categories of kidney function. These methods might not be practical for studying kidney function progression, often leading to information loss and significant reductions in statistical power when specific response categories are ignored (Ananth and Kleinbaum, 1997). Secondly, patients

with type 2 diabetes, parsimonious multi-biochemical, clinical, and demographic factors determine the incidence and progression of kidney disease (Tan et al., 2017)

Some of these clinical and biochemical risk factors are strongly interrelated and, at the same time, responsible for the risk of kidney disease. Including these variables as predictors in the regression model will lead to multicollinearity, potentially biasing effect estimates. Thirdly, researchers assessed the impact of various biochemical, clinical, and demographic variables on diabetic kidney disease in individuals with type 2 diabetes. Some variables alone do not significantly influence the incidence and development of the disease. However, combining correlated variables can reveal significant effects on disease progression.

Fourthly, many medical and epidemiologic studies often measure outcomes of interest on an ordinal scale, for example, tumor grade (grades I–III), stage of kidney function (stages I–IV), cardiac risk levels (low, medium, and high risk), and disease severity (low, medium, and severe). The order logit model is one of the most common methods in medical research for analysing ordinal-level response variables. Although researchers have developed various ordered logit models for ordinal response variables, such as proportional odds, continuation ratio, and adjacent-category models, the proportional odds model is probably the most frequently used in practice. When the predictor's variables violated the proportional odds assumption, this approach led to biased statistical inferences (Bender and Grouven, 1998).

To address this gap, our study used a relatively novel statistical method, PCA, combined with an ordered logit model. PCA is widely used to summarize information from high-dimensional data sets. It is used in medical science to find essential features in large amounts of data, like cancer and gene expression studies (Hsu et al., 2014) and (Adiwijaya et al., 2018) in cardiology to predict clinical cardiovascular events(Okin et al., 2002)and (Kristono et al., 2020). Principal component analysis is used in regression models (Çamdevỳren et al., 2005; Milewska et al., 2014; Zhang and Castelló, 2017) when many predictors are highly correlated, causing multicollinearity. This multicollinearity

increases the standard errors of the regression coefficients, leading to inaccurate effect estimates. Using PCA, we can reduce the dimensionality of correlated variables to a small number of uncorrelated candidate predictors, without losing valuable information, and thereby improve the regression model's effect estimation. Thus, this study uses PCA to construct new variables (principal components), which are linear combinations of the original variables, as predictors of kidney disease progression and to uncover hidden relationships among biochemical and clinical variables in individuals with T2D.

The general objective of this study is to identify risk factors for the progression of kidney disease in people with type 2 diabetes by applying advanced statistical methods to available clinical and biochemical variables. The study has the following specific objectives:

I) Apply principal component analysis combined with an ordered logit model to investigate how clinical and biochemical parameters interrelate and affect the progression of kidney disease.

II) Model the progression of kidney disease in people with type 2 diabetes to identify variables with significant effects and assess potential risk factors.

III) Demonstrate how to use and interpret principal component analysis when working with an ordered logit model that exhibits collinearity.

This study contributes novel statistical methodologies to analyse disease patterns and risk factors in chronic diseases, providing new insights into the progression of diabetic kidney disease. The research also assists healthcare professionals in developing and strategizing preventive interventions and in enhancing patients' understanding of disease progression. The results offer significant insights for medical researchers studying chronic diseases, particularly regarding the relationships between clinical and biochemical factors and their influence on disease progression.

## 3.2 Material and Methods

### 3.2.1 Study Design

This retrospective cross-sectional study retrieved data from a type 2 diabetic cohort at a polyclinic hospital in the University of Messina, Italy. We extracted all demographic, clinical, and biochemical variables from the electronic health records at the Polyclinic Hospital. The dataset contains various missing values, which were treated as noise and removed. Among 407 individuals, 84 were excluded due to missing data, so that 323 individuals were included in the study.

### 3.2.2 Study Variables

The outcome variable for the study is kidney function, as quantified by the eGFR. According to the 2012 Kidney Disease: Improving Global Outcomes (KDIGO) guidelines, individual kidney function based on eGFR is classified as normal (eGFR $\geq$ 90 mL/min/1.73 $m^2$), mildly reduced kidney function (eGFR 60 to 89 mL/min/1.73 $m^2$), moderately reduced kidney function (eGFR 30 to 59 mL/min/1.73 $m^2$), severely reduced kidney function (eGFR 15 to 29 mL/min/1.73 $m^2$) and kidney failure or end-stage kidney disease (eGFR <15 mL/min/1.73 $m^2$). Based on eGFR, our study classified kidney function as usual, mildly reduced (DKD), or moderately reduced (DKD). It does not include the last two categories because all individuals fall into the first three.

The biochemical and clinical variables: age, sex, duration of diabetes, BMI, ischemic heart disease, retinopathy, and metformin, gamma-glutamyl transferase (GGT), glutamate pyruvate transaminase (GPT), glutamate oxaloacetate transaminase (GOT), HbA1c, triglyceride (TG), HDL, LDL, and total cholesterol(TC).

### 3.2.3 Description of Motivating Dataset

Our primary motivation is the interrelationships among biochemical variables and their impact on the progression of kidney disease in individuals with type 2 diabetes. We

present a brief background of the medical dataset used in the study. The distribution of the biochemical variable and its basic descriptive statistics are presented in Table 3.1.

Table 3.1: Descriptive statistics for biochemical variables measured for diabetes patients ($N = 323$)

| Variables | Minimum | Maximum | Mean (SD) | Shapiro Test |
|---|---|---|---|---|
| HbA1C | 4.9 | 10.9 | 7.59 (1.06) | 0.06 |
| Creatinine | 0.03 | 2.2 | 1.04 (0.28) | 0.00183 |
| Glycemia | 77 | 277 | 162.64 (35.61) | 0.00187 |
| Total Cholesterol | 99 | 266 | 162.65 (32.59) | 0.0028 |
| HDL | 15 | 98 | 46.43 (12.34) | 0.00188 |
| LDL | 36 | 172 | 88.45 (26.45) | 0.0035 |
| Triglycerides | 20 | 297 | 139.05 (56.43) | 0 0.007 |
| GPT | 6 | 76 | 26.32 (11.64) | 0.0000216 |
| GOT | 10 | 67 | 22.22 (8.17) | 0.000022 |
| GGT | 8 | 85 | 31.45 (12.45) | 0.000022 |

In this study, we categorize kidney function as normal, mildly reduced, or moderately reduced (DKD). Due to space constraints, Figure 3.1 displays only two categories of kidney function; however, for graphical visualization and further analysis, the data were categorized into three groups. Our objective is to explore how the biochemical variables are interrelated and to observe how changes in these variables affect kidney function.

Figure 3.1 illustrates the interrelationships among biochemical variables and their relationship with diabetic kidney disease simultaneously. The scatter plot indicates that all biochemical variables exhibit an approximately linear relationship with one another. The diagonal density curves illustrate the distributions of biochemical variables for non-DKD and DKD. The Shapiro–Wilk test confirmed normality only for HbA1c $p > 0.05$ as shown Table 3.1. Although the biochemical variables did not meet normality assumptions, no transformations were applied, as the principal component and ordered logit models used

in this study do not require normally distributed variables.

Figure 3.1 also indicates the pattern of the relationship between biochemical variables and diabetic kidney disease. As the values of HbA1c, glycemia, creatinine, triglycerides, and LDL increase, the box plots of kidney function vary across categories, as shown in the figure. The figure clearly shows that people with type 2 diabetes who develop diabetic kidney disease have higher HbA1c, glycemia, creatinine, and triglycerides, as well as lower total cholesterol and LDL values than those with non-diabetic kidney disease. Conversely, as shown in the figure, despite fluctuations in the values of and HDL, GOT, GGT, and GPT, kidney function remains consistent across the two categories. This implies that kidney function is unaffected by the biochemical variables HDL, GOT, GGT, and GPT.

It also indicates that most of the biochemical variables have a strong correlation (Pearson correlation coefficient $p > 0.05$ with each other. We observed the strong correlations between glycemia and HbA1c, total cholesterol and low-density lipoprotein, HDL and triglycerides, and GOT and GPT. The strong correlation among the biochemical variables indicated multicollinearity. Using these variables as predictors in a regression model leads to multicollinearity issues. We cannot ignore this problem, as it significantly impacts statistical inference. One way to address multicollinearity is to use PCA on these variables. Despite the strong correlation between these biochemical variables (GOT and GPT), it is unnecessary to include them in PCA analyses, as they do not affect kidney function.

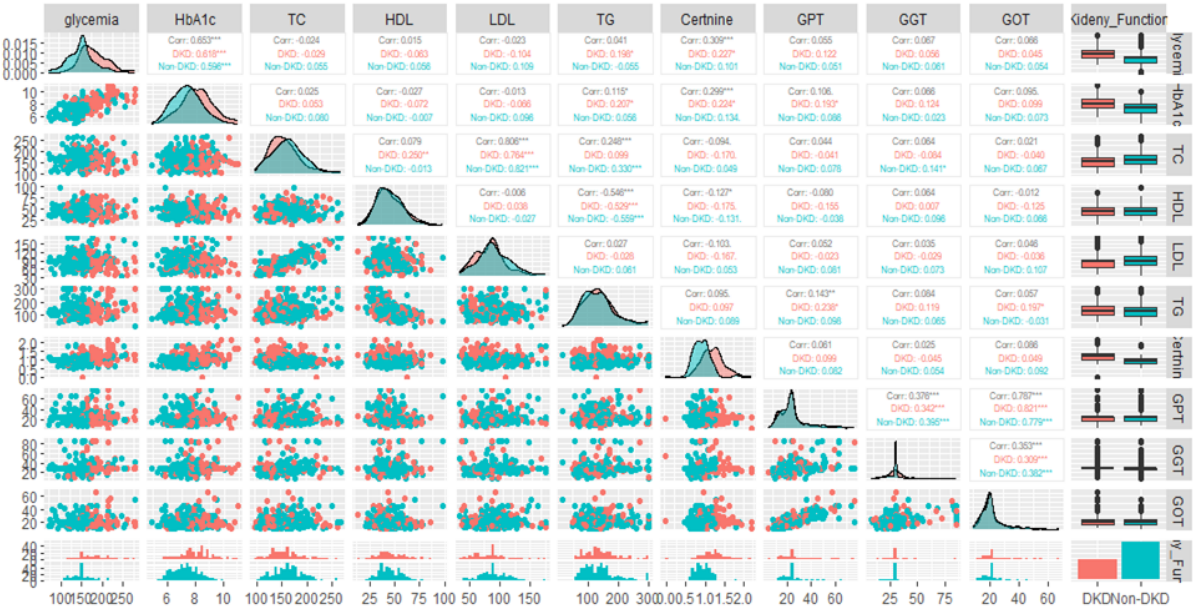



Figure 3.1: Scatter plot and correlation matrix between biochemical variables by kidney function

### 3.2.4 Statistical Analysis

### 3.2.5 Principal Component Analysis

In regression modeling, relying on only a few covariates may not adequately explain the response variable's variation. The model becomes more informative when a broader set of covariates is included; however, this set may contain correlated variables, which can introduce multicollinearity into the regression model. For instance, this study found strong correlations between most of the biochemical variables. One approach to solving this problem is to apply principal component analysis to these variables.

Principal component analysis is a dimension-reduction method for understanding a large data set and addressing multicollinearity to obtain unbiased parameter estimates in a regression model. This approach converts the original variables into a new set of orthogonal, uncorrelated variables, known as principal components. We performed principal component analysis on the correlated biochemical variables to rank their relative significance, describe their interrelation patterns, and understand their impact on kidney function.

To provide a clearer illustration of how PCA is applied, consider a data matrix $X$ with dimensions $n \times p$ where $n$ represents the number of observations and $p$ denotes the number of biochemical variables. Let $C$ be a $p \times p$ correlation matrix of these biochemical variables. Additionally, let $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_p$ be the eigenvalues of the correlation matrix $C$, and let $W = (w_1, w_2, \ldots, w_p)$ be the $p \times p$ matrix consisting of the normalized eigenvectors associated with each eigenvalue. Note that the eigenvalues are the solutions of the determinant equation.

$$|C - \lambda I| = 0$$

and $w_j$ are the solutions that satisfy the set of homogeneous equations

$$(C - \lambda_j I)w_j = 0.$$

Each $\lambda_j$ denotes the amount of variance explained in $C$ (the correlation matrix of the biochemical variables), and $w_j$ are the corresponding directions called the principal directions. The $i$-th principal component on the dataset $X$ in the direction of the principal direction $w_j$ is

$$PC_i = w_j X.$$

$$\begin{aligned}
PC_1 &= w_{11}x_1 + w_{12}x_2 + w_{13}x_3 + \cdots + w_{1p}x_p, \\
PC_2 &= w_{21}x_1 + w_{22}x_2 + w_{23}x_3 + \cdots + w_{2p}x_p, \\
&\vdots \\
PC_p &= w_{p1}x_1 + w_{p2}x_2 + w_{p3}x_3 + \cdots + w_{pp}x_p.
\end{aligned}$$

We refer to the newly created variables $PC_1$, $PC_2$, $\ldots$, $PC_p$ as principal components, and we identify the variable PC1 as the component corresponding to the greatest eigenvalue $\lambda_1$. This indicates that the first principal component captures the maximum variation in the data. Subsequently, each principal component captures the maximum remaining variance not explained by the previous components. We establish the $ith$ principal

component ($PC_i$) under conditions where the eigenvectors maintain unit length and orthogonality to each other.

$$\sum_{j=1}^{p} w_{ij}^2 = 1,$$

and the orthogonality condition between different eigenvectors is

$$\sum_{j=1}^{p} w_{ij}\, w_{kj} = 0 \quad \text{for } i \neq k.$$

We interpret the eigenvectors $w_1,\ w_2,\ \ldots,\ w_p$ as coefficients of principal components, and the sum of the variances of the principal components reveals the total variance of the initial variables. This provided clear information on each principal component's significance and contribution, along with the percentage of total variability it explained. We can calculate the percentage of variance that the $kth$ principal component contributed as:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}.$$

Although the number of principal components equals the number of original variables, we selected only the first few for further analysis, as they capture most of the variability in the original dataset. To decide which and how many principal components to retain, we used two methods: selecting components with an eigenvalue greater than one and examining a scree plot of eigenvalues against the corresponding numbers of components (Zwick and Velicer, 1986)

A crucial aspect of principal component analysis is that the loadings represent the correlations between the original variables and the principal components. These values provide clear information about the relationship between the biochemical variables and the given principal component; the greater the loading, the more influential the variable is in forming the principal component, and vice versa. For instance, we calculate the loading of a

biochemical variable $X_j$ for principal component i ($PCi$) as follows:

$$L_{ij} = w_{ij}\sqrt{\lambda_i}$$

where $L_{ij}$ is the correlation between the $i$-th principal component and the $j$-th variable, $w_{ij}$ is the principal component weight of the $j$-th variable in the $i$-th principal component, and $\lambda_i$ is the eigenvalue associated with the $i$-th principal component.

### 3.2.6 Regression Model for Ordinal Response Variables

To determine how clinical and biochemical variables affect the progression of kidney disease, ordinal logistic regression was used, with kidney function categorized as normal, mildly reduced, or moderately reduced. Several types of ordinal logistic regression models exist, each based on comparisons of the response category (cumulative, stage, or adjacent) and the failure to meet the proportional odds assumption (partial cumulative odds model, partial continuation ratio model, partial adjacent category models, and generalized ordinal logistic models). In this study, we conducted an in-depth investigation of the specific applications of each ordered logit model.

### 3.2.7 Proportional Odds Model

The proportional odds model (McCullagh, 1980), also known as the cumulative logit model, is the most widely used ordinal logistic regression model because of its simple interpretation. We used the model when the effect of each predictor variable is constant across the response variable's categories. This restriction, or assumption, is referred to as a proportional odds assumption. As previously stated, individual kidney function, denoted by Y (grouped from continuous variable eGFR), is classified as usual, stage II, and stage III or DKD, and $x_1,\ x_2,\ \ldots,\ x_p$ represent a vector of p-dimensional explanatory variables. Thus, the logit of the proportional odds model can be defined in the following way:

$$\log\left[\frac{\Pr(Y \leq j \mid X)}{\Pr(Y > j \mid X)}\right] = \alpha_j - \sum_{p=1}^{p} \beta_p X_p, \qquad j = 1, 2. \tag{3.1}$$

where $\boldsymbol{\beta}$ the $(p \times 1)$ vector containing the regression coefficients. The $\alpha_j$ are thresholds satisfying the condition $\alpha_1 < \alpha_2$.

### 3.2.8 Partial Proportional Odds Model

The partial proportional odds model is an extension of the proportional odds model, specifically designed for situations where one or more predictor variables violate the proportionality assumption (Peterson and Harrell Jr, 1990). If one or more predictor variables have varying effects on different stages of disease progression, the model will have two sets of regression coefficients, one set with proportional odds and the other with nonproportional odds. The logit of the partial proportional odds model can be defined as follows:

$$\log\left[\frac{\Pr(Y \leq j \mid X)}{\Pr(Y > j \mid X)}\right] = \alpha_j - \left(\sum_{p=1}^{p} \beta_p X_p + \lambda_j Z\right), \qquad j = 1, 2. \tag{3.2}$$

Where $X$ is the vector containing the full set of covariates and $Y$ is the response variable. $Z$ is a vector containing a subset of covariates that violate the assumption of proportional odds, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of regression coefficients, $\lambda_j$ $(j = 1, 2)$ are the regression coefficients associated with the variables in $Z$, and $\alpha_j$ are the thresholds.

### 3.2.9 Continuation Ratio Model

The continuation ratio model (Shrout, 1979) predicts the likelihood of a response falling into a specific category rather than the probability of a response falling into a higher category, based on the response variable's order category. This type of model is particularly useful in situations where individuals progress through successive response levels, such as the progression of kidney disease from a normal to a severe condition. The logit of the continuation ratio model can be defined as follows:

$$\log\left[\frac{\Pr(Y = j \mid X)}{\Pr(Y > j \mid X)}\right] = \alpha_j - \sum_{p=1}^{p} \beta_p X_p, \qquad j = 1, 2. \tag{3.3}$$

where $Y$ is the response variable, $X$ is the vector of $p$-dimensional explanatory variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of regression coefficients, and $\alpha_j$ are the thresholds.

### 3.2.10 Partial Continuation Ratio Model

The partial continuation ratio model is an extension of the continuation ratio model in which one or more predictor variables fail to meet the proportionality assumption. The logit of the partial continuation Ratio model incorporates coefficients for covariates that deviate from the proportionality assumption and can be defined in the following way:

$$\log\left[\frac{\Pr(Y=j \mid X)}{\Pr(Y>j \mid X)}\right] = \alpha_j - \left(\sum_{p=1}^{p} \beta_p X_p + \lambda_j Z\right), \qquad j = 1, 2. \tag{3.4}$$

Where $Y$ is the response variable and $X$ is the vector containing the full set of covariates. $Z$ is a vector containing a subset of covariates that violate the assumption of proportional odds, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of regression coefficients, $\lambda_j$ $(j = 1, 2)$ are the regression coefficients associated with the variables in $Z$, and $\alpha_j$ are the thresholds.

### 3.2.11 Adjacent Category Model

The adjacent-category model is a type of ordinal logistic regression used to predict the probability of adjacent categories of the response variable (Goodman, 1983). In our study, we used it to predict the likelihood of normal kidney function versus moderately reduced kidney function. The logit of the adjacent category model can be defined as follows:

$$\log\left[\frac{\Pr(Y=j \mid X)}{\Pr(Y=j+1 \mid X)}\right] = \alpha_j - \sum_{p=1}^{p} \beta_p X_p, \qquad j = 1, 2. \tag{3.5}$$

where $Y$ is the response variable, $X$ is the vector of $p$-dimensional explanatory variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of regression coefficients, and $\alpha_j$ are the thresholds.

### 3.2.12 Partial Adjacent Category Model

When certain independent variables in the adjacent-category model fail the proportionality assumption, a more concise model is needed to address the issue. The partial adjacent category model predicts the probability of adjacent categories, allowing the effects of one or two predictor variables to vary across categories of the response variable. By modifying the equation of the logit of the adjacent category model, we can obtain the following equation for the logit of the partial adjacent category model.

$$\log\left[\frac{\Pr(Y=j \mid X)}{\Pr(Y=j+1 \mid X)}\right] = \alpha_j - \left(\sum_{p=1}^{p} \beta_k X_k + \lambda_j Z\right), \qquad j = 1, 2. \tag{3.6}$$

where $Y$ is the response variable and $X$ is the vector containing the full set of covariates. $Z$ is a vector containing a subset of covariates that violate the assumption of proportional odds, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of regression coefficients, $\lambda_j$ are the regression coefficients associated with the variables in $Z$, and $\alpha_j$ are the thresholds.

### 3.2.13 Generalised Ordered Logit Model

The generalized ordered logit model is a generalization of the proportional odds model that allows all predictor variables to violate the parallel assumption. This model enables the regression coefficients to vary across the j-1 categories of the response variables.

$$\log\left[\frac{\Pr(Y=j \mid X)}{\Pr(Y=j+1 \mid X)}\right] = \alpha_j - \left(\sum_{p=1}^{p} \beta_j X + \lambda_j Z\right), \qquad j = 1, 2. \tag{3.7}$$

where $Y$ is the response variable, $X$ is the $p$-dimensional vector of explanatory variables, $\beta_j$ is a $(p \times 1)$ vector of regression coefficients, and $\alpha_j$ are thresholds.

## 3.3 Parameter Estimation

We used the maximum likelihood estimation (MLE) method to estimate the parameters of various types of ordinal logistic regression models. Let $(x_i, y_j)$ be a sample of size $n$,

where the vector $x_i$ contains the observed values of the $p$ explanatory variables and $y_j$ is the ordinal response variable with $j$ categories.

We denote the probability of an observation falling into category $j$ or lower (cumulative probability) as $P(Y \leq y_j \mid x_i)$. The likelihood function $L$ is then the product of these cumulative probabilities for all observations:

$$L(\alpha_1, \alpha_2, \ldots, \alpha_{j-1}, \beta) = \prod_{i=1}^{n} P(Y_i \leq y_j \mid x_i). \tag{3.8}$$

where $n$ is the total number of observations, $Y_i$ is the response variable for observation $i$, $x_i$ is the matrix of predictor variables for observation $i$, and $y_j$ is the category of the response variable for observation $i$.

The MLE principle is employed to estimate the parameter vector.

$$V = (\alpha_1, \alpha_2, \ldots, \alpha_{j-1}, \beta)$$

by maximizing the likelihood function. To simplify the calculation, the likelihood function is transformed into the natural logarithm of the likelihood, which is formulated as:

$$\ell(\alpha_1, \alpha_2, \ldots, \alpha_{j-1}, \beta) = \sum_{i=1}^{n} \log \big(P(Y_i \leq y_j \mid x_i)\big). \tag{3.9}$$

Maximizing this log-likelihood function with respect to the parameters $\alpha_1, \alpha_2, \ldots, \alpha_{j-1}, \beta$ yields the Maximum Likelihood Estimates of these parameters.

## 3.4 Results

### 3.4.1 Descriptive Statistics

Descriptive statistics were used to summarize the demographic and clinical characteristics of people with type 2 diabetes. Table 3.2 summarizes the distributions of the demographic

and clinical variables for each stage of kidney function in people with T2D. Among these people, 34.7% had moderately reduced kidney function, or stage III, defined by an eGFR lower than 60 ml/min/1.73 $m^2$, and 51.1% had mildly reduced kidney function, or stage II, defined by an eGFR between 60 to 89 ml/min/1.73 $m^2$, whereas 14.2% had normal kidney function (eGFR greater than 90 ml/min/1.73 $m^2$). The study included 60.4% males and 39.6% females, with a mean age of 68 years. The average length of time since being diagnosed with type 2 diabetes was 12 years, with a standard deviation of 7 years. Approximately 75.1% of people with type 2 diabetes use metformin as a medication, and 18% and 14% have a history of ischemic heart disease and retinopathy, respectively.

Table 3.2: Descriptive statistics for clinical variables measured for diabetes patients (N = 323). The proportions for categorical variables and the means (SD) for quantitative variables

| **Variables** | **Total** | **Normal** | **Stage II** | **DKD** |
|---|---|---|---|---|
| Number of Patients | 323 | 46 (14.2) | 165 (51.1) | 112 (34.7) |
| **Gender** | | | | |
| Male | 195 (60.4%) | 31 (9.6%) | 97 (30.0%) | 67 (20.7%) |
| Female | 128 (39.6%) | 15 (4.6%) | 68 (21.1%) | 45 (13.9%) |
| **Retinopathy** | | | | |
| Yes | 48 (14.9%) | 3 (0.9%) | 23 (7.1%) | 22 (6.8%) |
| No | 275 (85.1%) | 43 (13.3%) | 142 (44.0%) | 90 (27.9%) |
| **Ischemic heart disease** | | | | |
| Yes | 61 (18.9%) | 12 (3.7%) | 29 (9.0%) | 20 (6.2%) |
| No | 262 (81.1%) | 34 (10.5%) | 136 (42.1%) | 92 (28.5%) |
| **Metformin treatment** | | | | |
| Yes | 275 (85.1%) | 42 (13.0%) | 148 (45.8%) | 85 (26.3%) |
| No | 48 (14.9%) | 4 (1.2%) | 17 (5.3%) | 27 (8.4%) |
| Age (years) | 68.92 (6.18) | 67.14 (4.93) | 68.15 (5.77) | 70.77 (6.78) |
| Body mass index (kg/m$^2$) | 30.73 (6.18) | 31.98 (4.90) | 30.85 (5.26) | 30.07 (5.66) |
| Duration of diabetes (years) | 12.54 (7.11) | 9.91 (4.73) | 11.84 (6.97) | 14.64 (7.58) |

Figure 3.2 shows box plots of the distributions of biochemical and clinical variables at each stage of kidney function. The plots reveal that changes in age, diabetes duration, BMI, HbA1c, glycemia, creatinine, total cholesterol, and LDL are associated with variations in kidney function. However, despite fluctuations in triglyceride and HDL values, kidney function exhibits only slight differences across all three stages. The pie chart shows the frequency distribution of kidney function in people with diabetes.

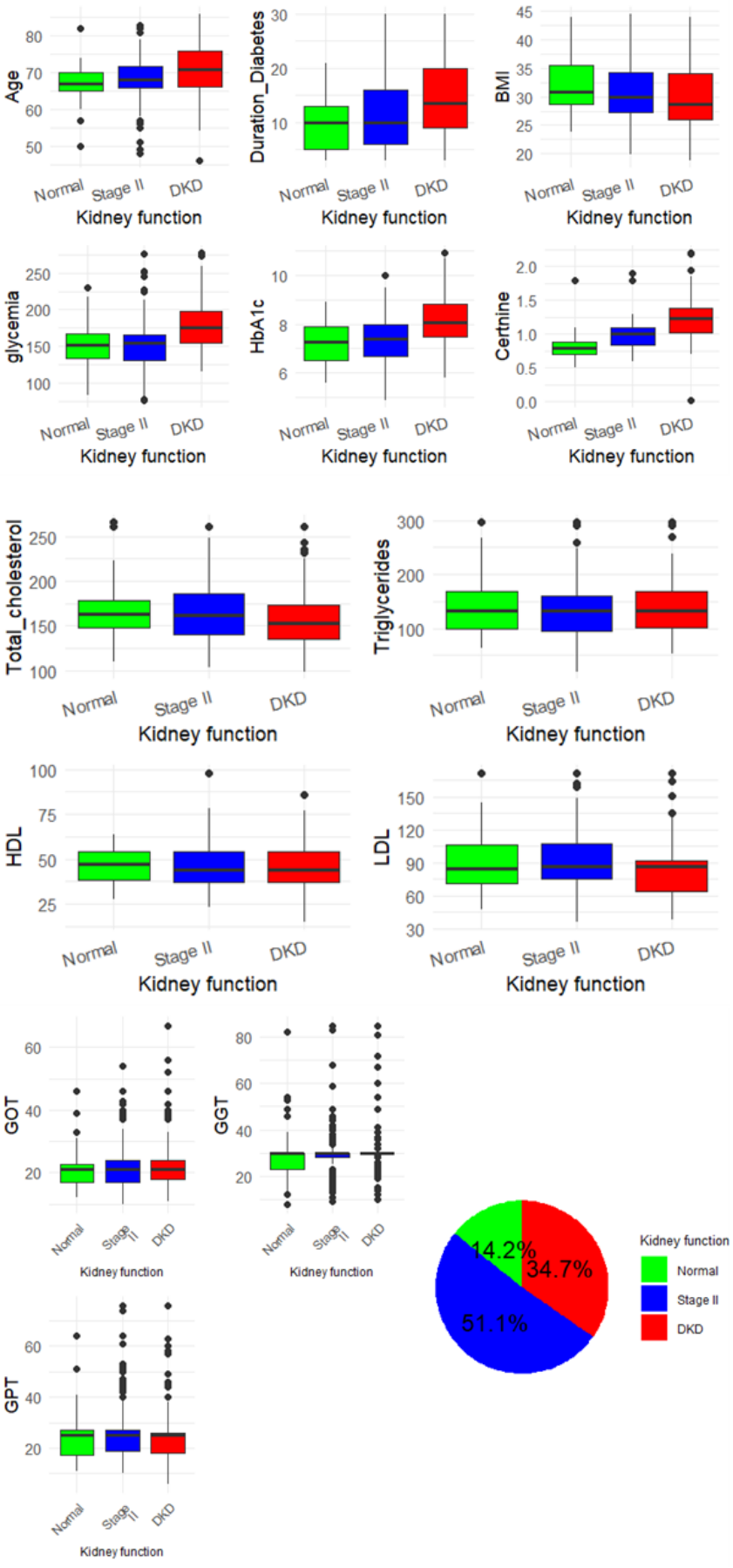



Figure 3.2: Distribution of clinical and biochemical variables for each stage of kidney function.

### 3.4.2 Results of Principal Component Analysis

Table 3.3 provides a comprehensive summary of the principal component analysis results, including the standard deviation, proportion of variance, and cumulative percentage explained by each component. The table reveals that the first three components account for 27.7%, 26.6%, and 21.1% of the variation in the biochemical variables, respectively.

Table 3.3: The proportion of variance explained by each component for biochemical variables.

| | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** |
|---|---|---|---|---|---|---|---|
| Std. Dev. | 1.3931 | 1.364 | 1.216 | 0.871 | 0.7067 | 0.5846 | 0.3341 |
| Prop. Var. | 0.277 | 0.266 | 0.211 | 0.108 | 0.071 | 0.048 | 0.016 |
| Cum. Prop. | 0.2772 | 0.5432 | 0.7547 | 0.8632 | 0.9346 | 0.9834 | 1.0000 |

Table 3.4 displays the loadings that represent the strength and direction of the relationships between the biochemical variables and the principal components. Regardless of sign, the greater the loading magnitude, the greater the contribution of the biochemical variables to the formation of the principal components. For example, -0.7513, -0.6294, and -0.7503 indicate the contributions of the biochemical variables glycemia, creatinine, and HbA1c, respectively, to component one (PC1). This shows that the first component is a composite variable that captures the overall trend of decreases in glycemia, creatinine, and HbA1c. Similarly, we can interpret component two as a composite variable that captures the overall trend of reducing low-density lipoprotein cholesterol and total cholesterol. We can interpret component three as a composite variable that captures the overall trend of decreasing high-density lipoprotein cholesterol and increasing triglycerides. This insight helps in understanding the underlying relationships among biochemical variables and identifying key patterns.

Table 3.4: Principal Component Loadings

| **Variables** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** |
|---|---|---|---|---|---|---|---|
| Glycemia | -0.7513 | -0.2402 | -0.3859 | 0.2267 | -0.1037 | -0.4078 | 0.0132 |
| Creatinine | -0.6249 | -0.0980 | -0.0059 | -0.7711 | 0.0712 | 0.0090 | -0.0032 |
| HbA1c | -0.7503 | -0.2233 | -0.3211 | 0.2665 | 0.0050 | 0.4148 | -0.0050 |
| TC | 0.3749 | -0.8596 | -0.1962 | -0.0570 | 0.1632 | -0.0067 | 0.2279 |
| LDL | 0.4104 | -0.7987 | -0.2252 | -0.1198 | -0.2705 | 0.0059 | -0.2051 |
| HDL | 0.2523 | 0.2657 | -0.8061 | -0.0251 | 0.4534 | -0.0287 | -0.0936 |
| TG | -0.2233 | -0.5038 | 0.6899 | 0.1548 | 0.4220 | -0.0694 | -0.1151 |

Table 3.4 also shows that the number of PCs equals the number of variables included in the analyses. As described in the methodology section, the main objective of PCA is to reduce the dimension of the intercorrelated biochemical variables to a small number of uncorrelated variables while retaining all pertinent information. To determine the number of principal components, we considered the principal component standard deviations and eigenvalues, as well as the scree plot. Table 3.3 displays the eigenvalue or standard deviations for each of the seven principal components, while Figure 3.3 presents the scree plot. The study selected the first three components for further analysis since their standard deviations exceeded one. A scree plot visualizes the proportion of variance of each PC, revealing a deep drop for PC1, PC2, and PC3, which stabilizes from PC4 onward. This pattern suggests that the first three principal components capture most of the total variance. Both the standard deviation and scree plot methods for determining the number of principal components indicate that three PCs were sufficient for further analysis (in this case, as covariates for an ordered logit regression model), as they explain about 76% of the variance in the considered biochemical variables.

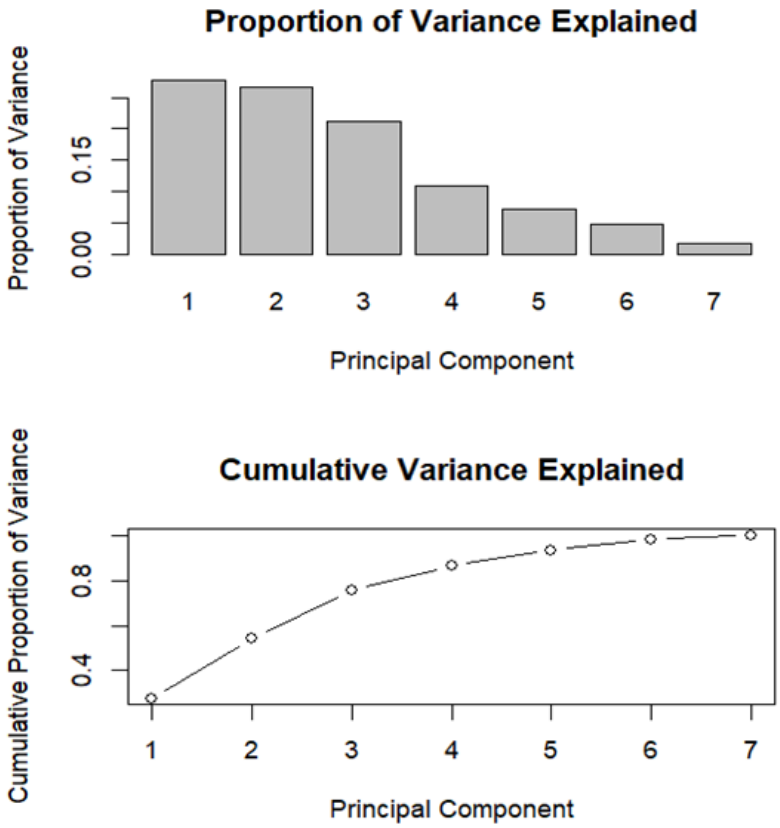



Figure 3.3: Scree plot for the percentage of explained variances against each component.

Figure 3.4 illustrates the relationship between biochemical variables and their role in forming the first two principal components. Lines that are closely aligned in the same quadrant of the plot indicate a positive correlation, whereas segments positioned on the opposite side of the quadrant indicate a negative correlation. The proximity of the lines to the circle signifies the strength of the relationship. When the lines are perpendicular, the variables are uncorrelated. TG and HDL have a negative relationship because they are in opposite quadrants of the plot. Positively correlated biochemical variables primarily load on the first principal component, with negative loadings for HbA1c, glycemia, creatinine, and TG, and positive loadings for TC, HDL, and LDL. This component explains 27.7% of the total variance. The second principal component is mainly composed of positively correlated variables with negative loadings of TC and LDL. This component explains 26.6% of the variance.

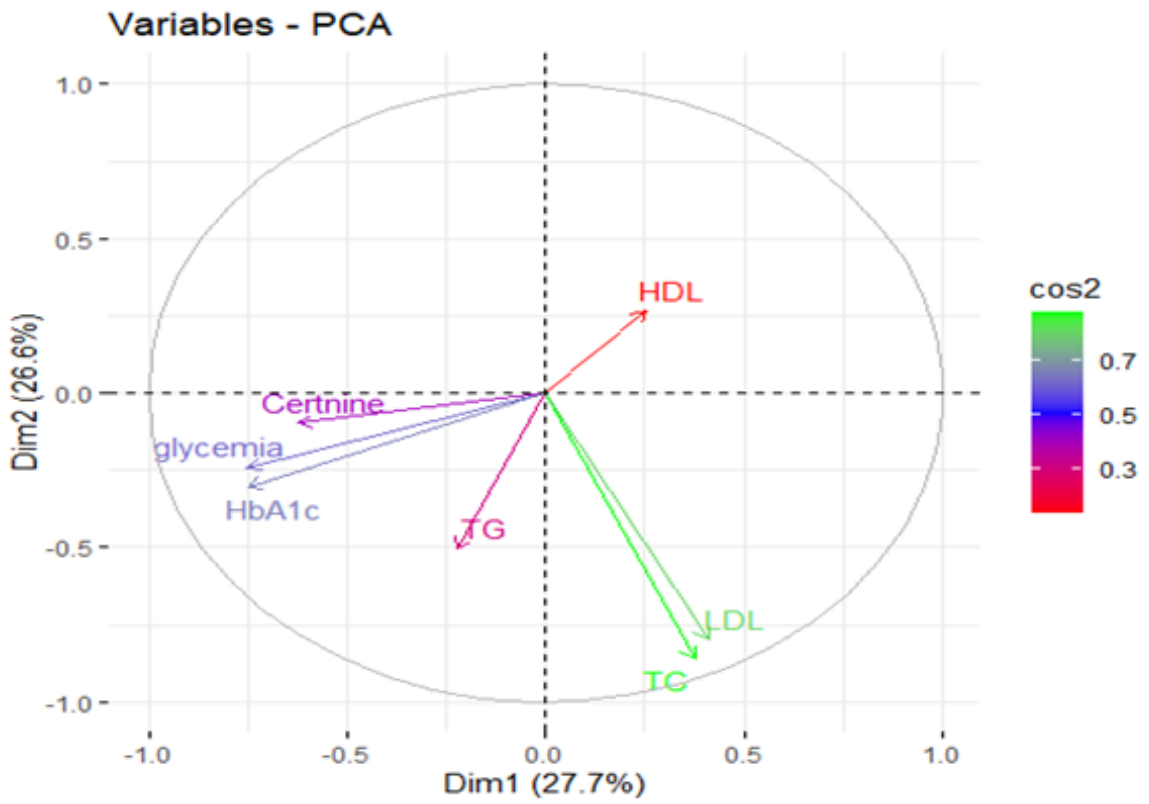



Figure 3.4: Correlation between biochemical variables and principal components.

Another notable feature of principal components is their capacity to compute individual score values or individual measurements. Each principal component score is a linear combination of the original variables weighted by the corresponding loadings for that principal component. Table 3.5 displays the eigenvectors or weights assigned to variables for each principal component. These values indicate each variable's contribution to the principal component score. In the ordered logit regression model, we used PCA-derived score values as covariates to identify the most significant components associated with kidney disease progression. The methodology section details the process of obtaining each principal score or data value. This involves entering the standardized data values of the biochemical variables into the estimated linear functions that constitute each principal component. For example, the principal component scores for the first individual (PC11)

can be computed as follows:

$$PC_1 = -0.5393\,\text{glycemia}_1 - 0.4485\,\text{creatinine}_1 - 0.5386\,\text{HbA1c}_1 + 0.2691\,\text{TC}_1 + 0.2946\,\text{LDL}_1 + 0.1811\,\text{HDL}_1 - 0.0160\,\text{TG}_1.$$

Table 3.5: Weights (eigenvectors) of the principal components

| **Variables** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** |
|---|---|---|---|---|---|---|---|
| Glycemia | -0.5393 | -0.1760 | -0.3171 | 0.2601 | -0.1468 | -0.6976 | 0.0388 |
| Creatinine | -0.4485 | -0.0718 | -0.0048 | -0.8849 | 0.1007 | 0.0154 | -0.0095 |
| HbA1c | -0.5386 | -0.2233 | -0.2641 | 0.3059 | 0.0072 | 0.7045 | -0.0177 |
| TC | 0.2691 | -0.6300 | -0.1612 | -0.0655 | 0.2310 | -0.0115 | 0.6684 |
| LDL | 0.2946 | -0.5853 | -0.2078 | -0.1375 | -0.3828 | 0.0102 | -0.6015 |
| HDL | 0.1811 | 0.1947 | -0.6625 | -0.0288 | 0.6416 | -0.0490 | -0.2746 |
| TG | -0.1603 | -0.3692 | 0.5670 | 0.1777 | 0.5971 | -0.1188 | -0.3377 |

### 3.4.3 Results of Ordinal Logistic Regression Models

The study used three uncorrelated principal components as covariates in the ordinal logistic regression model, rather than seven correlated biochemical variables. Cumulative proportional odds regression, continuation ratio regression, and adjacent category regression were used to model the associations between kidney disease progression and the principal components and clinical variables. Table 3.6 provides the estimated regression coefficients $\beta$ for the clinical variables and components, along with their associated odds ratios. The results revealed that age, sex, body mass index, component 1 (a linear combination of HbA1c, glycemia, and creatinine), and metformin treatment had significant effects on kidney disease progression across all three models. The magnitude of these associations varied slightly across the models. At the 5% level of significance, diabetes duration, component 2 (a linear combination of total cholesterol and LDL), and component 3 (a linear combination of HDL and triglycerides) did not have significant effects on

the progression of kidney disease.

Table 3.6: Summary of estimates, standard errors in brackets, and odds ratios of the proportional odds, continuation ratio, and adjacent category model

| | **Cumulative Odds Model** | | **Continuation Ratio Model** | | **Adjacent Category Model** | |
|---|---|---|---|---|---|---|
| **Covariates** | Estimate (Std. Error) | Odds Ratio | Estimate (Std. Error) | Odds Ratio | Estimate (Std. Error) | Odds Ratio |
| Intercept 1 | -1.339 (1.65) | 0.262 | 1.312 (1.56) | 3.713 | 1.114 (1.44) | 3.046 |
| Intercept 2 | 1.858 (1.66) | 6.410 | -1.588 (1.56) | 0.204 | -1.523 (1.45) | 0.218 |
| Component 1 | 0.848*** (0.10) | 2.335 | 0.795*** (0.10) | 2.215 | 0.740** (0.10) | 2.111 |
| Component 2 | 0.045 (0.08) | 1.046 | 0.044 (0.08) | 1.045 | 0.042 (0.07) | 1.043 |
| Component 3 | 0.148 (0.10) | 1.159 | 0.137 (0.09) | 1.147 | 0.144 (0.08) | 1.155 |
| Age | -0.045* (0.02) | 0.955 | -0.042* (0.02) | 0.957 | -0.038* (0.01) | 0.961 |
| Duration | -0.028 (0.02) | 0.971 | -0.029 (0.02) | 0.970 | -0.027 (0.02) | 0.972 |
| BMI | 0.063** (0.02) | 1.064 | 0.061** (0.02) | 1.062 | 0.050** (0.01) | 1.056 |
| Sex | 0.727** (0.25) | 2.069 | 0.709** (0.23) | 2.033 | 0.650** (0.21) | 1.910 |
| Metformin | 0.690* (0.35) | 1.994 | 0.632 (0.33) | 1.881 | 0.573* (0.30) | 1.774 |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

### Assumptions of the Proportion Odds Model

We assessed the proportional odds assumption for the three ordered logit models using the Brant and Wald tests. According to the Brant test, all the covariates satisfy the proportional odds assumption except component 1 ( see Table 3.7). Similarly, the Wald test indicates that all covariates, except component 1 and body mass index, satisfy the proportionality assumption. This implies that the effects of component 1 and body mass index vary across stages of kidney disease progression.

Table 3.7: Results of the test of proportionality via the Brant test and the Wald test for the three models.

| | **Proportional Odds Model** | | **Adjacent Category Model** | | **Continuation Ratio Model** | |
|---|---|---|---|---|---|---|
| **Covariates** | Chi-square | $p$-value | Wald value | $p$-value | Wald value | $p$-value |
| Component 1 | 3.97 | 0.040 | 54.84 | 0.0000 | 58.10 | 0.0000 |
| Component 2 | 0.01 | 0.940 | 0.33 | 0.5630 | 0.17 | 0.5760 |
| Component 3 | 0.21 | 0.650 | 2.82 | 0.0930 | 2.58 | 0.1360 |
| Age | 0.62 | 0.430 | 5.08 | 0.2450 | 5.41 | 0.2040 |
| Duration | 0.08 | 0.780 | 3.33 | 0.0680 | 2.00 | 0.3600 |
| BMI | 0.02 | 0.960 | 7.79 | 0.0052 | 8.79 | 0.0039 |
| Sex | 0.13 | 0.720 | 9.04 | 0.1236 | 5.98 | 0.2360 |
| Metformin | 0.87 | 0.350 | 3.52 | 0.0604 | 4.73 | 0.0530 |

The lack of proportionality across the stages of kidney function rendered the proportional odds, continuation ratio, and adjacent category models unsuitable for analyzing the effects of component 1 and body mass index on kidney disease progression. To address problems with proportionality and improve effect estimates, we used flexible ordered logit models, including partial cumulative odds, partial adjacent-category, and partial continuation-ratio models. If all the covariates violate the proportional odds assumption, we might use the generalized ordered logit model or another logit model, depending on our research questions. Therefore, we excluded the generalized ordered logit model from further analysis because it does not violate the parallel assumption across all covariates.

Table 3.8: Summary of estimates, standard errors in brackets, and odds ratios of the partial proportional odds, partial continuation ratio, and partial adjacent category model.

| | **Partial Proportional Odds Model** | | **Partial Continuation Ratio Model** | | **Partial Adjacent Category Model** | |
|---|---|---|---|---|---|---|
| **Covariates** | Estimate (Std. Error) | Odds Ratio | Estimate (Std. Error) | Odds Ratio | Estimate (Std. Error) | Odds Ratio |
| Intercept 1 | -1.511 (1.67) | 0.220 | 1.393 (1.70) | 4.026 | 1.006 (1.61) | 2.734 |
| Intercept 2 | 1.541 (1.68) | 4.669 | -0.874 (1.74) | 0.417 | -1.015 (1.64) | 0.362 |
| Component 1:1 | -0.964*** (0.12) | 0.381 | 0.990*** (0.10) | 2.708 | 0.940*** (0.13) | 2.575 |
| Component 1:2 | -0.582* (0.15) | 0.558 | 0.325 (0.17) | 1.384 | 0.360* (0.17) | 1.433 |
| Component 2 | -0.042 (0.09) | 0.958 | 0.040 (0.08) | 1.041 | 0.036 (0.08) | 1.037 |
| Component 3 | -0.144 (0.10) | 0.865 | 0.127 (0.09) | 1.136 | 0.136 (0.09) | 1.146 |
| Age | 0.048* (0.02) | 1.049 | -0.045* (0.01) | 0.955 | -0.038* (0.02) | 0.962 |
| Duration | 0.030 (0.02) | 1.031 | -0.031 (0.02) | 0.969 | -0.029 (0.02) | 0.970 |
| BMI 1:1 | -0.065** (0.02) | 0.936 | 0.066* (0.02) | 1.068 | 0.059* (0.03) | 1.061 |
| BMI 1:2 | | | 0.056 (0.03) | 1.058 | 0.048 (0.03) | 1.044 |
| Sex | -0.732** (0.24) | 0.481 | 0.690** (0.23) | 1.994 | 0.626** (0.21) | 1.871 |
| Metformin | -0.713* (0.35) | 0.490 | 0.645 (0.33) | 1.906 | 0.582* (0.31) | 1.791 |

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

In Table 3.8, we can see that age, sex, body mass index, component 1 (a linear combination of HbA1c, glycemia, and creatinine), and metformin all have a significant effect on the progression of kidney disease in all three of the partially ordered logit models. Conversely, we found no significant impact at the 5% significance level for diabetes duration, component 2 (a linear combination of TC and LDL), and component 3 (a linear combination of HDL and TG). The odds ratio indicates the effect of clinical and biochemical variables on the progression of kidney disease in people with type 2 diabetes. The partial proportional odds model uses metformin as an example; the odds ratio of 0.490 shows that diabetic individuals who take metformin are almost 50% less likely to have mildly or moderately reduced kidney function while all other predictors remain the same. An increase of one year in the age of diabetic people leads to a 4.9% rise in the odds of mildly reduced kidney function or moderately reduced kidney function. As their body mass index increased by one unit, their chances of having moderately reduced kidney function or developing diabetic kidney disease decreased by 6.4%.

Component 1 (a linear combination of HbA1c, glycemia, and creatinine) had a significant

effect on the progression of kidney function in individuals with T2D (P value = 0.000). Unlike other covariates, the impacts of Component 1 varied significantly across different levels of kidney function. For example, in the partial adjacent category model, an odds ratio of 2.575 indicates that a one-unit increase in Component 1 decreased the likelihood of mildly reduced kidney function by 2.575 times. Conversely, an odds ratio of 1.433 indicates that a one-unit increase in Component 1 decreased the possibility of moderately reduced kidney function by 1.5 times.

From the principal component analysis, we observe that Component 1 is primarily a linear combination of three variables: HbA1c, glycemia, and creatinine, with a negative relationship. In other words, a decrease in HbA1c, glycemia, and creatinine results in an increase in Component 1, and vice versa. This suggests that higher levels of HbA1c, glycemia, and creatinine increase the risk of declining kidney function in diabetic individuals.

To select the model that best fits the data, the deviance value, Akaike information criterion (AIC), and Bayesian information criterion (BIC) are applied. Note that these criteria aim to facilitate comparisons between models that fit the data reasonably well, not to identify the 'correct' model. Table 3.9 presents the deviance, AIC, and BIC values for the six types of ordered logit models. The results indicate that the partial adjacent category model is preferred, as it has the smallest deviance, AIC, and BIC values among the models.

Table 3.9: Comparison of the fitted models.

| **Model** | **Deviance value** | **AIC** | **BIC** |
|---|---|---|---|
| Proportional Odds model | 514.3475 | 538.3475 | 583.6793 |
| Continuation ratio model | 511.9666 | 535.9666 | 581.2984 |
| Continuation ratio model | 514.7567 | 538.7567 | 584.0885 |
| Partial Proportional odds oodel | 509.5034 | 535.5034 | 584.6128 |
| Partial continuation ratio model | 508.9252 | 534.9252 | 584.0347 |
| Partial adjacent category model | 508.0683 | 534.0683 | 583.1778 |

### 3.4.4 Discussion

Diabetic kidney disease is one of the most chronic complications in people with T2D and the leading cause of end-stage kidney disease (kidney failure). In this retrospective cross-sectional cohort study, 85.8% of elderly individuals with type 2 diabetes had either mildly reduced kidney function (eGFR, 60 to 89 mL/min/1.73 $m^2$) or moderately reduced kidney function (eGFR, 30 to 59 mL/min/1.73 $m^2$). There are modifiable and non-modifiable clinical and biochemical factors that affect the development and progression of diabetic kidney disease in people with type 2 diabetes. This study used principal component analysis and an ordered logit model to investigate the effects of these factors on the progression of kidney function in individuals with T2D.

We found a strong collinearity between some of the biochemical variables. This collinearity increases the standard errors of the regression coefficients, resulting in inaccurate effect estimates. To mitigate collinearity and improve model accuracy, we applied PCA to these variables before the regression. Principal component analysis reduces the correlated risk factors for kidney function decline into three uncorrelated principal components. These three principal components explained 76% of the variation in the biochemical variables. For example, the first principal component explained 24% of the variation in biochemical variables related to glycaemic markers (HbA1c, glycemia, and creatinine). The second

principal component explained 22% of the variation in the biochemical variables related to total cholesterol and LDL. The third principal component accounted for 21% of the variation in biochemical variables associated with HDL and triglycerides.

Our results were consistent with earlier research that used principal component analysis to cluster the risk factors for chronic disease. For example, (Hillier et al., 2006; Stuckey et al., 2014; Tsai et al., 2020) used principal components to investigate the effect of metabolic syndrome on the occurrence of diabetes and cardiovascular disease. Their research revealed that TG and HDL were loaded onto one principal component, whereas total cholesterol and LDL were loaded onto another, consistent with our study.

Despite the development of various ordered logit models, medical researchers often use the proportional odds model to examine the effects of predictors on the order-level outcome variable. When one or more predictor variables violated the assumption of proportional odds, this approach led to biased statistical inferences (Bender and Grouven, 1998). The empirical results of the Brant and Wald tests in our study also supported this; for instance, component 1 failed to meet the proportional odds assumption, indicating that it does not have the same effect across the three stages of kidney function. As a result, conclusions about how component 1 affects kidney function progression via the proportional odds model, continuation ratio model, or adjacent category model are biased.

To mitigate the bias introduced by proportionality, we employed alternative ordered logit models, including the partial proportional odds model, the partial continuation ratio model, and the partial adjacent category model. The choice among these models depends on the specific research question. For instance, the partial continuation ratio model or the partial adjacent category model is preferable for predicting the probability that diabetic individuals have moderately reduced kidney function compared to those with mildly reduced kidney function. Conversely, if the interest lies in predicting the probability of diabetic individuals having moderately reduced kidney function relative to other groups (normal kidney function and mildly reduced kidney function), the partial proportional odds model would be the preferred choice.

Studies have noted that HbA1c has a significant impact on the progression of kidney function (Ceriello et al., 2017; Gao et al., 2022; Russo et al., 2018; Yu et al., 2019). Our research highlight the composite effect of HbA1c, creatinine, and glycemia on the development of diabetic kidney disease. The results show that component 1 (glycaemic markers) is the primary and most significant risk factor for the progression of kidney disease. These results imply that as HbA1c, creatinine, and glycemia increase collectively, the risk of developing diabetic kidney disease is also high. These biochemical variables are modifiable, and people can prevent the progression of diabetic complications by controlling their HbA1c, creatinine, and glycemia levels.

Our study revealed that age and sex differences were significantly related to the progression of kidney disease in people with T2D. This result is supported by (Joshi et al., 2023) and (Russo et al., 2018), who reported that age significantly affects the progression of kidney disease, as older age is associated with a higher risk of chronic disease. Women are more likely to experience moderately decreased kidney function or diabetic kidney disease. The results suggest that practitioners should provide special attention to older individuals, who should also manage their glucose levels and properly take all necessary treatments to prevent diabetic kidney disease.

Researchers have identified the duration of type 2 diabetes as a risk factor for the incidence and progression of diabetic kidney disease (Siddiqui et al., 2022). Our study was unable to confirm this finding in the ordered logit models at the 5% level of significance, despite this risk factor for the incidence and progression of diabetic kidney disease. Additionally, lipid profiles (LDL, HDL, triglycerides, and total cholesterol) are directly associated with diabetic kidney disease, even though the associations were not statistically significant at the 5% level, which is inconsistent with some of the literature (Yun et al., 2016). Possible reasons include the sample size, the number and type of covariates included in the model, the use of elderly diabetic participants, or the methodology we applied. Based on the results, we recommend that healthcare professionals pay greater attention to the proper management of biochemical variables, such as HbA1c, creatinine, glycemia, LDL, HDL,

triglycerides and total cholesterol, which can cause DKD in people with T2D.

## 3.5 Conclusion

The study results indicate that 85.8% of individuals with T2D have either stage II kidney function (eGFR, 60 to 89 mL/min/1.73 $m^2$) or diabetic kidney disease (eGFR, 30 to 59 mL/min/1.73 $m^2$). Principal component analysis reduced the correlated biochemical variables into three uncorrelated components, with component one, a linear combination of HbA1c, glycemia, and creatinine, showing a strong effect on the progression of kidney disease. These biochemical variables are modified, and people with diabetes can prevent complications by managing their HbA1c, creatinine, and glycemia levels. The effects of clinical and biochemical variables differ across various stages of disease progression. Flexible ordered logit models, including partial proportional odds, cumulative odds, adjacent-category models, and continuation-ratio models, are appropriate to provide more precise and unbiased results.

# Bibliography

Adiwijaya, W. U., Lisnawati, E., Aditsania, A., Kusumo, D. S., et al. (2018). Dimensionality reduction using principal component analysis for cancer detection based on microarray data classification. *Journal of Computer Science*, 14(11):1521–1530.

Alicic, R. Z., Rooney, M. T., and Tuttle, K. R. (2017). Diabetic kidney disease: challenges, progress, and possibilities. *Clinical journal of the American Society of Nephrology*, 12(12):2032–2045.

Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333.

Bender, R. and Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of clinical epidemiology*, 51(10):809–816.

Çamdevỳren, H., Demỳr, N., Kanik, A., and Keskỳn, S. (2005). Use of principal component scores in multiple linear regression models for prediction of chlorophyll-a in reservoirs. *Ecological Modelling*, 181(4):581–589.

Ceriello, A., De Cosmo, S., Rossi, M. C., Lucisano, G., Genovese, S., Pontremoli, R., Fioretto, P., Giorda, C., Pacilli, A., Viazzi, F., et al. (2017). Variability in hba1c, blood pressure, lipid parameters and serum uric acid, and risk of development of chronic kidney disease in type 2 diabetes. *Diabetes, Obesity and Metabolism*, 19(11):1570–1578.

De Boer, I. H. and Steffes, M. W. (2007). Glomerular filtration rate and albuminuria: twin manifestations of nephropathy in diabetes. *Journal of the American Society of Nephrology*, 18(4):1036–1037.

Gao, M., Zhong, Z., Yue, Y., and Liu, F. (2022). Correlation between glycaemic variability and prognosis in diabetic patients with ckd. *Endokrynologia Polska*, 73(6):947–953.

Goodman, L. A. (1983). The analysis of dependence in cross-classifications having ordered categories, using log-linear models for frequencies and log-linear models for odds. *Biometrics*, pages 149–160.

Gregorich, M., Heinzel, A., Kammer, M., Meiselbach, H., Böger, C., Eckardt, K.-U., Mayer, G., Heinze, G., and Oberbauer, R. (2021). A prediction model for the decline in renal function in people with type 2 diabetes mellitus: study protocol. *Diagnostic and Prognostic Research*, 5(1):19.

He, A., Shi, C., Wu, X., Sheng, Y., Zhu, X., Yang, J., and Zhou, Y. (2024). Clusters of body fat and nutritional parameters are strongly associated with diabetic kidney disease in adults with type 2 diabetes. *Diabetes Therapy*, 15(1):201–214.

Hillier, T. A., Rousseau, A., Lange, C., Lepinay, P., Cailleau, M., Novak, M., Calliez, E., Ducimetière, P., and Balkau, B. (2006). Practical way to assess metabolic syndrome using a continuous score obtained from principal components analysis: The desir cohort. *Diabetologia*, 49(7):1528–1535.

Hsu, Y.-L., Huang, P.-Y., and Chen, D.-T. (2014). Sparse principal component analysis in cancer research. *Translational cancer research*, 3(3):182.

Joshi, R., Subedi, P., Yadav, G. K., Khadka, S., Rijal, T., Amgain, K., and Rajbhandari, S. (2023). Prevalence and risk factors of chronic kidney disease among patients with type 2 diabetes mellitus at a tertiary care hospital in nepal: a cross-sectional study. *BMJ open*, 13(2):e067238.

Kristono, G. A., Holley, A. S., Hally, K. E., Brunton-O'Sullivan, M. M., Shi, B., Harding, S. A., and Larsen, P. D. (2020). An il-6-il-8 score derived from principal component analysis is predictive of adverse outcome in acute myocardial infarction. *Cytokine: X*, 2(4):100037.

Liu, X. Z., Duan, M., Huang, H. D., Zhang, Y., Xiang, T. Y., Niu, W. c., Zhou, B., Wang, H. L., and Zhang, T. T. (2023). Predicting diabetic kidney disease for type 2

diabetes mellitus by machine learning in the real world: a multicenter retrospective study. *Frontiers in Endocrinology*, 14:1184190.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Milewska, A. J., Jankowska, D., Citko, D., Więsak, T., Acacio, B., and Milewski, R. (2014). The use of principal component analysis and logistic regression in prediction of infertility treatment outcome. *Studies in Logic, Grammar and Rhetoric. Logical, Statistical and Computer Methods in Medicine*, 39(52):7–23.

Okin, P. M., Devereux, R. B., Fabsitz, R. R., Lee, E. T., Galloway, J. M., and Howard, B. V. (2002). Principal component analysis of the t wave and prediction of cardiovascular mortality in american indians: the strong heart study. *Circulation*, 105(6):714–719.

Peterson, B. and Harrell Jr, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(2):205–217.

Russo, G. T., De Cosmo, S., Viazzi, F., Mirijello, A., Ceriello, A., Guida, P., Giorda, C., Cucinotta, D., Pontremoli, R., Fioretto, P., et al. (2018). Diabetic kidney disease in the elderly: prevalence and clinical correlates. *BMC geriatrics*, 18(1):38.

Russo, G. T., De Cosmo, S., Viazzi, F., Pacilli, A., Ceriello, A., Genovese, S., Guida, P., Giorda, C., Cucinotta, D., Pontremoli, R., et al. (2016). Plasma triglycerides and hdl-c levels predict the development of diabetic kidney disease in subjects with type 2 diabetes: the amd annals initiative. *Diabetes care*, 39(12):2278–2287.

Shrout, P. E. (1979). Book review: The analysis of cross-classified categorical data stephen fienberg, cambridge, ma: Mit press, 1977. *Applied Psychological Measurement*, 3(2):275–277.

Siddiqui, K., George, T. P., Joy, S. S., and Alfadda, A. A. (2022). Risk factors of chronic kidney disease among type 2 diabetic patients with longer duration of diabetes. *Frontiers in Endocrinology*, 13:1079725.

Stuckey, B. G., Opie, N., Cussons, A. J., Watts, G. F., and Burke, V. (2014). Clustering of metabolic and cardiovascular risk factors in the polycystic ovary syndrome: a principal component analysis. *Metabolism*, 63(8):1071–1077.

Tan, J., Zwi, L. J., Collins, J. F., Marshall, M. R., and Cundy, T. (2017). Presentation, pathology and prognosis of renal disease in type 2 diabetes. *BMJ Open Diabetes Research & Care*, 5(1).

Tsai, T.-Y., Hsu, P.-F., Lin, C.-C., Wang, Y.-J., Ding, Y.-Z., Liou, T.-L., Wang, Y.-W., Huang, S.-S., Chan, W.-L., Lin, S.-J., et al. (2020). Factor analysis for the clustering of cardiometabolic risk factors and sedentary behavior, a cross-sectional study. *Plos one*, 15(11):e0242365.

Tziomalos, K. and Athyros, V. G. (2015). Diabetic nephropathy: new risk factors and improvements in diagnosis. *The review of diabetic studies: RDS*, 12(1-2):110.

Ye, W., Ding, X., Putnam, N., Farej, R., Singh, R., Wang, D., Kuo, S., Kong, S. X., Elliott, J. C., Lott, J., et al. (2022). Development of clinical prediction models for renal and cardiovascular outcomes and mortality in patients with type 2 diabetes and chronic kidney disease using time-varying predictors. *Journal of Diabetes and its Complications*, 36(5):108180.

Yu, Z.-B., Wang, J.-B., Li, D., Chen, X.-Y., Lin, H.-B., and Chen, K. (2019). Prognostic value of visit-to-visit systolic blood pressure variability related to diabetic kidney disease among patients with type 2 diabetes. *Journal of hypertension*, 37(7):1411–1418.

Yun, K.-J., Kim, H. J., Kim, M. K., Kwon, H.-S., Baek, K.-H., Roh, Y. J., and Song, K.-H. (2016). Risk factors for the development and progression of diabetic kidney disease in patients with type 2 diabetes mellitus and advanced diabetic retinopathy. *Diabetes & metabolism journal*, 40(6):473.

Zhang, Z. and Castelló, A. (2017). Principal components analysis in clinical studies. *Annals of translational medicine*, 5(17):351.

Zwick, W. R. and Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432.

Convergence diagnostic checking results

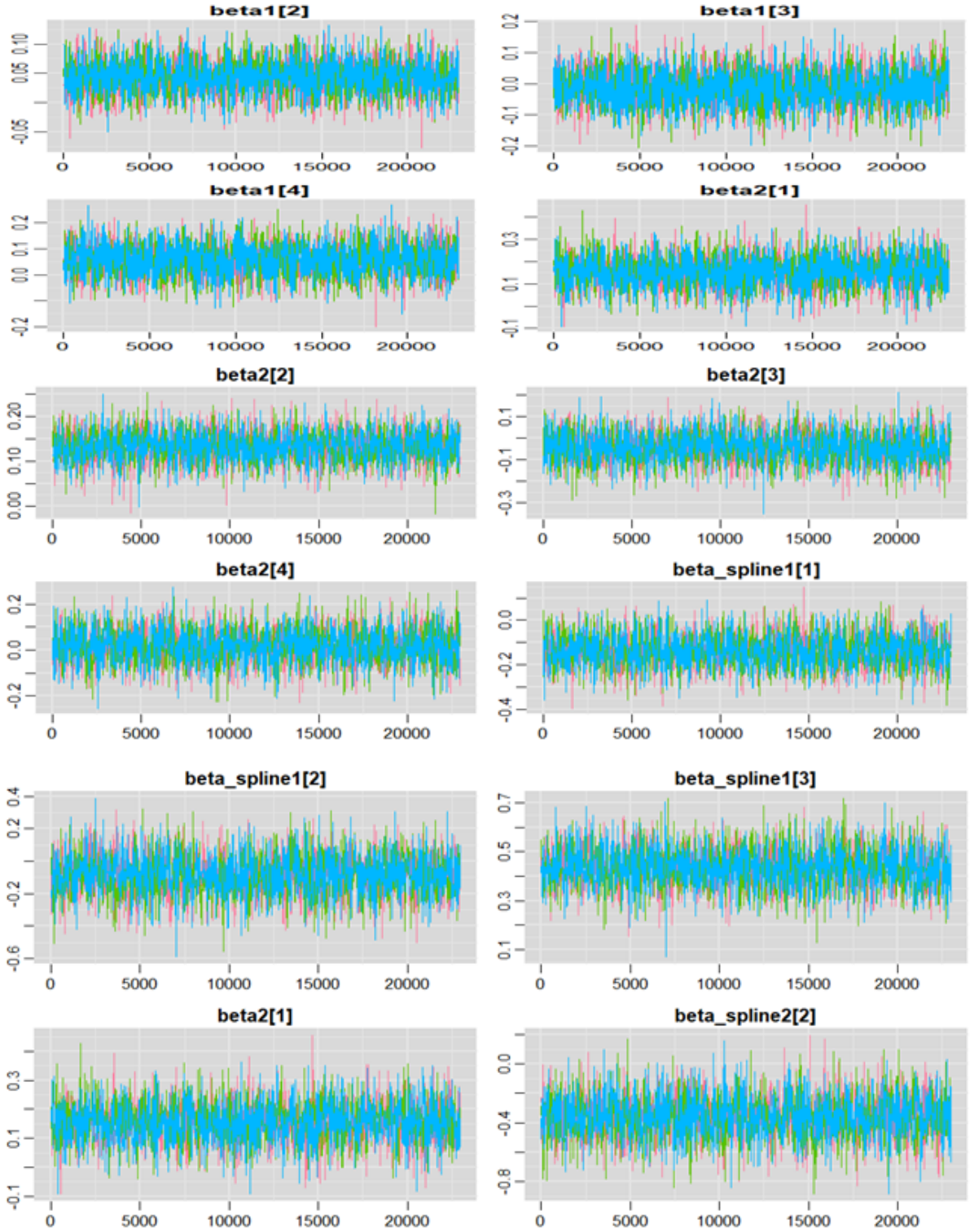
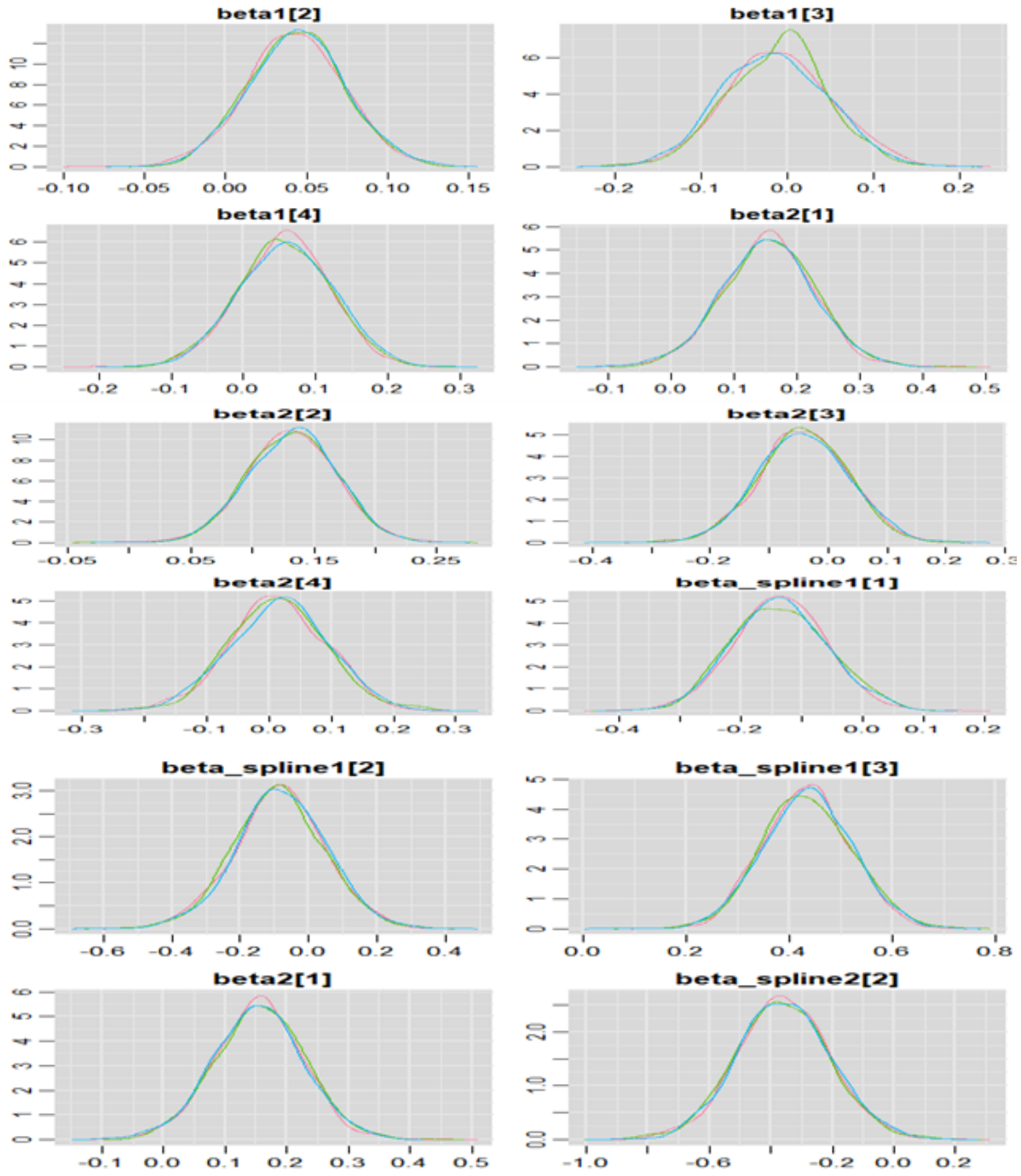
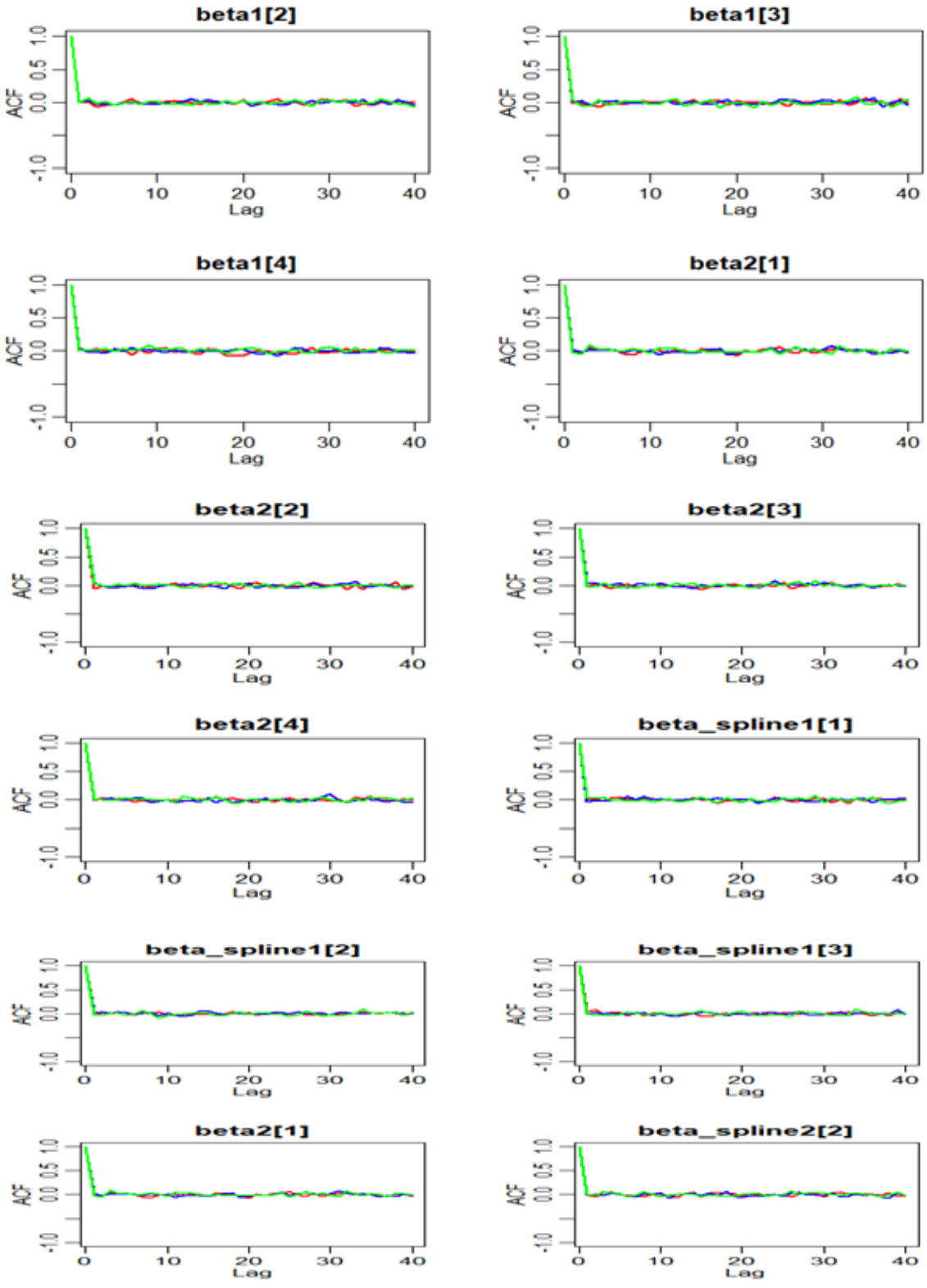



Figure 5: Appendix plot




Density plots of all parameters from the chosen model except the model error and the random effects.




Autocorrelation plots all parameters from the chosen model except the model error and the random effects.