

Joint Modeling of Longitudinal and Time-to-Event Data with the Application to Kidney Transplant Data

by

Jianghu Dong

M.Sc.in Statistics, University of Alberta, 2005

M.Sc.in Statistics, Renmin University of China, 2003

B.Sc.in Math, Beijing Normal University.1997

Thesis Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in the

Department of Statistics and Actuarial Science

Faculty of Mad Science

© Jianghu Dong 2018

SIMON FRASER UNIVERSITY

Dec 2018

Copyright in this work rests with the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Approval

Name: **Jianghu Dong**

Degree: **Doctor of Philosophy (Statistics)**

Title: **Joint Modeling of Longitudinal and Time-to-Event Data with the Application to Kidney Transplant Data**

Examining Committee: **Chair:** Lloyd T. Elliott
Assistant Professor

Jiguo Cao
Senior Supervisor
Associate Professor

Liangliang Wang
Supervisor
Assistant Professor

Jagbir Gill
Internal Examiner
Associate Professor
Department of Medicine
University of British Columbia

Jun Yan
External Examiner
Professor
Department of Statistics
University of Connecticut, USA

Date Defended: **Dec 13, 2018**

Abstract

The main thesis develops the novel and powerful statistical methodology in functional principal component analysis and joint models with the application to solve the problems in kidney transplant. This thesis can be divided broadly into five parts.

Firstly, we use functional principal component analysis (FPCA) through conditional expectation to explore major sources of variations of GFR curves. The estimated FPC scores can be used to cluster GFR curves. Ordering FPC scores can detect abnormal GFR curves. FPCA can effectively estimate missing GFR values and predict GFR values. Secondly, we propose new joint models with mixed-effect and Accelerated Failure Time (AFT) submodels, where the piecewise linear function is used to calculate the non-proportional dynamic hazard ratio curve of a time-dependent side event. The finite sample performance of the proposed method is investigated in simulation studies. Our method is demonstrated by fitting the joint model for some clinical kidney data. Thirdly, we develop a joint model with FPCA and multi-state model to fit the longitudinal and multiple time-to event outcomes together. FPCA is efficient in reducing the dimensions of the longitudinal trajectories. Multistate submodel can be used to describe the dynamic process of multiple time-to-event outcomes. The relationships between the longitudinal and time-to-event outcomes can be assessed based on the shared latent features. The latent variables FPC scores are significantly related to time-to-event outcomes in the application example, and Cox model may cause bias for multiple time-to event outcomes compared with multi-state model. Fourthly, we develop a flexible class joint model of generalized linear latent variables for multivariate responses, which has an underlying Gaussian latent processes. The model accommodates any mixture of outcomes from the exponential family. Monte Carlo EM is proposed for parameter estimation and the variance components of the latent processes. We demonstrate this methodology by kidney transplant studies. Finally, in many social and health studies, measurement of some covariates are only available from units of subjects, rather than from individual. Such kind of measures are referred as to aggregate average exposures. The current method fails to evaluate high-order or nonlinear effect of aggregated exposures. Therefore, we develop a nonparametric method based on local linear fitting to overcome the difficulty. We demonstrate this methodology by kidney transplant studies.

Finally, future work

Keywords: Functional Data Analysis and FPCA; Accelerated Failure Time; Missing data and Outlier; Latent Features; Joint modelling; Kidney Transplant; GFR Trajectory

Dedication

I am grateful for the support in various forms from my family: my wife, Wencong Wang, for her understanding and support while I was working to complete this dissertation, my parents-Mr. Yuexing Dong and Mrs. Yuehong Pan, and my old brother-Mr.Jiangwen Dong, my young sister-Mrs. Jianghua Dong.

Acknowledgements

I am deeply grateful for my supervisors Dr. Jiguo Cao and Dr. Liangliang Wang for their inspirational instruction, tremendous supports and invaluable guidance. They give me many wonderful real academic and life supports. We always discover so many interesting statistical problems during weekly meeting. This dissertation has benefited from their insights and intellectual acumen.

I am deeply grateful to Dr. Jagbir Gill at University of British Columbia for being my mentor in both academia and real life, his invaluable advice, and his wonderful instruction in medical knowledge. My sincere thanks are also extended to other members of my examining committee. I would like to sincerely thank Professor Dr. Jun Yan of Department of Statistics at University of Connecticut for taking time from his busy schedule to serve as my external examiner, and Dr. Lloyd T. Elliott for his kindness to chair my defence.

I am deeply grateful for all wonderful supports from my previous supervisors Dr.Scott Klarenbach at University of Alberta, Dr. Peng Zhang at Zhejiang University, and Dr. Xi Chen at University of Alberta. I would like to sincerely thank Dr. Xi Chen for his encouraging me to take on this PhD study.

I believe that the research and teaching experience I have gained during my graduate study with them has had a profound influence and impact upon my academic career.

Last but not least, I would like to thank all of those who have supported and helped me throughout my Ph.D studies at Simon Fraser University.

Table of Contents

Approval	ii
Abstract	iii
Dedication	v
Acknowledgements	vi
Table of Contents	vii
List of Tables	xi
List of Figures	xii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Methods for longitudinal data	4
1.2.1 Functional principal component analysis	4
1.2.2 Parameter models for longitudinal data	6
1.3 Survival analysis	8
1.3.1 Types of censoring	8
1.3.2 The Accelerated Failure Time model (AFT model)	9
1.3.3 Multi-state survival models	10
1.4 Methods for joint modelling the longitudinal and multiple time-to event outcomes	12
1.5 Algorithm Material on Monte Carlo	13
1.5.1 Gibbs Sampler	13
1.5.2 Importance Sampling	14
1.5.3 Laplace Approximation	14
1.6 Outline of this dissertation	15
2 Functional Principal Component Analysis of GFR Curves after Kidney Transplant	16

16	section.2.1	
2.2	Methods	19
2.2.1	Functional Principal Component Analysis	19
2.2.2	Clustering	21
2.2.3	Detection of GFR trajectory outliers	22
2.2.4	Prediction for Future GFR	22
2.3	Results on Kidney Transplant Data	22
2.3.1	Functional Principal Component Analysis	22
2.3.2	Patient clustering	25
2.3.3	Detection of GFR trajectory outliers	25
2.3.4	Prediction	26
2.4	Conclusions and Discussion	26
3	A Joint Model of a Longitudinal and Accelerated Failure Time Data and its Application to Transplant Patients with an ESRD and a Diabetes	33
33	section.3.1	
3.2	The Joint Model	36
3.2.1	The Survival Submodel	37
3.3	Estimation Method	40
3.3.1	E-step	41
3.3.2	M-step	42
3.4	Application to Clinical Transplant Data	42
3.4.1	Main Results from the Joint Model	42
3.4.2	Effect of Pancreas Transplant on Allograft	45
3.5	Simulations	47
3.5.1	Simulation 1	47
3.5.2	Simulation 2	48
3.5.3	Simulation 3	50
3.6	Conclusions and Discussion	51
4	Jointly Modelling Multiple Outcomes by Functional Principal Component Analysis via a Multistate Model	52
4.1	Introduction	52
4.2	A Joint Model	55
4.2.1	Functional Principal Component Analysis	56
4.2.2	Multi-state models	56
4.3	The Estimation Method	58
4.3.1	The joint likelihood functions	58
4.3.2	Parameter estimation	60
4.4	The application of the proposed joint model	61

4.4.1	Results from Functional Principal Component Analysis	63
4.4.2	Results from multi-state submodel	63
4.5	Simulations	64
4.6	Conclusions and discussion	69
5	Jointly Modelling Multiple Continuous and Discrete Outcomes by a Flexible Class of Generalized Linear Latent Variable Models	71
5.1	Introduction	71
5.2	Model Specification	72
5.3	The covariance structure of the latent variables	73
5.3.1	The autoregressive structure in time series frame	74
5.4	The joint likelihood function	76
5.4.1	Monte Carlo EM	77
5.4.2	Information Matrix	80
5.5	The application to Clinical Transplant Data	80
5.5.1	Model specification in the application example	80
5.5.2	Model results	81
5.6	Conclusion	81
6	A Predict Model with a Polynomial Effects Covariate in Presence of Measurement Errors	83
6.1	Motivation	83
6.2	Introduction	83
6.3	Model Specification	84
6.4	Method	85
6.5	Local linear fitting approach	86
6.6	Statistical inference	88
6.7	Simulation	90
6.7.1	Simulation setting	90
6.7.2	Simulation results	90
6.8	The application to kidney transplant data	92
6.8.1	Data Resource	92
6.8.2	Model specification	92
6.8.3	Model results	96
6.9	Conclusion	98
7	Future works	99
7.1	Motivation	99
7.2	Current work and future research	99
7.2.1	Functional data analysis (FDA)	99

7.2.2	Joint modeling	100
7.2.3	Measurement error models	101
7.2.4	Cost-effectiveness analysis	102
Bibliography		103
Appendix A Supplementary material for Functional Principal Component Analysis of GFR Curves after Kidney Transplant		111
Appendix B Supplementary material for A Joint model of a longitudinal and Accelerated Failure Time data and its application to transplant patients with an ESRD and a diabetes		114
B.1	Monte Carlo EM algorithm	114
B.1.1	M-step	114
B.2	The result from simulation 1 when N=500	115

List of Tables

Table 3.1	Estimates for parameters in Model (5.1). The standard errors of the estimates are given in brackets.	44
Table 3.2	Means, biases, root mean square errors (RMSEs) of the parameter estimates for the joint model (5.1) using our proposed MCEM algorithm in Simulation 1.	48
Table 3.3	Means and standard deviations (STD) of the parameter estimates for our proposed joint model (3.1) and the model (3.8) in Simulation 2. .	49
Table 3.4	The mean, bias, standard deviation (STD), and root mean squared error (RMSE) of the parameter estimates for the joint model (3.1) when the model assumption is correct or misspecified in Simulation 3.	51
Table 4.1	Kidney transplanted recipient characteristics in some kidney transplant data	62
Table 4.2	Estimated hazard ratios of kidney failure post kidney transplant in the joint model with different survival sub-models. 95% confidence interval are given in brackets.	65
Table 4.3	Estimated hazard ratios of death post kidney transplant from different survival sub-models and 95% confidence interval are given in brackets.	66
Table 4.4	Means and standard deviations (STD) in three different scenarios. Each scenario has 100 simulation replicates and 100 subjects in each simulation replicate	69
Table 5.1	Estimates for parameters in Model (5.14). The standard errors of the estimates are given in brackets.	82
Table B.1	Mean, bias, RMSE of the parameter estimates using our proposed MCEM algorithm for Model using 100 simulation replicates in the first simulation study ($N = 500$).	116

List of Figures

Figure 1.1	Kidney transplantation	2
Figure 1.2	Kidney function progression	3
Figure 1.3	The three states of kidney transplant recipients. All patients start from the date of the kidney transplant (state 1), then they may move to state 2 (kidney failure). If not, they directly move to state 3 when die	11
Figure 2.1	Observed GFR trajectory curves with various circumstances and trends. Patients in the upper left panel (a) have missing data records. Patients in the upper right panel (b) have flat GFR trends. Patients in the lower left panel (c) have strong fluctuating trends. Patients in the lower right panel (d) have increasing or decreasing trends. Each color represents one individual patient in each panel.	17
Figure 2.2	The mean curve of GFR in the left panel and the correlation function of GFR in the right panel. They are estimated from the total patients.	23
Figure 2.3	The first four leading functional principal components (FPCs) estimated from the GFR curves.	28
Figure 2.4	GFR curves when their FPC scores are extreme. The thick blue curve in each panel is the average of individual GFR curves in that panel, which represents the common trend in that panel. The left four panels, from top to bottom, are GFR curves when their first, second, third, and fourth FPC scores are smaller than the 5% quantiles, respectively. The right four panels, from top to bottom, are GFR curves when their first, second, third, and fourth FPC scores are larger than the 95% quantiles, respectively.	29
Figure 2.5	Part of the GFR curves in six clusters.	30
Figure 2.6	Some abnormal GFR curves.	31
Figure 2.7	The predicted GFR curves for four patients. The dots are observed GFR data.	32

Figure 3.1	Observed individual GFR trajectory curves. The left two panels, from top to bottom, are GFR curves for patients with All-cause graft loss (ACGL) events or without ACGL events, respectively, when they don't have a pancreas transplantation. The right two panels, from top to bottom, are GFR curves for patients with ACGL event or without ACGL events, respectively, when they have a pancreas transplantation. Each color represents the individual patient in each panel.	35
Figure 3.2	The three statuses of kidney transplant patients. All patients start from the date of the kidney transplant (Status 1), then they may move to Status 2 (pancreas transplantation) when a matched pancreas organ is available during the followed-up time period. If not, they directly move to Status 3 when the time-to-event outcome of all-cause graft loss happens, or they still are on the waiting-list for the pancreas transplant.	38
Figure 3.3	The cumulative Nelson-Aalen estimate of all-cause graft loss by patient status of pancreas transplantation. The red line is the cumulative Nelson-Aalen estimate of all-cause graft loss for patient with a pancreas transplant and the blue line is the cumulative Nelson-Aalen estimate of all-cause graft loss for patient without a pancreas transplant.	39
Figure 3.4	The curve of hazard ratios of all-cause graft loss for patients with a pancreas transplant with the 95% confidence intervals at 14, 45, 90, 152, 180, 365, 730 days from the date of pancreas transplant. The reference group are patients without a pancreas transplant. The hazard Ratio curve reaches 1.00 at 152 days from the date of pancreas transplant.	46
Figure 4.1	GFR curves when their FPC scores are extreme. The thick blue curve in each panel is the average of individual GFR curves in that panel, which represents the common trend in that panel. The four panels are GFR trajectory curves are donated by their first, second, third, and fourth FPC scores respectively.	53
Figure 4.2	The three states of kidney transplant recipients. All patients start from the date of the kidney transplant (state 1), then they may move to state 2 (kidney failure). If not, they directly move to state 3 when die	57

Figure 4.3	The first two leading functional principal components (FPCs) account for 95.24% of the total variability of GFR curves, and the four leading FPCs account for 99.82%	67
Figure 4.4	The first four leading functional principal components (FPCs) estimated from the GFR curves.	68
Figure 6.1	Comparison of Models for Different Pooling Scenarios	93
Figure 6.2	Evolution Effect of Different Pooling Scenarios on Models	94
Figure 6.3	Evolution Effect of Different Pooling Scenarios on Models	95
Figure 6.4	Probability of kidney transplant versus logarithm of income. Solid line represents the curve with local logistic regression, dashed line is the curve of regular logistic regression and the dotted line is the fitted logistic regression with quadratic term.	97
Figure A.1	Part of GFR trajectory curves for the first 20 cluster groups	112
Figure A.2	Part of GFR trajectory curves for the last 20 cluster groups	113

Chapter 1

Introduction

1.1 Background and Motivation

The incidence and prevalence of end-stage renal disease (ESRD) is increasing worldwide. In Canada, N=37,457 patients live with ESRD, compared with N=594,000 in the United States in 2015. ESRD is an important public health problem due to the high cost of renal dialysis, a high mortality rate and a decreased quality of life. Kidney transplantation is the preferred treatment for ESRD. Kidney and pancreas transplant is the preferred treatment for patients with the type one diabetes and ESRD. The population of transplant survivors has been increasing rapidly as a result of advances in treatment, but the demand supply of organs is not sufficient to meet the increasing demand. There are three following strategies to address this problem.

1. Decrease the incidence of ESRD
2. Increase the number of deceased and living organ donors
3. Maximize the utility of the available organ supply

The first strategies have far been inadequate to narrow the gap between supply and demand so far; therefore the second and third strategies should be highlighted. Kidney donation has sustained kidney transplantation activity, but there is significant regional variability noted. Deceased/Living donor kidneys are routinely shared within seven geographically defined regions, but are infrequently shared between regions. For example, the donor rate per million population (RPMP) in 2004 varied from 6.0 in Manitoba to 18.0 in Quebec. The reasons for this variability remain unclear. The RPMP does not account for population differences between regions that may impact organ donation, making it difficult to determine if regional variation is due to differences in the number potential organ donors, differences in organ procurement practices, or differences in family consent rates for organ donation, community social work, household income, donor race, etc. A major barrier to understand regional differences in deceased organ donation is lack of an informative metric

of donor activity. Understanding why some regions have higher deceased/living donor rates than others will inform health policy to improve kidney donation in all regions. However, due the current unavailability of kidney donor dataset, we only focus on develop the new statistics model in the area of the third strategies about how to maximize the utility of the available organ supply, but I will continue the second strategy research in the future work.

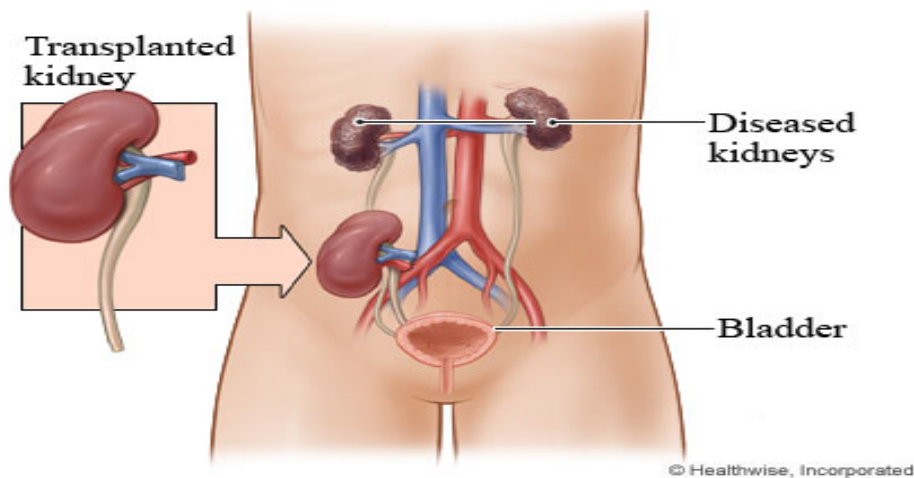


Figure 1.1: Kidney transplantation

It is known that the kidney transplantation as shown in 1.1 can prolong the survival of patients with end-stage renal disease in the papers by Levey et al.⁷⁵ and Wolfe et al.⁷⁶. However, how to extend the long-term survival of the kidney graft still remains the main challenge for transplant. Transplanted kidneys have a limited lifespan despite advancements in pharmaceuticals for the acute rejection, the long-term survival life time of the grafted kidney has not been increased. Kidney allograft failure after transplantation significantly adds to the demand for kidney transplantation. Therefore, it is important to identify clinical markers to predict the kidney allograft loss. If the kidney graft failure can be prevented, then we can maximize the utility of all available kidney organs recipients. Therefore, several






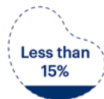
STAGES OF CHRONIC KIDNEY DISEASE		GFR*	% OF KIDNEY FUNCTION
Stage 1	Kidney damage with normal kidney function	90 or higher	 90-100%
Stage 2	Kidney damage with mild loss of kidney function	89 to 60	 89-60%
Stage 3a	Mild to moderate loss of kidney function	59 to 45	 59-45%
Stage 3b	Moderate to severe loss of kidney function	44 to 30	 44-30%
Stage 4	Severe loss of kidney function	29 to 15	 29-15%
Stage 5	Kidney failure	Less than 15	 Less than 15%

Figure 1.2: Kidney function progression

possible surrogate markers have been proposed. For example, the papers by Marcén⁷⁷ and Moranne⁷⁸ proposed to use the slope of GFR to predict the graft failure by a Cox model.

In fact, it is a clinical problem in the longitudinal continuous outcomes and multiple time-to event outcomes data, where the trajectories of kidney function progression recorded as repeated GFR measurements and other multiple outcomes like the transplant failure, death after the transplant failure, and death without the kidney failure. Since the repeated measures on each individual may be correlated and there are large variations in GFR across individuals and within individuals, a longitudinal model may be appropriate to model the GFR trajectories. Marcén et al.⁷⁷ and Moranne et al.⁷⁸ used a mixed model for the GFR trajectories, and they proposed to use the slope of GFR trajectories to predict the graft failure in a Cox model. However, Cox model may cause some bias for multiple time-to event outcomes in the present of competing-risk events. Studies such as Prentice⁸³ and Putter⁸⁴ show that the Kaplan-Meier or Cox method for multiple outcomes may yield unreliable results in the presence of competing risks. The kidney transplant failure is a

competing risk for death because the kidney transplant failure increases the probability of death. Furthermore, if a large proportion of trajectories of GFR are nonlinear, this simplified linear assumption may cause the result to be biased. This longitudinal continuous repeated outcome and multiple time-to event outcomes motivate us to develop the new statistics model in this thesis. The proposed statistical models are created on the different scenarios in this clinical question. I hope that the results from the proposed new models can supply some references for the future kidney research, especially for how to predict the long-term transplant outcomes.

1.2 Methods for longitudinal data

The defining feature of the clinical longitudinal data is that the biomarker measurements of the same patients are taken repeatedly during the followed-up time period, thereby allowing the researcher to observe the interesting outcome over time.

The primary goal of a longitudinal study is to characterize the outcome change over time and to identify the factors that influence change. If the longitudinal data is clustered, then the observations within a cluster will typically exhibit the correlation, which have to be accounted for in the analysis. Alternatively, clustered data can arise from random sampling of naturally occurring groups in the population. Family, hospital medical practices, and schools are all instances of naturally occurring clusters in the population.

According to these features of longitudinal data, many statistical models have been developed. For example, Laird and Ware proposed the use of the EM algorithm to fit a class of linear mixed effects models in the early 1980s. Recently more methods in the analysis of longitudinal and multilevel data continue to develop. New and more flexible models such as the generalized estimating equations by Liang and Zeger. The new algorithm such as Markov Chain Monte Carlo (MCMC) have been developed. Also, the non-parameter method such as functional principal component analysis (FPCA) provides another way to look at the the variance-covariance correlation structure and dominant modes of the longitudinal trajectory. Therefore, FPCA has become a hot topic in statistical research such as climatology, medicine, and economics.

1.2.1 Functional principal component analysis

Functional principal component analysis (FPCA) is becoming a popular statistical method when we want to investigate the dominant modes of variation of functional data. In this method, a random function is represented in the eigenbasis, which is an orthonormal basis of the Hilbert space L^2 that consists of the eigenfunctions of the autocovariance operator. FPCA represents functional data in the most parsimonious way, in the sense that when using a fixed number of basis functions, the eigenfunction basis explains more variation than any other basis expansion. FPCA can be applied for representing random functions

or in functional regression and classification. Asymptotic convergence properties of these estimates have been investigated.

FPCA can be applied for displaying the modes of functional variation in scatterplots of FPCs against each other when modeling sparse longitudinal data or for functional regression and classification. Scree plots can be used to determine the number of included components. This methodology is being adapted from traditional multi-variate techniques to carry out analysis on financial data sets such as stock market indices, generation of implied volatility graphs and so on. Since being introduced by Rao¹⁰ for comparing growth curves, FPCA has attracted considerable attention. For instance, Castro et al.¹¹ related FPCA to the Karhunen-Loève theorem and the best m -dimensional functional linear model. Dauxois et al.¹² studied the asymptotic properties of empirical eigenfunctions and eigenvalues when sample curves are fully observable. A very nice example of the advantages of the functional approach is the Smoothed FPCA (SPCA), proposed by Silverman [1996] and studied by Pezzulli and Silverman [1993] that enables direct combination of the FPCA analysis together with a general smoothing approach that makes the use of the information stored in a linear differential operators possible.

An important application of the FPCA already known from multivariate PCA, is motivated by the Karhunen-Loève decomposition of a random function to the set of functional parameters factor functions and corresponding factor loadings (scalar random variables). This application is much more important than in the standard multivariate PCA since the distribution of the random function is in general too complex to be directly analyzed and the Karhunen-Loève decomposition reduces the analysis to the interpretation of the factor functions and the distribution of scalar random variables. Due to dimensionality reduction as well as its accuracy to represent data, there is a wide scope for further developments of functional principal component techniques in the financial and medical field . Zhang and Chen¹³ and Benko et al.¹⁴ extended this work to a more practical setting where sample curves are observed at finitely many design points. Hall and Hosseini-Nasab^{15,16} studied the estimation errors of empirical eigenfunctions and eigenvalues. To overcome excessive variation of empirical eigenfunctions, Rice and Silverman¹⁷ proposed smoothing estimators of eigenfunctions via a roughness penalty. Consistency of these estimators was established by Pezzulli and Silverman¹⁸. Subsequently, Silverman¹⁹ proposed an alternative way to obtain smoothing estimators of eigenfunctions through modifying the norm structure, and established the consistency of the estimators. A kernel-based method for smoothing eigenfunctions was proposed by Boente and Fraiman²⁰.

The extension of FPCA to sparse data such as longitudinal data was studied by James et al.⁸¹ and Yao et al.⁸⁵. James et al.²³, Tian and James²⁴, and Lin et al.²⁵ proposed to increase the interpretability of FPCA by adding some sparse constraints on functional principal components. FPCA has been used to explore variations of curves in a sundry groups of applications in subjects such as biology and medicine. For instance, Feng et al.²⁶

applied FPCA to explore spatial and temporal variations of cadmium concentrations in Pacific oysters from British Columbia. Luo et al.²⁷ used FPCA to detect the major modes of variations among ward admission intensity functions in hospital emergency departments. An excellent introduction on FPCA can be found in Chapters 8 and 9 of Ramsay and Silverman²⁸

1.2.2 Parameter models for longitudinal data

As mentioned in the background, there are two sources of variation in longitudinal data: one is the variation from within-individual, and the other between-individual variation. Modelling the variation within-individual allows one to see the change of the interest longitudinal outcome over time during the followed-up time period, while modelling between-individual variation allows one to understand the differences between individuals.

Regression models such as the Generalized Linear Effects Models and the Generalized Mixed Linear Effects Models are often used to approximate the relationship between the longitudinal data, using the responses and covariates terminology, for prediction or scientific exploration purpose. There are two types of covariates: time-invariant covariates such as a patients sex and race and time-varying covariates such as age or the status of the organ transplant. The book applied longitudinal analysis by Fitzmaurice et al. (2002) provided a comprehensive overview of various longitudinal models: a linear model or a generalized linear model. Both types of model have the mixed effects model (LME) as a special case when consider the random effect or missing value. The mixed effects model is a popular approach to model such type of data arising in clinical trials and epidemiological studies of cancer and other diseases. This thesis is focus on the application of our proposed joint models in the renal diseases.

The Generalized Linear Effects Models

Let y_1, y_2, \dots, y_n be a sample of i.i.d. observations from a distribution in the exponential family. The general probability density function of y_i can be written as if we choose the joint distribution $f(y_i)$ in the exponential family.

$$f(y_i|u_i, \boldsymbol{\alpha}\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta)}{a(\phi)} + c(y_i, \phi)\right\},$$

where θ is the natural or canonical parameter, and ϕ is the dispersion parameter, and a , b , and c are specific functions. It can be shown (Molenberghs and Verbeke, 2005) that Y has mean and variance

$$E(y_i) = \mu = \partial b(\theta)/\partial\theta,$$

$$Var(y_i) = a(\phi)\frac{\partial^2 b(\phi)}{\partial\theta^2}$$

The Generalized Mixed Linear Effects Models

The Generalized Linear Mixed Effects Models (GLMMs) are the most frequently used random effects models in the context of discrete repeated measurements. It is a useful tool to analyze longitudinal data with different individual variation.

GLMMs assume that the response is linked to a function of covariates with fixed regression coefficients and random coefficients. Let y_{ij} be the response variable of the i^{th} subject at time j , and \mathbf{x}_{ij} be the vector of covariates associated with the response y_{ij} , where $i = 1, \dots, N$, and $j = 1, \dots, n_i$.

$$Y_{ij}|u_{ij} \sim f(y_{ij}|\mu_{ij}, \phi), \text{ where } g(\mu_{ij}) = \boldsymbol{\alpha}^T \mathbf{X}_i + u_{ij}, i = 1, \dots, n,$$

They are an extension of the class of generalized linear models in which random effects are added to the linear predictor. This modification extends the broad class of generalized linear models to accommodate correlation via random effects, while retaining the ability to model non-normal distributions and allowing non-linear models of specific form. The class of GLMMs includes the special cases of linear mixed models, random coefficient models, random effects logistic regression, and random effects Poisson regression, and etc. The incorporation of random effects is a natural way to model or accommodate correlation in the context of a linear/nonlinear model for normal/nonnormal data. It generates a rich class of correlated data models that would be difficult to specify directly. We will review it in the Chapter 5. Readily available, flexible, multivariate distributions analogous to the multivariate normal distribution do not exist for most normal/nonnormally distributed data. For example, a longitudinal model for repeated measurement outcome $Y_i(t_{ij})$ as follows is used to fit the longitudinal outcomes GFR in our application example:

$$Y_i(t) = \boldsymbol{\alpha}^T \mathbf{Z}_i + \boldsymbol{\beta}_i^T \boldsymbol{\xi}(t) + \boldsymbol{\epsilon}_i(t), i = 1, \dots, n, \quad (1.1)$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_P)^T$ is a vector of coefficients for the fixed effects of $\mathbf{Z}_i = [Z_{i1}, \dots, Z_{iP}]^T$, and $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{iL})^T$ is a vector of coefficients for the random effects of $\boldsymbol{\xi}(t) = (\xi_1(t), \dots, \xi_L(t))^T$. Here, $\xi_\ell(t), \ell = 1, \dots, L$, is a parametric function of t . For example, $\xi_1(t) = 1$ and $\xi_2(t) = t$. We assume that $\boldsymbol{\beta}_i \sim \text{Normal}(\mathbf{b}, \mathbf{B})$. The vector of measurement errors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{im_i})$ are assumed to be multivariate normal distributed with the mean *zero* and the variance-covariance matrix $\sigma^2 \mathbf{I}_n$.

Inferences

Inferences for these models including a linear model or a generalized linear mixed effects model can be of the usual variety, that is, modeling the effect of predictors on the mean, in which case the random effects and correlation are nuisance features of the model. In other

situations, however, both estimation and testing of the variances of the random effects, as well as prediction of the realized values of the random effects, may be of interest.

Let Y_{ij} be the j^{th} outcome for the cluster i or repeated measurements of i patient, $i = 1, \dots, N$; $j = 1, \dots, n_i$ and Y_i is the vector of all measurements for the cluster i . We formulate the generalized linear mixed effects models using a two-step specification. First-step: Assume that the conditional distribution of each Y_{ij} , given the random effects b_i , belongs to the exponential family with conditional mean. The parameter approach for accounting for the within-subject association via the latent features. The likelihood-based inference is the standard approach for these models including a linear model or a generalized linear model. Suppressing the covariates, we could write the marginal distribution f for y_i in the following unified way

$$f(y_i|\boldsymbol{\theta}, D) = \int f(y_i|\mathbf{b}_i; \boldsymbol{\theta})f(\mathbf{b}_i|D)d\mathbf{b}_i.$$

So the likelihood is in the following

$$L(\boldsymbol{\theta}, D|y) = \prod_{i=1}^n f(y_i|\mathbf{b}_i; \boldsymbol{\theta})f(\mathbf{b}_i|D).$$

where $y = (y_1, \dots, y_n)^T$, and $\boldsymbol{\theta}$ is the collection of all parameters except D . But for a NLME model or a GLMM model, the likelihood involves an intractable multi-dimensional integral with respect to the random effects. We will discuss it in more detail in the chapters 3 and 5. The maximum likelihood method or the restricted maximum likelihood method can be used for LME. For GLMM, likelihood methods include the exact methods based on Gauss-Hermite quadratic integration techniques, EM algorithms, and approximate methods based on Taylor approximations or Laplace approximations. Dean and Nielsen (2007) provided a recent review of these methods for GLMM.

1.3 Survival analysis

1.3.1 Types of censoring

This section is to introduce and explain the concepts of survival analysis. The main outcome under assessment in the survival analysis is the time to an event of interest. If the event occurred in all individuals, the time survived from complete remission to relapse or progression as equally as to the time from diagnosis to the date of event. However, it is usual that at the end of follow-up some of the individuals have not had the event of interest, and thus their true time to event is unknown. The data on these individuals are said to be right censored. The right-censoring is said to be independent if the failure rates that apply to individuals on trial at each time $t > 0$ are the same as those without censoring. Individuals can also be subject to left censoring if the individual is observed to fail prior to some time

t , but the actual time of failure is unknown. For example, the observe $T \in (0, t)$ in the right censoring, while $T \in (t, \infty)$.

1.3.2 The Accelerated Failure Time model (AFT model)

In the statistical area of survival analysis, an accelerated failure time model (AFT model) is a parametric model that provides an alternative to the commonly used proportional hazards models. Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. This is especially appealing in a technical context where the disease is a result of some mechanical process with a known sequence of intermediary stages. The interpretation of means that everything in the relevant life history of an individual happens twice as fast.

Cox and Oakes⁶¹ extended a AFT model with time-dependent covariates.

$$\lambda(t|Z(t)) = \lambda_0 \left\{ \int_0^t \exp(\beta Z(s)) ds \right\} \exp(\beta Z(t)),$$

James⁶² provided a method to estimate the time-dependent AFT model in the presence of confounding factors. Reid⁶³ and Kay⁶⁴ mentioned that the AFT models were more appealing than the the proportional hazard models because they could give direction physical interpretations. For example, as mentioned in the paper by Kay⁶⁴, the AFT model can supply a more straightforward interpretation of the treatment effect on the time to event data because the coefficient of the treatment indicator can be estimated across various intervals defined by the cut time points from the date of treatment." We also explain how to interpret the model parameters in the second paragraph of Page 8. In the statistical area of survival analysis, an accelerated failure time model (AFT model) is a parametric model that provides an alternative to the commonly used proportional hazards models. Whereas a proportional hazards model assumes that the effect of a covariate is to multiply the hazard by some constant, an AFT model assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. This is especially appealing in a technical context where the disease is a result of some mechanical process with a known sequence of intermediary stages.

The hazard function of the accelerated failure time model can be specified as $\lambda(t|\phi) = \phi \lambda_0(t\phi)$, where ϕ denotes the joint effect of covariates. Then the survival function can be expressed as $S(t|\phi) = S_0(t\phi)$. From this, we can see that the moderated life time T is distributed such that $T\phi^{-1}$ and the unmoderated life time T_0 have the same distribution. So $\log(T) = \log(\phi) + \log(T\phi^{-1}) = \log(\phi) + \tau$, where the last term τ has the same distribution as $\log(T_0)$. The interested parameters need to be estimated, for example, γ in the $\phi = \exp\left(\sum_{p=1}^P \gamma_{1p} Z_{ip} + \gamma_{21} \beta_{i1} + \gamma_{22} \beta_{i2} + w(t|\gamma_3)\right)$ in the application example in the Chapter 3.

For the distributions of τ used in AFT models, the log-logistic distribution provides the most commonly used AFT model. Unlike the Weibull distribution, it can exhibit a non-monotonic hazard function which increases at early times and decreases at later times. It is somewhat similar in shape to the log-normal distribution but it has heavier tails. The log-logistic cumulative distribution function has a simple closed form, which becomes important computationally when fitting data with censoring. For the censored observations one needs the survival function, which is the complement of the cumulative distribution function, i.e. one needs to be able to evaluate . The Weibull distribution (including the exponential distribution as a special case) can be parameterized as either a proportional hazards model or an AFT model, and is the only family of distributions to have this property. The results of fitting a Weibull model can therefore be interpreted in either framework. However, the biological applicability of this model may be limited by the fact that the hazard function is monotonic, i.e. either decreasing or increasing. Other distributions suitable for AFT models include the log-normal, gamma and inverse Gaussian distributions, although they are less popular than the log-logistic, partly as their cumulative distribution functions do not have a closed form. Finally, the generalized gamma distribution is a three-parameter distribution that includes the Weibull, log-normal and gamma distributions as special cases.

1.3.3 Multi-state survival models

Several multi-state survival models have been developed. We are focus on the competing risks models and the progressive illness-death models in this paper. In the competing risks framework, two popular competing risks models are used. One is the cause-specific hazard model proposed by Prentice⁸³ and Putter⁸⁴, and the other is the sub-distribution hazards regression introduced by Fine and Gray⁸⁸. The cause-specific hazard model calculates the occurrence rate of specific event types in subjects who are currently event free. For example, there are 2 types of events in this application example: death with the kidney function from other reasons and death from the kidney transplant failure. The cause-specific hazard of the kidney failure death denotes the instantaneous rate of the kidney failure death in alive subjects who have not yet experienced either event. The sub-distribution hazard model calculates the instantaneous risks of the specific event type in subjects who have not yet experienced this event type. Each method has its own specific purpose. For example, If the progressive illness-death model is used as shown in Figure 1.3 for three state survival model,

the progressive illness-death model can determine the incidence of kidney transplant failure, the mortality rate for alive patients after kidney transplant, and mortality rate for patients with kidney transplant failure. If the transition intensities of multi-state models can be specified as

$$\lambda_{jm}(t) = \lim_{\Delta t \rightarrow 0} \frac{P_{jm}(t, t + \Delta t)}{\Delta t}$$

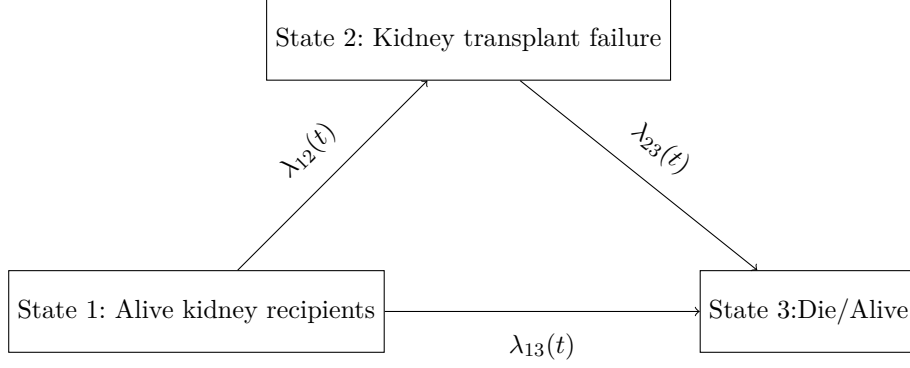


Figure 1.3: The three states of kidney transplant recipients. All patients start from the date of the kidney transplant (state 1), then they may move to state 2 (kidney failure). If not, they directly move to state 3 when die

, $j \neq m$, and

$$\lambda_{mm} = - \sum_{j \neq m} \lambda_{jm}(t),$$

then the transition intensities can be specified in a matrix. For the convenient notation by setting $M = 3$, the matrix of transition intensities area can be specified as follows:

$$Q(t) = \begin{bmatrix} -(\lambda_{12}(t) + \lambda_{13}(t)) & \lambda_{12}(t) & \lambda_{13}(t) \\ 0 & -\lambda_{23}(t) & \lambda_{23}(t) \\ 0 & 0 & 0 \end{bmatrix}$$

where

$$\lambda_{12} = \lim_{\Delta t \rightarrow 0} \frac{P_{12}(t, t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(\text{state 2 at time } t + \Delta t | \text{state 1 at time } t)}{\Delta t},$$

$$\lambda_{13} = \lim_{\Delta t \rightarrow 0} \frac{P_{13}(t, t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(\text{state 3 at time } t + \Delta t | \text{state 1 at time } t)}{\Delta t},$$

$$\lambda_{23} = \lim_{\Delta t \rightarrow 0} \frac{P_{23}(t, t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(\text{state 3 at time } t + \Delta t | \text{state 2 at time } t)}{\Delta t}.$$

If the probability distribution on the state space of a Markov chain is discrete and the Markov chain is homogeneous, then the Chapman-Kolmogorov equations can be expressed in terms of matrix multiplication:

$$P(s, t) = P(s, u)P(u, t), s < u < t$$

The transition probability $P(s, t)$ is the unique solution of the Kolmogorov forward differential equation:

$$\frac{\partial}{\partial t} P(s, t) = P(s, t)Q(t); P(s, s) = I$$

$P(s, t)$ can be recovered from the transition intensities through product integration

$$P_{11}(s; t) = \exp\left(-\int_s^t (\lambda_{12}(u) + \lambda_{13}(u))du\right)$$

$$P_{22}(s; t) = \exp\left(-\int_s^t \lambda_{23}(u)du\right)$$

$$P_{23}(s, t) = 1 - P_{22}(s, t)$$

$$P_{12}(s, t) = \int_s^t P_{11}(s, u)\lambda_{12}(u)P_{22}(u, t)du$$

$$P_{13}(s, t) = 1 - P_{11}(s, t) + P_{12}(s, t)$$

$$P_{jm}(s, t) = 0, \quad \text{when } j > m$$

1.4 Methods for joint modelling the longitudinal and multiple time-to event outcomes

As mentioned before, no studies have used a joint model to predict the long-term kidney function, which includes the longitudinal continuous outcome of glomerular filtration rate (GFR) and the time-to-event outcome of all-cause graft loss (ACGL). From our preliminary result as in Figure 3.1, the levels of the observed GFR trajectories for patients with ACGL are higher than those of patients without ACGL, and we find that the slopes of GFR trajectories for patients with ACGL are steeper in comparison with patients without ACGL. In addition, all patients may have a pancreas transplant at any time post kidney transplant to treat the diabetic disease, and they are in different statuses at different time points during the followed-up period. For example, they are in status 1 (alive without pancreas transplant) at the time of the admission to the waiting list for a pancreas. They move to status 2 (alive with pancreas transplant) if a matched pancreas organ becomes available for them before ACGL, or they directly move to status 3 with ACGL. In fact, the failure rates of the time-to-event outcome are different as patients change their status. As shown in Figure 3.1, it seems that patients who have a pancreas transplant are less likely to have ACGL. In short, the above scenarios motivate us to develop a new dynamical joint model to predict the long-term outcome, since there are at least two advantages in using a joint model. Firstly, joint modelling multiple outcomes together can increase the power and decrease the Type I error^{46,47}. Secondly, this joint model can estimate the parameters in the longitudinal component by incorporating the time-to-event information through censoring, and similarly, for the converse situation, the estimation of the time-to-event is incorporated by the longitudinal data information.

Several joint models for multiple outcomes have been developed such as the papers^{48,49,52,54,55,59,89,92}. One major challenge in jointly modelling multiple outcomes is the lack of a suitable multivariate joint distribution. Two approaches are proposed for jointly modelling

multiple outcomes. The first approach directly specifies the joint distribution by factorizing it into the conditional distribution of one outcome and a marginal distribution of the other outcome. For instance,^{54,89} parameterized the model such that the joint distribution is factorized as the product of the marginal distribution of a continuous response and the conditional distribution of a discrete response given the continuous response or latent variables. Another case is that the binary response is related to an unobserved continuous latent variable, and the latent variable and the continuous response have a joint Gaussian distribution. The second approach directly formulates a joint model for both types of outcomes. For instance,⁵⁵ used a copula to construct the joint distribution. Another challenge in joint models is the intensive computation because of the complex correlation structure of latent variables or measurement errors in covariates such as in the papers^{49,54,89}. Most of above studies apply the EM or the Monte Carlo EM algorithm to estimate parameters, and some use Bayesian method such as in the paper⁵⁹.

1.5 Algorithm Material on Monte Carlo

In this section, we review the basics of Monte Carlo methods for the remainder of the chapter. In statistics, we are often tasked with computing the expected value of a function $f(x)$ with respect to a probability density function $p(x)$, where $x \in R_n$, especially when n is not small. If a cumulative distribution is non-decreasing and easily invertible then we can draw samples from its distribution by using inverse sampling. However, many distributions are difficult or impossible to invert, and in some cases a closed-form representation might not exist or be computationally intractable to obtain. This is a problem since finding expected values of functions is often a step in a lot of statistical problems. We outline several methods, Gibbs Sampling, importance sampling, and rejection sampling, that are useful when direct simulation from p is difficult or impossible but direct simulation from another distribution $q(x)$ is possible. We refer to the distribution similar to $p(x)$ as the instrumental distribution, and label it $q(x)$.

1.5.1 Gibbs Sampler

The Gibbs sampler (Gelfand and Smith, 1990) is a popular Markov Chain Monte Carlo algorithm to generate samples from a complicated multi-dimensional distribution by sampling from lower dimension full conditionals, in turn, until convergence. Here the Gibbs sampler can be used to simulate the missing random effects such as β_{1i} and β_{2i} . Firstly we set the initial values $(\beta_{1i}^0, \beta_{2i}^0)$. If the current generated values are $(\beta_{1i}^k, \beta_{2i}^k)$, $k = 0, 1, 2, \dots$, we can obtain $(\beta_{1i}^{k+1}, \beta_{2i}^{k+1})$ as follows:

1. Draw a sample for the missing or random effect β_{1i}^{k+1} from the full conditional distribution

$$f(\beta_{1i}^k | x_i, y_i, \beta_{2i}^k; \theta^{(t)}),$$

2. Draw a sample for the missing or random effect β_{2i}^{k+1} from the full conditional distribution

$$f(\beta_{2i}|x_i, y_i, \beta_{1i}^{k+1}; \theta^{(t)}).$$

3. Repeat the above steps k times.

We assess the convergence of the Gibbs sampler by examining sample autocorrelation function plots and time series plots. After a sufficiently large burn-in of r iterations, the sample values will achieve a steady state as reflected by the time series plots. Then, $(\beta_{1i}(r), \beta_{2i}(r))$ can be treated as a sample from the multidimensional density function $f(\beta_{1i}, \beta_{2i}, x_i, y_i; \theta^{(t)})$.

1.5.2 Importance Sampling

Importance sampling can be used when the density say $f(x)$ is difficult to sample. Basically it draws from a similar distribution other than $p(x)$, say $q(x)$, and then the bias is corrected if sample from the wrong distribution.

We estimate the expectation of $f(x)$ with respect to $p(x)$ by

$$I = \int f(x)p(x) = \int f(x) \frac{p(x)}{q(x)} q(x) = \int \frac{f(x)p(x)}{q(x)} q(x).$$

We can see the bias correction, or the importance weight $p(x)/q(x)$ can be determined exactly for a given sampling point x . In practice, the actual $p(x)$ or $q(x)$ will often be unnormalized.

Hence, given an iid sample x^1, \dots, x^N from $q(x)$, our estimator of \hat{I} becomes

$$\hat{I} = N^{-1} \sum_{i=1}^N \frac{p(x^i)f(x^i)}{q(x^i)}.$$

As the number of samples is increased, the variance of the estimate I will decrease. The selection of $q(x)$ will have a huge impact on the accuracy of our estimation. For example we can select $q(x)$ that has a similar shape to $f(x)$, but with thicker tails. In fact, one of the biggest problems with using the importance sampling method is that a poor selection of the sampling distribution will lead to a high-variance estimate I , that yields the wrong answer without any indication.

1.5.3 Laplace Approximation

The Laplace approximation is very useful for Monte Carlo as it may be used to construct accurate instrumental density, $q(x)$. The Laplace approximation is an analytic approximation to the expectation with respect to a distribution $p(x)$. We assume $l(x) = \log p(x)$ admits a second-order Taylor expansion about the mode of $p(x)$. Let x_0 denote the maximize of $l(x)$ satisfying the equation $l''(x_0) < 0$. We can expand $l(x)$ around x_0 by Taylor's theorem,

$$I(x) = I(x_0) + I'(x_0)(x - x_0) + \frac{1}{2}I''(x_0)(x - x_0)^2 + R$$

where $R = O((x - x_0)^3)$

The Laplace method can be applied to approximate integrals of the form.

1.6 Outline of this dissertation

This dissertation is motivated by some end-stage renal disease and kidney transplant data, and we illustrate the proposed models and the associated inference procedures using some clinical dataset. The proposed statistical methodologies are not only limited to this specific program and can be applied to other clinical or medical studies. The rest of this thesis is organized as follows:

1. Chapter 2 introduces Functional Principal Component Analysis through the conditional expectation for the longitudinal Curves and its application GFR curves after Kidney Transplant, which have been published in Statistical Methods in Medical Research (2017).
2. Chapter 3 develops a Joint model of a longitudinal and Accelerated Failure Time data and its application to transplant patients with an ESRD and a diabetes, which have been published in Statistical Methods in Medical Research (2018)
3. Chapter 4 is about a Joint model for Multiple Outcomes by Functional Principal Component Analysis via a Multistate Model, which uses functional principal component analysis (FPCA) to fit the longitudinal outcome and proposes the multi-state model to describe multiple time-to event outcomes together. The FPCA method is efficient in reducing the dimension of the longitudinal trajectories. Multistate submodel can be used to describe the dynamic process of multiple time-to-event outcomes. The longitudinal trajectories and the multiple time-to-event outcomes is linked with the shared latent features.
4. Chapter 5 jointly modelling multiple mixed continuous and discrete outcomes through a flexible class of generalized linear latent variables.
5. Chapter 6 introduces statistical inference in a predict model with a polynomial effect covariate in presence of measurement errors. We apply this method to predict the kidney donor incidence rate.
6. Chapter 7 is about future work.

Chapter 2

Functional Principal Component Analysis of GFR Curves after Kidney Transplant

2.1 Introduction¹

Kidney transplantation supplies a preferred therapy to extend survival time for patients with end-stage renal disease. Rates of acute rejection and graft failure that occur in the first 3-6 months after kidney transplant have been improved over the past couples of decades, but kidney transplant recipients still confront the high probability of the loss of graft function after kidney transplant. How to assess and extend the long-term kidney function remains a crucial research goal. GFR can provide a more precise measure of kidney function than serum creatinine alone⁵. Retrospective studies^{6,7} have shown that one-year GFR is a good predictor for the long-term graft function after renal transplant. Klahr et al.⁸ and Marcén et al.⁹ recommended that the change in GFR should be used to assess the progression of the kidney function and to identify the risk for kidney graft failure. Consequently, a natural candidate for a surrogate marker for the progression of the kidney function is the GFR progression curve as shown in Figure 3.1.

In the studies above, linear least squares regression or linear mixed effect models were used to calculate the change of GFR. However, these statistical models can not completely characterize complexity of GFR trajectories especially when the non-linear trends exist. Furthermore, it is difficult to estimate the change of GFR when the GFR trajectory is sparsely and irregularly observed. For instance, some patients have missing data records as shown in the upper left panel (a) of Figure 3.1. In addition, all the curves in the upper right panel (b) of Figure 3.1 are very flat. Conversely, the lower left panel (c) of Figure 3.1 shows that the curves of these patients have strong fluctuating curves. In other words, these

¹This chapter has been published in *Statistical Methods in Medical Research* (2017). Dong J, and Wang S, and Wang L, and Gill J, and Cao J

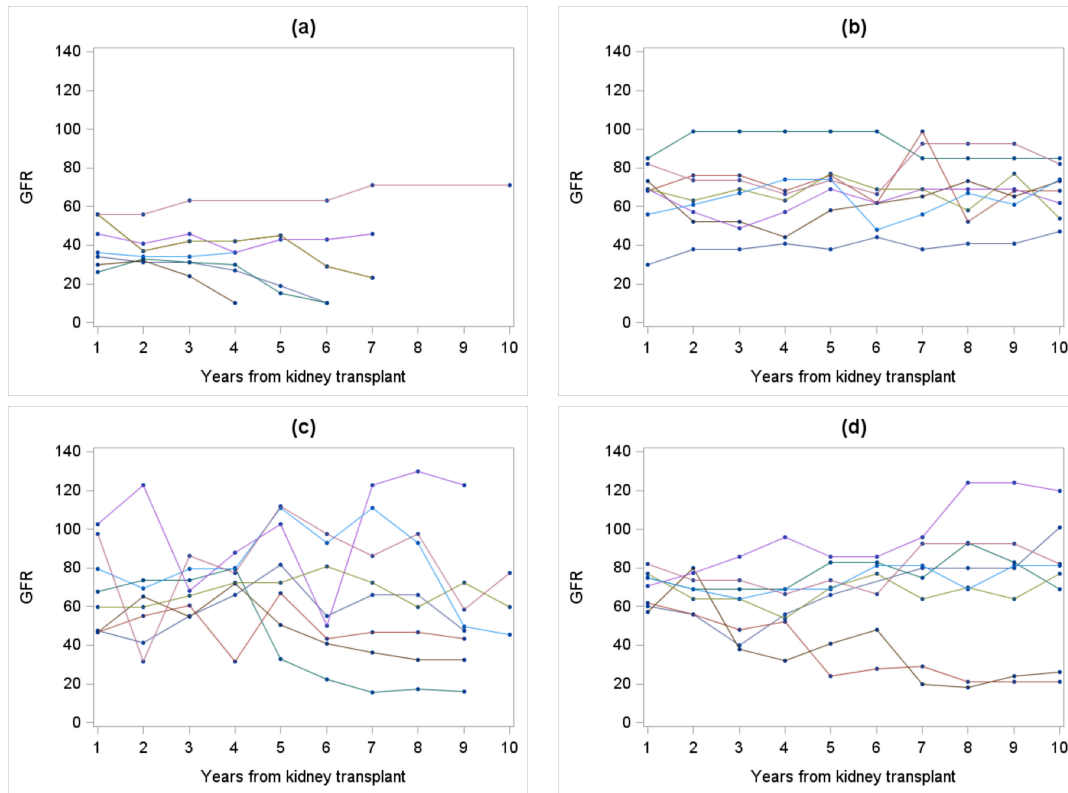


Figure 2.1: Observed GFR trajectory curves with various circumstances and trends. Patients in the upper left panel (a) have missing data records. Patients in the upper right panel (b) have flat GFR trends. Patients in the lower left panel (c) have strong fluctuating trends. Patients in the lower right panel (d) have increasing or decreasing trends. Each color represents one individual patient in each panel.

patients are in an unstable stage with a large amount of variations. Patients in the lower right panel (d) of Figure 3.1 have a negative or positive slope.

Our goal in this article is exploring the major source of variations of GFR curves, clustering GFR curves, detecting outlying GFR curves, estimating missing GFR values, and predicting future GFR values. To explore these variations of GFR trajectories, we propose to use functional principal component analysis (FPCA) to fit and predict GFR trajectories in this paper.

FPCA is a cutting-edge method for detecting the major source of variations in curves and projecting infinite-dimensional curves into low-dimensional vectors. Since being introduced by Rao¹⁰ for comparing growth curves, FPCA has attracted considerable attention. For instance, Castro et al.¹¹ related FPCA to the Karhunen-Loève theorem and the best m -dimensional functional linear model. Dauxois et al.¹² studied the asymptotic properties of empirical eigenfunctions and eigenvalues when sample curves are fully observable. Zhang and Chen¹³ and Benko et al.¹⁴ extended this work to a more practical setting where sample curves are observed at finitely many design points. Hall and Hosseini-Nasab^{15,16} studied the estimation errors of empirical eigenfunctions and eigenvalues. To overcome excessive variation of empirical eigenfunctions, Rice and Silverman¹⁷ proposed smoothing estimators of eigenfunctions via a roughness penalty. Consistency of these estimators was established by Pezzulli and Silverman¹⁸. Subsequently, Silverman¹⁹ proposed an alternative way to obtain smoothing estimators of eigenfunctions through modifying the norm structure, and established the consistency of the estimators. A kernel-based method for smoothing eigenfunctions was proposed by Boente and Fraiman²⁰. The extension of FPCA to sparse data such as longitudinal data was studied by James et al.⁸¹ and Yao et al.⁸⁵. James et al.²³, Tian and James²⁴, and Lin et al.²⁵ proposed to increase the interpretability of FPCA by adding some sparse constraints on functional principal components. FPCA has been used to explore variations of curves in a sundry groups of applications in subjects such as biology and medicine. For instance, Feng et al.²⁶ applied FPCA to explore spatial and temporal variations of cadmium concentrations in Pacific oysters from British Columbia. Luo et al.²⁷ used FPCA to detect the major modes of variations among ward admission intensity functions in hospital emergency departments. An excellent introduction on FPCA can be found in Chapters 8 and 9 of Ramsay and Silverman²⁸.

In this paper, the FPCA method is applied to explore the major variations of GFR trajectories. To the best of our knowledge, it is the first time that FPCA is applied to kidney transplant research. We find that FPCA can project the complex GFR trajectories into simple functional principal component (FPC) scores. These FPC scores enable us to cluster the GFR trajectories in such a way that each cluster contains homogeneous GFR trajectories, and allows us to detect GFR trajectory outliers. At the same time, FPCA method can effectively impute missing GFR information and predict future GFR based on all available data information from all kidney recipients.

The rest of this paper is organized as follows: Section 2 briefly introduces methods to analyze our kidney transplant recipient data. Section 3 provides the data analysis results from our data. Some concluding remarks are presented in Section 4.

2.2 Methods

2.2.1 Functional Principal Component Analysis

Functional principal component analysis is used to investigate the dominant modes of GFR curve variations during the followed-up time frame. Let $X_i(t)$ be the GFR trajectory of the i -th patient, where $i = 1, 2, \dots, n$, $t \in \mathcal{T}$, and \mathcal{T} is the bounded time-frame range. From the Karhunen-Lovève theorem, $X_i(t)$ can be expressed as

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t), \quad (2.1)$$

where $\mu(t) = E(X_i(t))$ is the mean trajectory, $\phi_k(t)$ is the k -th functional principal component (FPC), and $\xi_{ik} = \int_{\mathcal{T}} (X_i(t) - \mu(t)) \phi_k(t) dt$ is the associated functional principal component score. Then the variance-covariance function $G(s, t)$ can be expressed as:

$$G(s, t) = \text{Cov}(X_i(s) - \mu(s), X_i(t) - \mu(t)) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t), \quad (2.2)$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$.

In practice, $X_i(t)$ is usually well approximated by the first few leading FPCs and FPC scores:

$$X_i(t) = \mu(t) + \sum_{k=1}^K \xi_{ik} \phi_k(t). \quad (2.3)$$

The first FPC $\phi_1(t)$ displays the dominant mode of variations of $X_i(t)$. In other words,

$$\phi_1(t) = \arg \max_{\|\phi\|=1} \left\{ \text{Var} \left(\int_{\mathcal{T}} (X_i(t) - \mu(t)) \phi(t) dt \right) \right\}, \quad (2.4)$$

where $\|\phi\| = \left(\int_{\mathcal{T}} \phi(t)^2 dt \right)^{\frac{1}{2}}$. The k -th FPC $\phi_k(t)$ ($k = 2, \dots, K$) is the dominant mode of curve variation orthogonal to $\phi_1(t), \dots, \phi_{k-1}(t)$. It can be expressed as

$$\phi_k(t) = \arg \max_{\|\phi\|=1, \langle \phi, \phi_j \rangle = 0 \text{ for } j=1, \dots, k-1} \left\{ \text{Var} \left(\int_{\mathcal{T}} (X_i(t) - \mu(t)) \phi(t) dt \right) \right\}, \quad (2.5)$$

where $\langle \phi, \phi_j \rangle = \int_{\mathcal{T}} \phi(t) \phi_j(t) dt$, for $j = 1, \dots, k-1$. After obtaining the k -th FPC $\phi_k(t)$, the corresponding FPC score of the i -th curve $X_i(t)$ is calculated as

$$\xi_{ik} = \int_{\mathcal{T}} \phi_k(t) (X_i(t) - \mu(t)) dt.$$

However, the above method can not be applied to our kidney transplant recipient data directly, since there exists a measurement error or missing GFR at some time points for some recipients. In this case, the principal components analysis through the conditional expectation (PACE) method⁸⁵ can be used.

The PACE method estimates the top FPCs and FPC scores as follows. Let $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$ be the measured GFR at time t_{ij} , where ϵ_{ij} are identically and independently distributed normal random variables with mean 0 and variance σ^2 , and $j = 1, 2, \dots, m_i$. To estimate the mean function $\mu(t)$, we can consider observations in two different cases. If GFR observations are available on a regular grid, we take the average at each location t_{ij} : $\hat{\mu}(t_{ij}) = \frac{1}{n} \sum_{i=1}^n Y_{ij}$. If GFR observations are sparse, the mean function is obtained by smoothing the data from all observations based on the local linear smoother method⁸². Let $G_i(t_{ij}, t_{il}) = (Y_{ij} - \hat{\mu}(t_{ij}))(Y_{il} - \hat{\mu}(t_{il}))$, $j \neq l, i = 1, \dots, n$, be the sample covariance. It can be shown that $E(G_i(t_{ij}, t_{il})) = \text{Cov}(X(t_{ij}), X(t_{il})) + \sigma^2 \delta_{jl}$. Therefore, we only use the off-diagonal sample covariances $G_i(t_{ij}, t_{il})$ as input data to obtain the smooth covariance surface estimate $\hat{G}(s, t)$. Let $\hat{V}(t)$ be a smoothed version of the diagonal elements $G_i(t_{ij}, t_{ij})$ of the sample covariances. Then $\hat{V}(t)$ is an estimate of $G(t, t) + \sigma^2$. Therefore, an estimate of σ^2 is obtained by

$$\hat{\sigma}^2 = \frac{2}{|\mathcal{T}|} \int_{\mathcal{T}_1} \{\hat{V}(t) - \hat{G}(t, t)\} dt, \quad (2.6)$$

where $|\mathcal{T}|$ is the length of \mathcal{T} , and $\mathcal{T}_1 = [\inf\{x : x \in \mathcal{T}\} + |\mathcal{T}|/4, \sup\{x : x \in \mathcal{T}\} - |\mathcal{T}|/4]$. To ensure that the variance is nonnegative, $\hat{\sigma}^2$ is set to 0 if $\hat{\sigma}^2 < 0$.

The FPCs, $\phi_k(t)$, $k = 1, \dots, K$, are eigenfunctions of the eigenequation

$$\int_{\mathcal{T}} \hat{G}(s, t) \phi_k(s) ds = \lambda_k \phi_k(t), \quad (2.7)$$

with the constraints $\int_{\mathcal{T}} \phi_k^2(t) dt = 1$ and $\int_{\mathcal{T}} \phi_k(t) \phi_m(t) dt = 0$ for $m < k$. The eigenfunctions are estimated by discretizing the smoothed covariance $\hat{G}(s, t)$. We denote $\hat{\phi}_k(t)$ ($k = 1, \dots, K$) as the estimated FPCs.

The FPC score of the i -th curve $X_i(t)$ on the k -th FPC can be obtained by the conditional expectation

$$\hat{\xi}_{ik} = \mathbb{E}(\xi_{ik} | \mathbf{Y}_i) = \lambda_k \hat{\boldsymbol{\phi}}_k^T \hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_i}^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}),$$

where $\hat{\boldsymbol{\phi}}_k$ and $\hat{\boldsymbol{\mu}}$ are vectors by evaluating $\hat{\phi}_k(t)$ and $\hat{\mu}(t)$ at the grid points t_{ij} , $j = 1, 2, \dots, m_i$, $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im_i})^T$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}_i} = \tilde{\mathbf{G}} + \hat{\sigma}^2 \mathbf{I}_{m_i}$, and the matrix $\tilde{\mathbf{G}}$ is obtained by evaluating $\hat{G}(s, t)$ at the grid points t_{ij} , $j = 1, 2, \dots, m_i$.

The estimated trajectory of the i -th patient GFR is

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t). \quad (2.8)$$

The optimal number of FPCs, K , can be determined by various statistics methods. For instance, Shibata³¹ determined K by AIC based on a pseudo-Gaussian log-likelihood, and Rice and Silverman¹⁷ proposed the cross-validation (CV) score based on the leave-one-curve-out prediction error:

$$CV(K) = \sum_{i=1}^n \sum_{j=1}^{m_i} \{Y_{ij} - \hat{X}^{(-i)}(t_{ij})\}^2, \quad (2.9)$$

where $\hat{X}^{(-i)}(t)$ is the predicted GFR curve for the i -th subject after removing the i -th subject from the data. In this paper, we use the method proposed by Rice and Silverman¹⁷ as shown in the above formula, and then use the scree test plot³² to display the relationship between $CV(K)$ and K . We choose the optimal number of FPCs, K , according to the $CV(K)$ curve at the point where the curve starts levelling off.

2.2.2 Clustering

In clinical practice, it is more efficient to manage the clinical patients when they can be clustered into a small number of groups with homogeneous GFR curves. For example, a group of patients with flat GFR trends indicate that their current clinical treatments are effective and the kidney transplant is successful. On the other hand, a group of patients with decreasing GFR trends may require to be diagnosed, and consequently require an alternative clinical treatment. In this section, we illustrate how to cluster all kidney recipients into groups with similar GFR curve patterns.

As introduced in the section above, we can obtain a set of FPC scores $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_n)$, where $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iK})^T, i = 1, \dots, n$, is a K -dimensional vector of FPC scores for the i -th subject. Based on the distance between these FPC Scores, we use the k -means method^{33,34} to partition the individual GFR curves into Q sets $\mathbf{G} = (G_1, G_2, \dots, G_Q)$. We minimize the within-cluster sum of the distance of each FPC score vector in its cluster to its cluster center, which is defined by

$$\sum_{q=1}^Q \sum_{\boldsymbol{\xi}_i \in G_q} \|\boldsymbol{\xi}_i - \boldsymbol{\mu}_q\|^2, \quad (2.10)$$

where $\boldsymbol{\mu}_q$ is the mean vector in the set G_q . The optimal number of clusters is determined by the Silhouette method³⁵.

The algorithm of the k -means method is implemented in the following steps:

1. Split all kidney transplant recipients into Q initial clusters.
2. Assign each kidney recipient into the cluster whose centroid is the closest, and recalculate the centroid for the cluster once it receives ones or loses.
3. Repeat step 2 until no more reassignments take place in any clusters.

2.2.3 Detection of GFR trajectory outliers

In this section, we illustrate how to use an ordering method based on FPC scores to detect abnormal GFR curves. The rationality is that we can detect abnormal GFR curves by checking their FPC scores in the FPC space, because the abnormal FPC scores in the one-dimensional FPC space are easier to detect than the abnormal GFR curves in the infinite-dimensional functional space.

As mentioned in the subsection 2.1, FPC scores are defined as $\xi_{ik} = \int_{\mathcal{T}} (X_i(t) - \mu(t)) \phi_k(t) dt$. From this definition, FPC scores are the projection of GFR curves onto the space expanded by the FPCs. Therefore, it is reasonable to detect abnormal GFR curves by ordering $\sum_{k=1}^K (\xi_{ik}^2 / \lambda_k)$, where λ_k is the eigenvalue defined in the eigenequation (2.7). In fact, λ_k is also the variance of the FPC score ξ_{ik} . As shown in a later section, the result from our clinical data shows that abnormal curves are more visible in the FPC space than in the original functional space, because the FPC space has one dimension while the functional space has infinite dimensions. This is consistent with the result in Filzmoser et al.³⁶.

2.2.4 Prediction for Future GFR

After estimating the mean GFR curve $\hat{\mu}$, the FPC $\hat{\phi}_k$, and the FPC score $\hat{\xi}_{ik}$ from all available GFR data, we can estimate any missing GFR or predict future GFR using the following formula

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t), \quad (2.11)$$

where t can be any past or future time points.

2.3 Results on Kidney Transplant Data

The data resource for this paper is kidney transplant recipient data from the Organ Procurement Transplant Network/United Network for Organ Sharing (Optn/UNOS), which collect the kidney transplant recipient register form including patient description at the time of transplant, and other follow-up forms including patient description and GFR during the followed-up period.

2.3.1 Functional Principal Component Analysis

In this section, functional principal component analysis (FPCA) is applied to analyze the kidney transplant recipient data. Figure 2.2 displays the estimated mean and the correlation function of the GFR curves. The mean GFR curve reveals that the overall trend of kidney function is flat. The correlation function shows that GFR is temporally correlated with its adjoining GFR observation, and the correlation decreases to 0.6 as the time gap between two observations increases to 10 years.

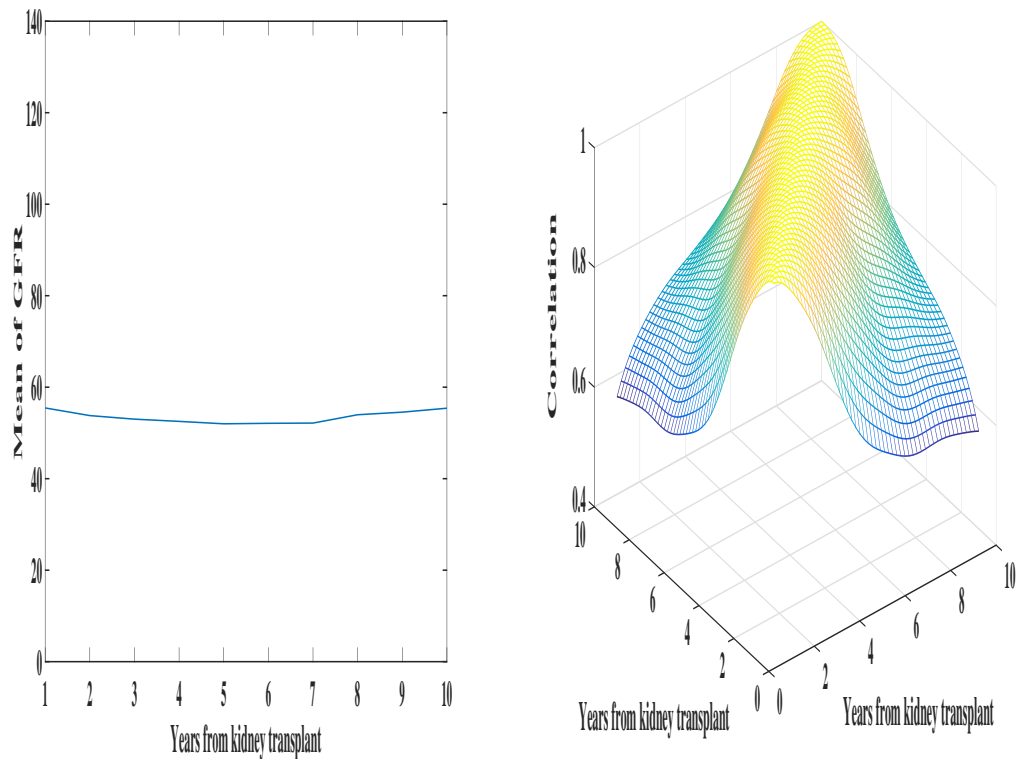


Figure 2.2: The mean curve of GFR in the left panel and the correlation function of GFR in the right panel. They are estimated from the total patients.

Figure 4.4 shows the four leading functional principal components (FPCs) estimated from the GFR curves. These four FPCs account for 99.8% of the total variation of GFR curves. Specifically, the first FPC explains around 84.6% of the total variation. The first FPC is positive throughout 10 years, with a slight increase at the beginning and then become flat. In other words, it represents that 84.6% of the total variation of GFR curves comes from the weighted average of GFR curves with more weight on GFR after 4.4 years from the date of kidney transplant, where the weighted average uses the first FPC as the weight function. The second FPC accounts for around 10.7% of the total variation. The second FPC is positive during the period from the beginning to 5.5 years, and then it becomes negative after 5.5 years. It can be interpreted as 10.7% of total variation comes from the change of GFR after 5.5 years in comparison with the early stage. The third FPC accounts for around 3.4% of the total variation. The third FPC is positive in $[3.2, 7.8)$ and negative for all other times. It represents that 3.4% of the total variation is from the difference of GFR in the middle stage $[3.2, 7.8)$ in contrast with the early and later stage. The fourth FPC is an S shape curve, which is positive in $[2.5, 5.6) \cup [8.5, 10.0]$ and negative in $[1.0, 2.5) \cup [5.6, 8.5)$. It stands for 1.1% of the total variation, and it comes from the change of GFR in $[2.5, 5.6) \cup [8.5, 10.0]$ in contrast with $[1.0, 2.5) \cup [5.6, 8.5)$. It is worth mentioning that many FPCA papers only consider the first two FPCs. In our case, the first two FPCs only account for 95.3% of the total variation; as a result it may neglect some important patient information. As shown in this Figure 4.1, a small number of patients have strong fluctuating curves, which can only be represented by the third and fourth FPCs. In clinical practice, these patients with abnormal trajectories should be monitored more closely, and they need to be diagnosed to find out the underlying reason.

Figure 4.1 displays the GFR curves for patients with extreme FPC scores. For instance, the top left panel in Figure 4.1 shows the GFR curves whose first FPC scores are smaller than the 5% quantile. All these GFR curves have low GFR during the 10 year period after kidney transplant. By contrast, the GFR curves in the top right panel in Figure 4.1 have their first FPC scores larger than the 95% quantile, and all of them have high GFR during the 10-year period since the kidney transplant. The GFR curves with their second FPC scores smaller than the 5% quantile have GFR increasing trends over time, while the GFR curves with their second FPC scores larger than the 95% quantile have decreasing trends over time. The GFR curves with their third FPC scores smaller than the 5% quantile start with high GFR values, decrease for the first 5 years, and rebound afterwards. By comparison, the GFR curves with their third FPC scores larger than the 95% quantile start with low GFR values, increase for the first 5 years, but decrease afterwards. The GFR curves with their fourth FPC scores smaller than the 5% quantile also have the opposite fluctuation to the GFR curves when their fourth FPC scores larger than the 95% quantile.

2.3.2 Patient clustering

In this section, we use the k -means method to cluster 5654 GFR curves. The Silhouette analysis method in Rousseeum and Silhouettes³⁵ is used to determine the optimal number of clusters to be 40. The k -means method clusters all curves into 40 groups, with the group size varying from 21 to 264. We also compare the k -means method with the model-based cluster method⁹¹. We use the adjusted-rand index³⁸ to evaluate the similarity of the clustering results from the two methods. The adjusted-rand index is 0.792, which indicates that the cluster results from these two methods are similar.

Generally, the renal function of patients is normal when GFR are over $90.0 \text{ mL}/\text{min}/1.73\text{m}^2$. Patients have the mildly decreased kidney function when GFR are in the range of $60.0\text{-}89.9 \text{ mL}/\text{min}/1.73\text{m}^2$. Patients have the moderately decreased kidney function when GFR are in the range of $30.0\text{-}59.9 \text{ mL}/\text{min}/1.73\text{m}^2$, and patients have the severely decreased kidney function when GFR are in the range of $15.0\text{-}29.9 \text{ mL}/\text{min}/1.73\text{m}^2$. Figure 2.5 displays part of GFR curves for six of these groups. The four groups in the top four panels show that all kidney recipients have the stable kidney function with flat GFR trends, but their GFR levels are very different. They are in various chronic kidney disease (CKD) stages, including CKD stage 1 with over $90.0 \text{ mL}/\text{min}/1.73\text{m}^2$, CKD stage 2 with the GFR range of $60.0\text{-}89.9 \text{ mL}/\text{min}/1.73\text{m}^2$, CKD stage 3a with the GFR range of $45.0\text{-}59.9 \text{ mL}/\text{min}/1.73\text{m}^2$, and CKD stage 3b with the GFR range of $30.0\text{-}44.9 \text{ mL}/\text{min}/1.73\text{m}^2$. By comparison, the two groups shown in the bottom two panels of Figure 2.5 have an almost monotonically increasing or decreasing trend, respectively.

As mentioned in the section 2.2, it is more efficient to manage the patients when they are clustered into a small number of groups with homogeneous GFR curves. For example, patients in the panel (a) of Figure 2.5 have flat normal GFR trends, indicating that their kidney transplantations are successful. On the other hand, patients in the panel (e) have decreasing GFR trends, and they may need alternative treatments. Partial patients in all 40 groups are shown in the supplementary document.

2.3.3 Detection of GFR trajectory outliers

Detecting abnormal GFR curves can be helpful in practice. As introduced in the subsection 2.3, the order statistic $\sum_{k=1}^K (\xi_{ik}^2/\lambda_k)$ can be used to detect abnormal GFR curves. Figure 2.6 displays some abnormal GFR curves. It shows one patient having the same GFR value of 131 during the 10-year followed-up period, which may be caused by erroneously data recording. Another GFR curve in Figure 2.6 has GFR at 130 in the beginning, then down to 10 after two years. It seems unreasonable for a patient with an extremely healthy kidney function to lose the kidney function so quickly. Another patient has GFR at 10 for 3 years and then increases to 100 in a short time period. On the other hand, in practice, not all

abnormal GFR curves are caused by incorrectly data recording, since some of them may be some interesting clinical cases.

2.3.4 Prediction

FPCA can be used to predict GFR values for the future time or to impute missing GFR values in the past time by using Equation (4.5). In clinical practice, the prediction of future GFR and the recovery of missing GFR can be useful. For instance, the predictions can help to raise a red flag for specific patients if their predicted GFR values continue to decrease or go below certain thresholds.

Figure 2.7 displays the predicted GFR curves for four patients. The top two panels show two patients with GFR measurements in the first few years. Equation (4.5) can be used to predict their GFR in the future. For instance, the patient at the top left panel has the measured GFR values at the first four years, but no GFR values are available afterwards. Our method predicts the future GFR value of this patient at the fifth year to be 67. The bottom two panels show that two patients have GFR missing in some years. Equation (4.5) can also be used to estimate the missing GFR for these two patients. For example, the patient corresponding to the bottom left panel has missing GFR values at the 2nd year. Our method estimates the missing GFR value of this patient in the second year to be 41. The predicted GFR of another patient, shown at the bottom right panel of Figure 2.7, continually decreases and reaches 30 at the seventh year, which means that the kidney function of this patient goes for CKD stage 4 ($\text{GFR} < 30$) at the seventh year. This prediction will raise a red flag for this patient.

2.4 Conclusions and Discussion

GFR curves are ideal biomarker measurements of the kidney function progression after kidney transplant. In this paper, we use the functional principal component analysis (FPCA) method to determine the major source of variations of GFR curves. Four functional principal components (FPCs) are estimated, and they account for 99.8% of the total variations of GFR curves. All these four FPCs have some interesting interpretation. For instance, the second FPC represents the change of GFR after 5.5 years in comparison with the early stage. In addition, the corresponding FPC scores can be used to cluster GFR curves. All 5654 GFR curves are clustered into 40 groups, and each of these 40 groups contains similar GFR curves. We also find that FPC scores can be used to detect abnormal GFR curves, which supplies a useful tool to detect data entry errors or interesting clinical cases in a large dataset. FPCA can also recover missing GFR values, and predict future long-term GFR trajectories.

As one reviewer pointed out, it is of great interest to predict the future GFR for patients with different therapies, complications, gender, or ages. This question can be solved

with two methods. One method is that we can do separate FPCA for different therapy/complication/gender/age groups. The other method is that we can first do the regression of GFR on all variables such as therapies, complications, gender, or ages. Then we do FPCA on the fitted GFR residuals after adjusting these variables. We can then predict the future GFR based on the estimated FPC scores on the GFR residuals and linear coefficients to these variables. This is a very interesting problem, and we will investigate it in our future research.

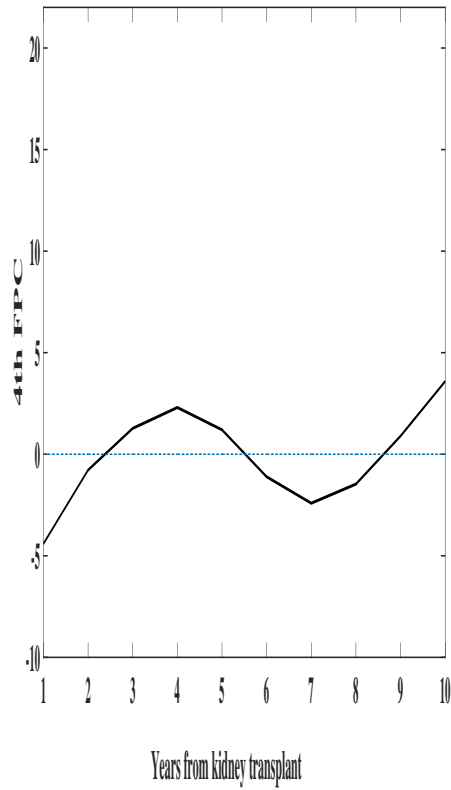
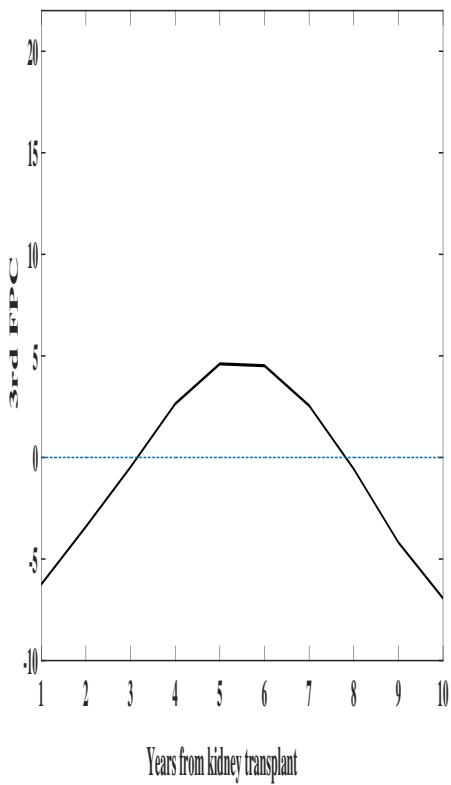
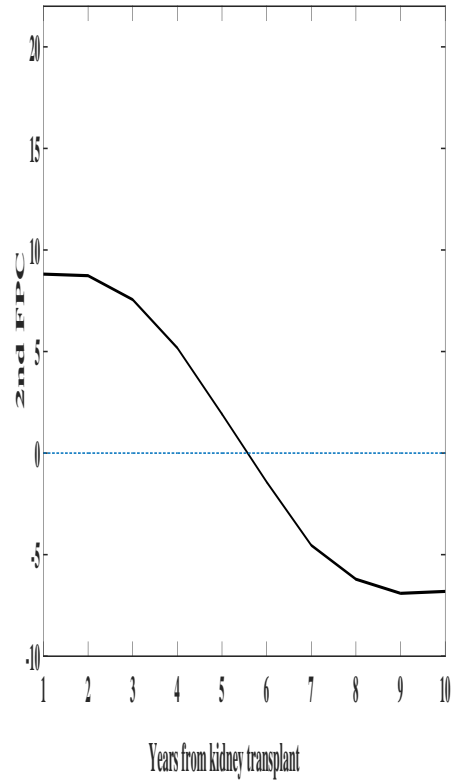
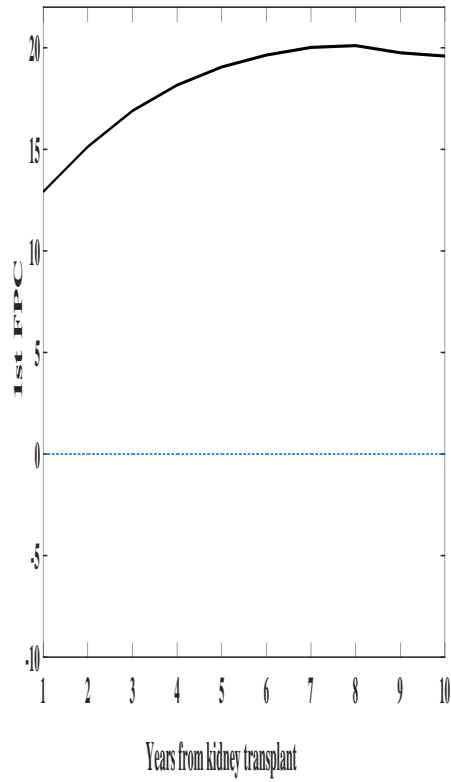


Figure 2.3: The first four leading functional principal components (FPCs) estimated from the GFR curves.

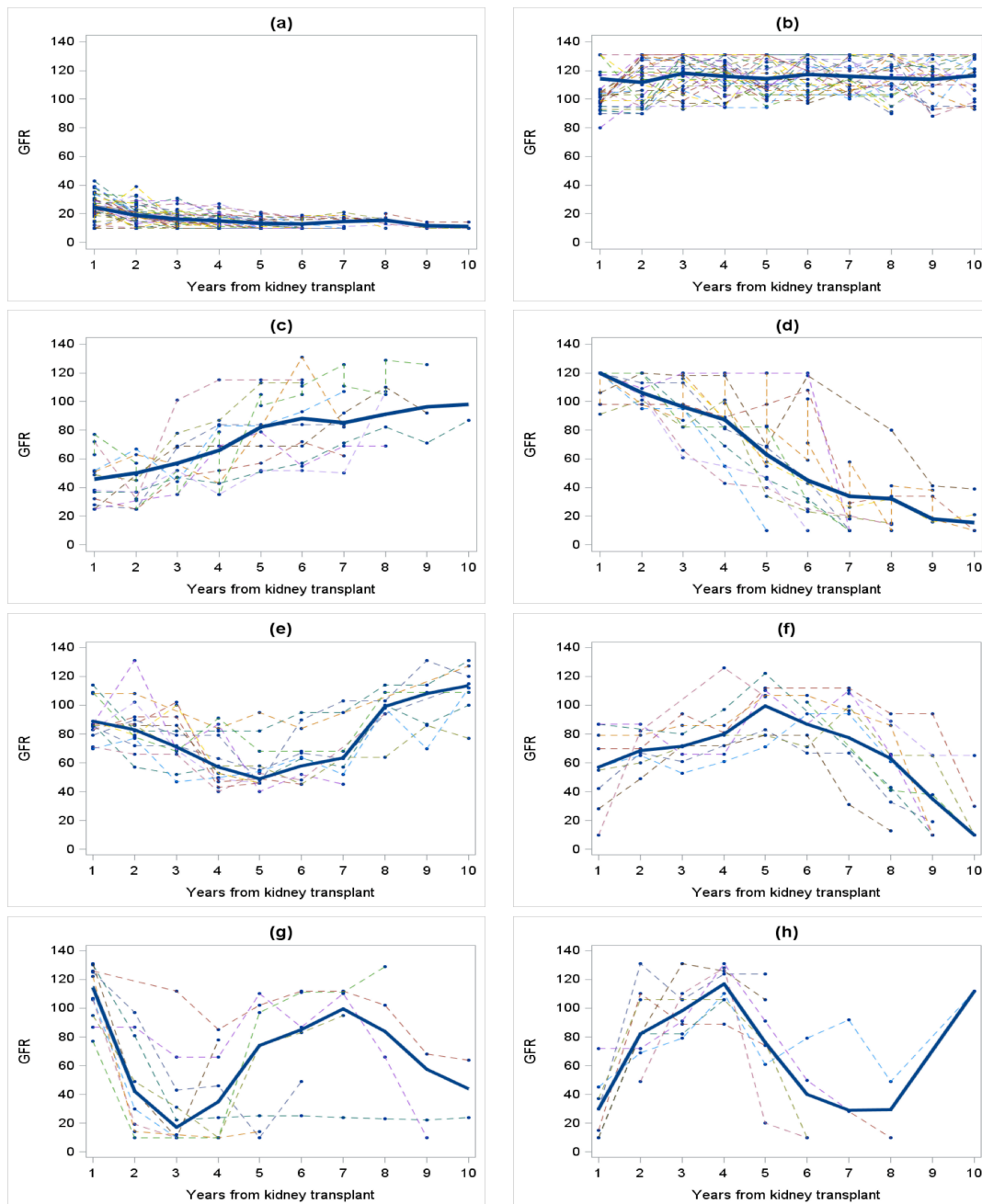


Figure 2.4: GFR curves when their FPC scores are extreme. The thick blue curve in each panel is the average of individual GFR curves in that panel, which represents the common trend in that panel. The left four panels, from top to bottom, are GFR curves when their first, second, third, and fourth FPC scores are smaller than the 5% quantiles, respectively. The right four panels, from top to bottom, are GFR curves when their first, second, third, and fourth FPC scores are larger than the 95% quantiles, respectively.

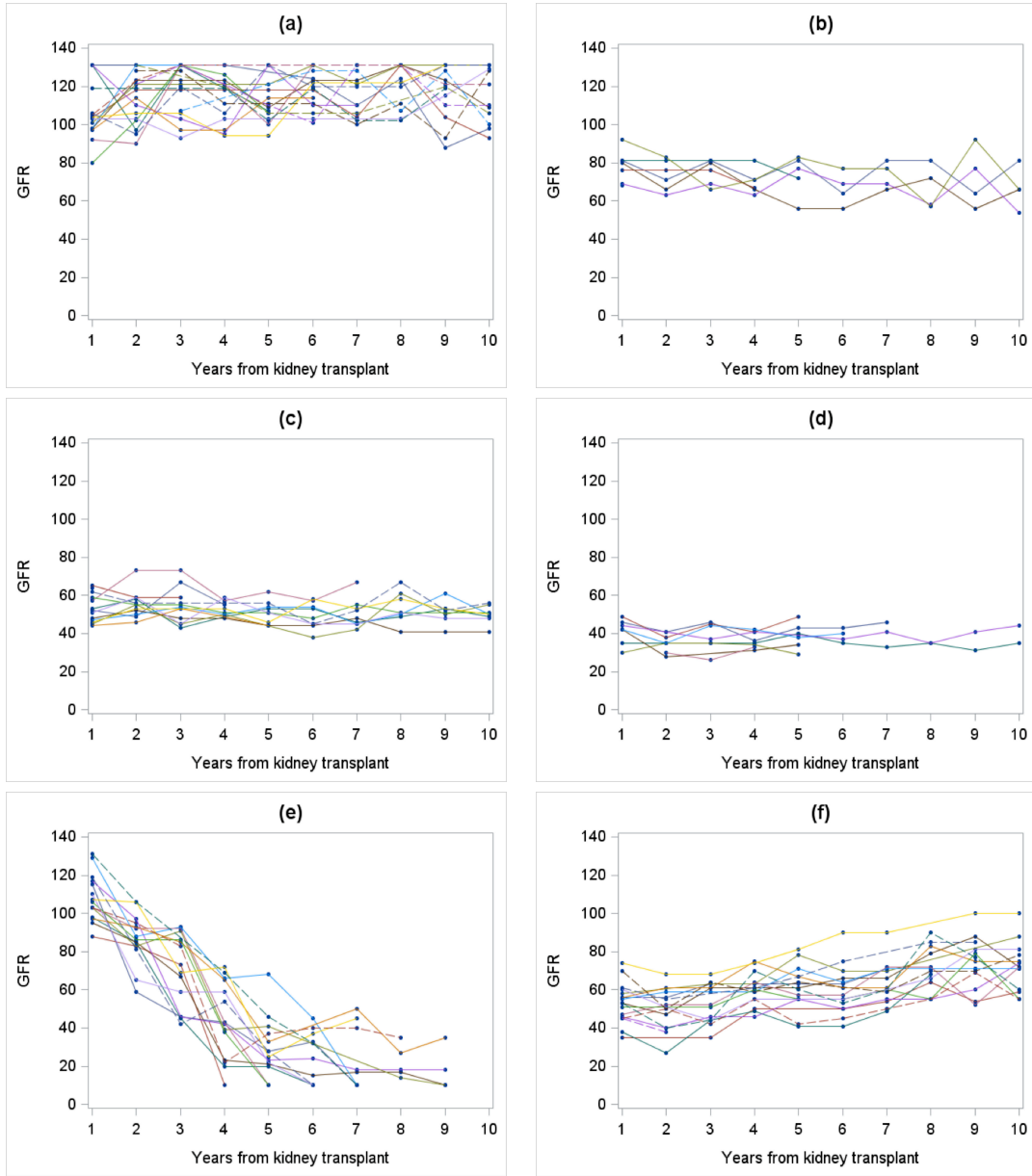


Figure 2.5: Part of the GFR curves in six clusters.

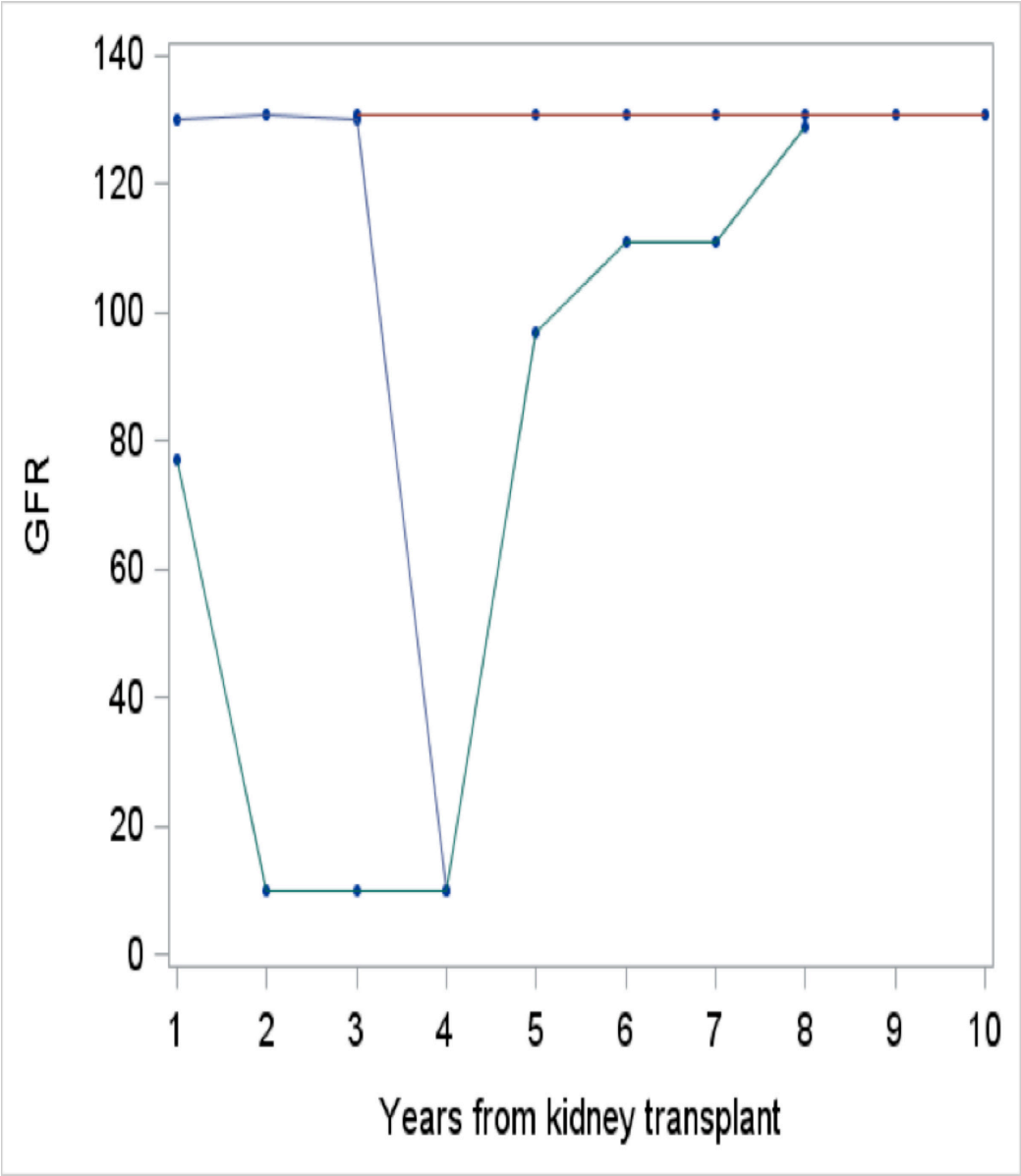


Figure 2.6: Some abnormal GFR curves.

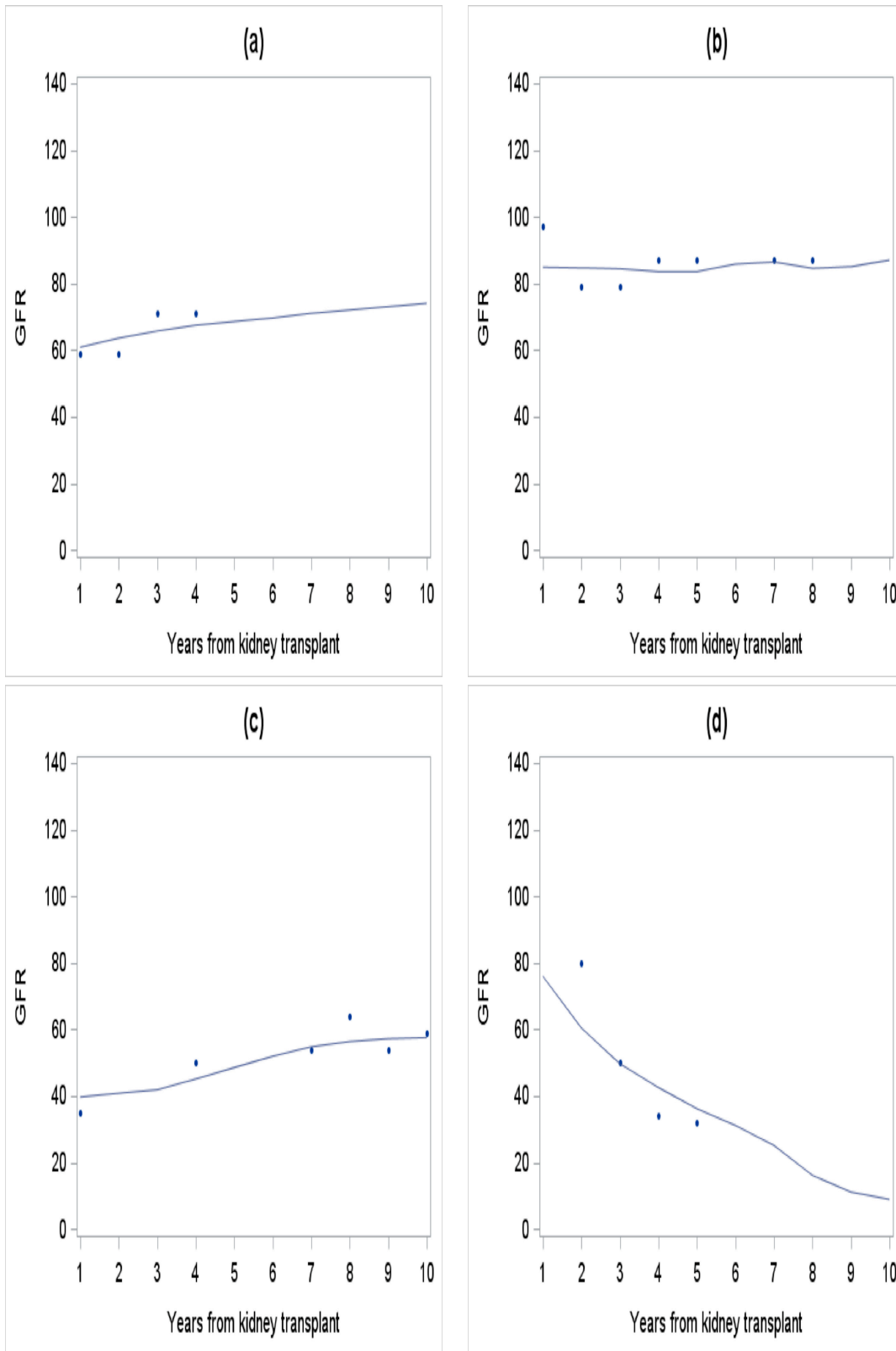


Figure 2.7: The predicted GFR curves for four patients. The dots are observed GFR data.

Chapter 3

A Joint Model of a Longitudinal and Accelerated Failure Time Data and its Application to Transplant Patients with an ESRD and a Diabetes

3.1 Introduction¹

This article is motivated by a longitudinal and time-to-event clinical dataset, where all patients have the end-stage renal disease (ESRD) and diabetes. All patients have already had a kidney transplantation from a living or deceased kidney donor, and all of them are on the waiting list for the pancreas transplant to treat the diabetes disease. Partial patients have a pancreas transplantation if a matched pancreas organ is available for them. It is known that the organ transplantation can prolong the survival of type 1 diabetic patients with ESRD^{40–45,76}. However, how to extend the long-term kidney function still remains the main challenge for transplantation.

No studies have used a joint model to predict the long-term kidney function, which includes the longitudinal continuous outcome of glomerular filtration rate (GFR) and the time-to-event outcome of all-cause graft loss (ACGL). From our preliminary result as in Figure 3.1, the levels of the observed GFR trajectories for patients with ACGL are higher than those of patients without ACGL, and we find that the slopes of GFR trajectories for patients with ACGL are steeper in comparison with patients without ACGL. In addition, all patients may have a pancreas transplant at any time post kidney transplant to treat the diabetic disease, and they are in different statuses at different time points during the

¹This chapter has been published in *Statistical Methods in Medical Research* (2018), Dong J, and Wang S, and Wang S, and Wang L, and Gill J, and Cao J

followed-up period. For example, they are in status 1 (alive without pancreas transplant) at the time of the admission to the waiting list for a pancreas. They move to status 2 (alive with pancreas transplant) if a matched pancreas organ becomes available for them before ACGL, or they directly move to status 3 with ACGL. In fact, the failure rates of the time-to-event outcome are different as patients change their status. As shown in Figure 3.1, it seems that patients who have a pancreas transplant are less likely to have ACGL. In short, the above scenarios motivate us to develop a new dynamical joint model to predict the long-term outcome, since there are at least two advantages in using a joint model. Firstly, joint modelling multiple outcomes together can increase the power and decrease the Type I error^{46,47}. Secondly, this joint model can estimate the parameters in the longitudinal component by incorporating the time-to-event information through censoring, and similarly, for the converse situation, the estimation of the time-to-event is incorporated by the longitudinal data information.

Several methods for estimating joint models of multiple outcomes have been developed. The main challenge in jointly modelling multiple outcomes is the lack of a suitable multivariate joint distribution. The first approach is a two-stage approach⁴⁸, where a random components model is developed to describe repeated longitudinal measures in the first stage, and a Cox proportional hazards model is estimated in the second stage. However, this approach may cause bias when the observation of the longitudinal process is interrupted by the event. To address this problem, the second approach⁴⁹ directly specifies the joint distribution by factorizing it into the conditional distribution of one outcome and a marginal distribution of the other outcome. This approach was reviewed with some insightful comments⁵². The accelerated failure time model is considered in their joint model rather than the Cox proportional hazards model^{54,89}. The third approach directly formulates a joint model for longitudinal repeated measurements and the time-to-event outcome. For instance, a copula is used to construct the joint distribution⁵⁵. Another challenge in jointly modelling multiple outcomes is the intensive computation due to the complex correlation structure of latent variables and measurement errors in covariates^{49,89}. So the EM or the Monte Carlo EM algorithm is developed to estimate parameters in the joint models^{54,89}. The Bayesian method is also developed to estimate the joint models⁵⁹.

Most of above joint models are based on the Cox proportional hazard regression, and only a few joint models such as the paper⁵⁴ use the accelerated failure time regression. As shown in Figure 3.3, the assumption of Cox proportional hazard model fails because the cumulative survival lines cross with each other for patients with/without a pancreas transplant. Therefore, the accelerated failure time submodel is used in our proposed joint model. On the other hand, the proposed joint model is different from the joint models in the paper⁵⁴, which treats the longitudinal component as a covariate in the survival analysis. In our proposed joint model, instead of using the whole longitudinal component as a covariate, we propose to use some latent features of the longitudinal component in

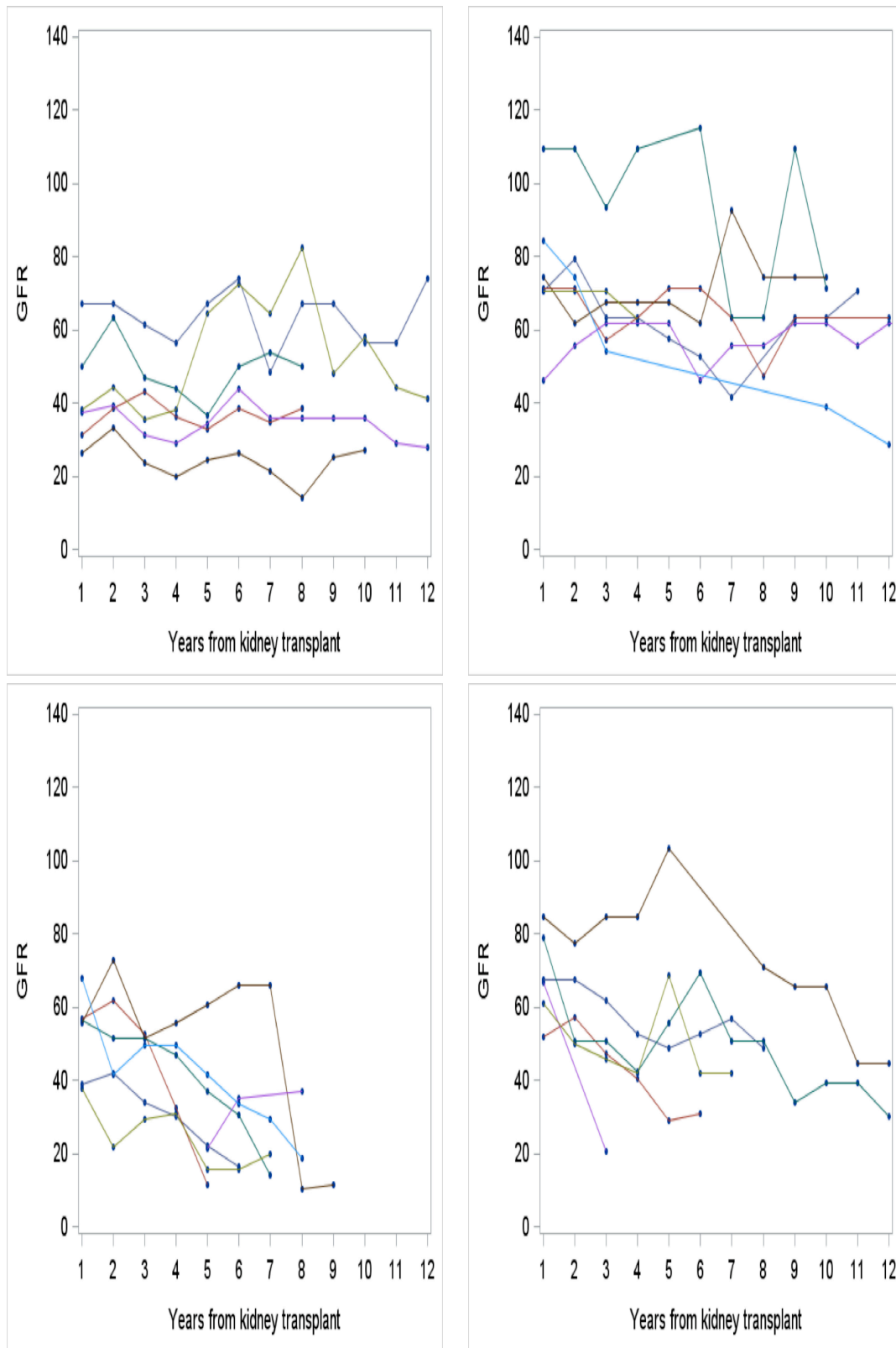


Figure 3.1: Observed individual GFR trajectory curves. The left two panels, from top to bottom, are GFR curves for patients with All-cause graft loss (ACGL) events or without ACGL events, respectively, when they don't have a pancreas transplantation. The right two panels, from top to bottom, are GFR curves for patients with ACGL event or without ACGL events, respectively, when they have a pancreas transplantation. Each color represents the individual patient in each panel.

the survival submodel. Finally, none of the above joint models has considered a method to obtain the dynamical non-proportional hazard ratio curve of a side event when hazard ratios are non-proportional during the followed-up time period.

The main contribution of this paper is that we review the clinical question in the transplantation data, and accordingly develop a new joint model to determine the relationships of multiple outcomes to account for correlations within/between subjects. To the best of our knowledge, it is the first time for the joint model to be applied to the organ transplantation research. Our proposed joint model has three advantages. Firstly, the survival submodel shares a vector of latent variables with the longitudinal submodel. The advantages of this model are that unnecessary noise can be filtered, and the effects of other covariates can be adjusted in the longitudinal submodel. In addition, it is easy to interpret the coefficients from the model results. For example, the latent features are the baseline and the slope of GFR trajectories in the application example. The coefficients in the survival component represent their corresponding relationship with the time-to-event outcome. Secondly, the survival submodel shares the data information together with the longitudinal submodel. For example, our proposed joint model in the application example has considered that the occurrence of death or transplant failure may lead to the censoring of GFR, which overcomes the drawback of separate analyses for each outcome. Finally, our proposed joint model includes a piecewise linear function to display the dynamical non-proportional hazard ratios of the side event on the time-to-event outcome.

The rest of this article is organized as follows. Our proposed joint models are introduced in Section 4.2. We present our estimation method for the joint model in Section 3.3. Section ?? demonstrates the application of our joint model in the transplantation clinical data. Section 4.4 presents three simulation studies to investigate the finite sample performance of our joint model. Conclusions and discussion are given in Section 4.6.

3.2 The Joint Model

Let $Y_i(t_{ij})$ be a repeated continuous measured outcome at times t_{ij} for the i -th subject, where $i = 1, \dots, n$, $j = 1, \dots, m_i$, and m_i is the number of repeated measurements for the i -th subject. For example, the longitudinal outcome $Y_i(t_{ij})$ is the repeated measurements of GFR at different time points in the application example of transplant clinical data. Let T_i be the i -th subject's survival time to the event of interest, C_i be a possible censoring time, $\delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$ be the censoring indicator, $S_i = \min(T_i, C_i)$ be the observed survival time, and $\mathbf{Z}_i = [Z_{i1}, \dots, Z_{iP}]^T$ be the observed covariates for the i -th subject. We propose the following joint model:

$$\begin{cases} Y_i(t_{ij}) = \boldsymbol{\alpha}^T \mathbf{Z}_i + \boldsymbol{\beta}_i^T \boldsymbol{\xi}(t_{ij}) + \epsilon_i, i = 1, \dots, n, \\ \lambda(t|\mathbf{Z}_i, \boldsymbol{\beta}_i, \mathbf{w}_i(t)) = \lambda_0 \left\{ \int_0^t \phi(s, \mathbf{Z}_i, \mathbf{w}_i(s), \boldsymbol{\beta}_i, \boldsymbol{\gamma}) ds \right\} \phi(t, \mathbf{Z}_i, \mathbf{w}_i(t), \boldsymbol{\beta}_i, \boldsymbol{\gamma}), \end{cases} \quad (3.1)$$

The first equation in the joint model (3.1) is the longitudinal submodel for repeated measurement outcome $Y_i(t_{ij})$, where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_P)^T$ is a vector of coefficients for the fixed effects of $\mathbf{Z}_i = [Z_{i1}, \dots, Z_{iP}]^T$, and $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{iL})^T$ is a vector of coefficients for the random effects of $\boldsymbol{\xi}(t) = (\xi_1(t), \dots, \xi_L(t))^T$. Here, $\xi_\ell(t), \ell = 1, \dots, L$, is a parametric function of t . For example, $\xi_1(t) = 1$ and $\xi_2(t) = t$ in our application example. We assume that $\boldsymbol{\beta}_i \sim \text{Normal}(\mathbf{b}, \mathbf{B})$. The vector of measurement errors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{im_i})$ are assumed to be multivariate normal distributed with the mean and the variance-covariance matrix $\sigma^2 \mathbf{I}_n$.

The second equation in the joint model (3.1) is the survival sub-model with the accelerated failure time hazard function, where $\phi(t, \mathbf{Z}_i, \mathbf{w}_i(t), \boldsymbol{\beta}_i, \boldsymbol{\gamma}) = \exp[\boldsymbol{\gamma}_1^T \mathbf{Z}_i + \boldsymbol{\gamma}_2^T \boldsymbol{\beta}_i + \mathbf{w}_i(t|\boldsymbol{\gamma}_3)]$, which represent the joint effects of covariates, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \boldsymbol{\gamma}_2^T, \boldsymbol{\gamma}_3^T)$ are the coefficients in the survival model, and $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{iL})^T$ are the vector of latent variables, which are shared in the longitudinal sub-model and the survival sub-model. The time-dependent indicator function $\mathbf{w}_i(t|\boldsymbol{\gamma}_3)$ captures the dynamic relative risk of the side event at different time points post the side event. Here, $\lambda_0 \{ \int_0^t \phi(s, \mathbf{Z}_i, \mathbf{w}_i(s), \boldsymbol{\beta}_i, \boldsymbol{\gamma}) ds \}$ is the baseline hazard function. This survival sub-model is justified with more details in Subsection 3.2.1.

This proposed joint model has two major advantages. Firstly, the mixed-effect submodel of the longitudinal outcome can adjust for other co-variables to filter some noise because we do not treat the longitudinal outcome as a covariate in the survival submodel. Secondly, the survival submodel shares the latent features $\boldsymbol{\beta}_i$ with the mixed-effect submodel. The estimates of the latent features in the joint model can offer an answer for specific clinical questions. For example, in our kidney transplant application, the latent variable β_{i1} is the baseline of GFR, the latent variable β_{i2} is the slope of GFR. Their corresponding coefficients (γ_{21} and γ_{22}) in the survival model show the effect of the baseline and the slope of GFR to the time-to-event outcome.

3.2.1 The Survival Submodel

This subsection provides the justification for the survival submodel in the proposed joint model with more details. In the transplant clinical data, all subjects have a kidney transplant, and only part of them have a pancreas transplant at a certain time after kidney transplant before death.

From our preliminary analysis, the clinical transplant data has several aspects. Firstly, as shown in Figure 3.1, patients with a pancreas transplant are less likely to have the time-to-event outcome in comparison with patients without a pancreas transplant. Secondly, patients have a dynamical status as shown in Figure 4.2. For instance, each individual is on the waiting-list program for the pancreas transplantation after kidney transplant (status 1). Then patients either move to status 2 (alive and pancreas transplant) when a matched pancreas organ is available, or they directly move to status 3 (ACGL or on the waiting). The hazard rates are different when moving from status 1 to status 3 in comparison with

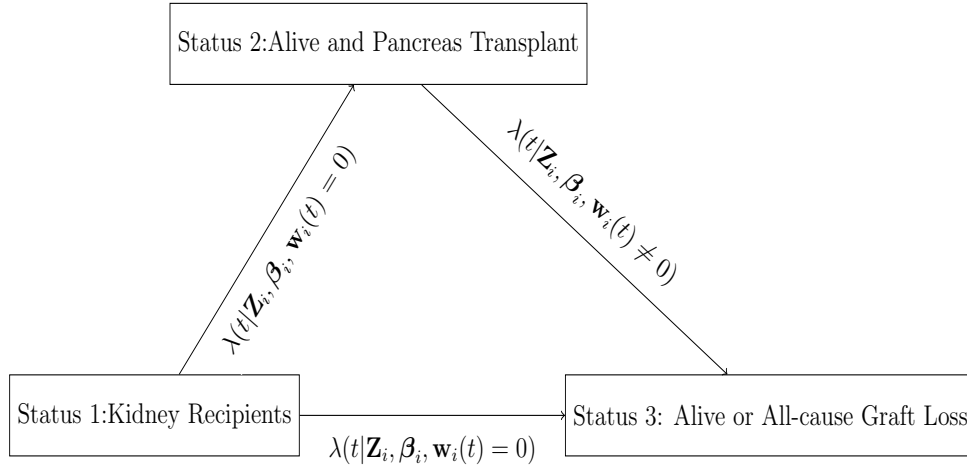


Figure 3.2: The three statuses of kidney transplant patients. All patients start from the date of the kidney transplant (Status 1), then they may move to Status 2 (pancreas transplantation) when a matched pancreas organ is available during the followed-up time period. If not, they directly move to Status 3 when the time-to-event outcome of all-cause graft loss happens, or they still are on the waiting-list for the pancreas transplant.

the other scenario when moving from status 2 to status 3. Thirdly, the assumption of Cox proportional hazard model fails as shown in Figure 3.3, because the cumulative survival line of patients with a pancreas transplant cross with the cumulative survival line of patients who have no pancreas transplant. Therefore, we recommend the alternative hazard model rather than Cox proportional hazards model in this paper.

We propose to use the accelerated failure time (AFT) model, which was firstly introduced by Cox⁶⁰ to determine whether the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. Cox and Oakes⁶¹ extended a AFT model with time-dependent covariates. James⁶² provided a method to estimate the time-dependent AFT model in the presence of confounding factors. Reid⁶³ and Kay⁶⁴ mentioned that the AFT models were more appealing than the the proportional hazard models because they could give direction physical interpretations. For example, as mentioned in the paper by Kay⁶⁴, the AFT model can supply a more straightforward interpretation of the treatment effect on the time to event data because the coefficient of the treatment indicator can be estimated across various intervals defined by the cut time points from the date of treatment.

In this paper, the hazard function of the accelerated failure time submodel is specified as:

$$\lambda(t|\mathbf{Z}_i, \boldsymbol{\beta}_i, \mathbf{w}_i(t)) = \lambda_0 \left\{ \int_0^t \phi(s, \mathbf{Z}_i, \mathbf{w}_i(s), \boldsymbol{\beta}_i, \boldsymbol{\gamma}) ds \right\} \phi(t, \mathbf{Z}_i, \mathbf{w}_i(t), \boldsymbol{\beta}_i, \boldsymbol{\gamma}), \quad (3.2)$$

where $\phi(t, \mathbf{Z}_i, \mathbf{w}_i(t), \boldsymbol{\beta}_i, \boldsymbol{\gamma}) = \exp[\boldsymbol{\gamma}_1^T \mathbf{Z}_i + \boldsymbol{\gamma}_2^T \boldsymbol{\beta}_i + \mathbf{w}_i(t)]$, $\mathbf{Z}_i = [Z_{i1}, \dots, Z_{iP}]^T$ is a vector of time-independent covariates, $\boldsymbol{\beta}_i = (\beta_{i1}, \dots, \beta_{iL})^T$ is a vector of latent variables, which

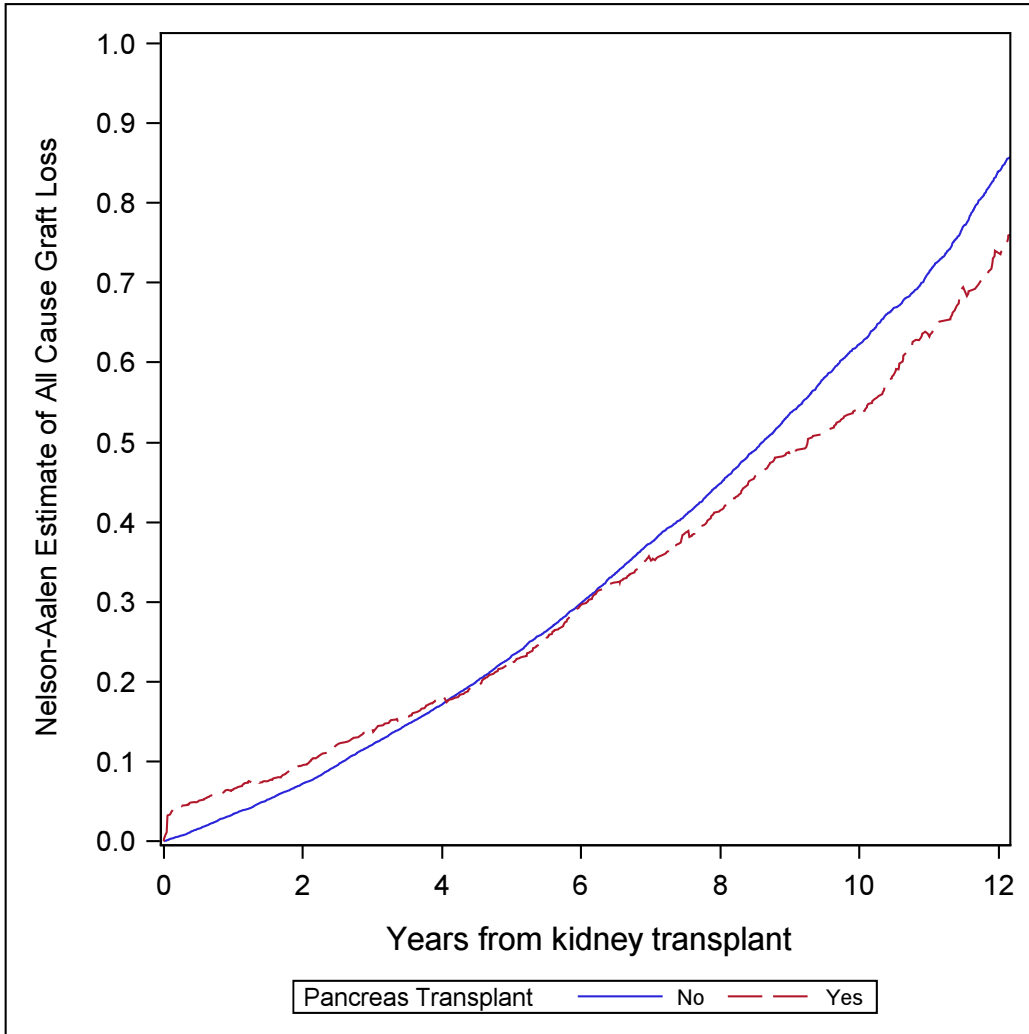


Figure 3.3: The cumulative Nelson-Aalen estimate of all-cause graft loss by patient status of pancreas transplantation. The red line is the cumulative Nelson-Aalen estimate of all-cause graft loss for patient with a pancreas transplant and the blue line is the cumulative Nelson-Aalen estimate of all-cause graft loss for patient without a pancreas transplant.

are random factors shared with the longitudinal sub-model, and $\mathbf{w}_i(t)$ is a time-dependent indicator function. We assume $\mathbf{w}_i(t)$ to be a piecewise linear function, which is used to capture the dynamic relative risk at different time points post the side event. We express $\mathbf{w}_i(t)$ as

$$\mathbf{w}_i(t|\boldsymbol{\gamma}_3) = \begin{cases} 0 & \text{if } t \leq W_i, \\ \gamma_{30} & \text{if } W_i < t \leq W_i + \frac{D_0}{365}, \\ \gamma_{3k}(t - W_i - \frac{D_{k-1}}{365}) & \text{if } W_i + \frac{D_{k-1}}{365} < t \leq W_i + \frac{D_k}{365}, k = 1, \dots, K, \\ \gamma_{3(K+1)} & \text{if } t > W_i + \frac{D_K}{365}, \end{cases} \quad (3.3)$$

where W_i is the time of the side event for the i -th subject, D_0, D_1, \dots, D_K are denoted as the specified number of days post the side event, and $\boldsymbol{\gamma}_3 = (\gamma_{30}, \dots, \gamma_{3(K+1)})^T$ are the coefficients in the piecewise function. In our application example of the clinical transplant data, W_i is the time from the kidney transplant to the pancreas transplant for those patients who have pancreas transplant. For patients without pancreas transplant, W_i is set to be larger than the end date of the study cohort minus the date of kidney transplant.

It is worth mentioning that the hazard function defined in (3.2) can be a strata hazard model because it can provide different hazard functions when patients are in different statuses. For example, the hazard function $\lambda(t|\mathbf{Z}_i, \boldsymbol{\beta}_i, \mathbf{w}_i(t|\boldsymbol{\gamma}_3) = 0)$ is the hazard rate when patients move from status 1 to status 3 if $t \leq W_i$. The hazard function $\lambda(t|\mathbf{Z}_i, \boldsymbol{\beta}_i, \mathbf{w}_i(t|\boldsymbol{\gamma}_3) \neq 0)$ is the hazard rate when patients move from status 2 to status 3 if $t > W_i$.

More importantly, the piecewise linear function (3.3) in the proposed joint model can be used to determine the time-dependent hazard ratios of the side event when the effect of the side event on the time-to-event outcome is non-proportional. For example in our application example, this piecewise linear function can be used to calculate the dynamic relative risk of the pancreas transplant on all-cause graft loss at different time points (D_k) after a pancreas transplant. For instance, we set D_k as $D_0 = 14$ days, $D_1 = 45$ days, $D_2 = 90$ days, $D_3 = 180$ days, $D_4 = 365$ days, and $D_5 = 730$ days from the date of pancreas transplant, and then we can obtain the relative hazard ratios at each time point from the joint model. These relative hazard ratios can supply some references to control the potential risk of the pancreas transplant in the clinical practice. We give the change curve of the relative hazard ratios over time D_k and discuss it in more details in Section 3.4.

3.3 Estimation Method

We discuss the likelihood function in a general framework for this proposed joint model with latent variables. Let $\boldsymbol{\Theta} = (\boldsymbol{\gamma}^T, \boldsymbol{\alpha}^T, \boldsymbol{\lambda}^T, \boldsymbol{T}, \mathbf{B}^T, \sigma^2, \lambda_0)^T$ be the parameters to estimate.

The overall likelihood function based on the observed information is given by:

$$L(\Theta) = \prod_{i=1}^n \left[f(t_i, \mathbf{w}_i(t_i), S_i, \delta_i | Y_i, \mathbf{Z}_i, \gamma, \lambda_0) \left\{ \prod_{j=1}^{m_i} f(Y_{ij} | \mathbf{Z}_i, \beta_i, t_i, \boldsymbol{\alpha}, \sigma^2) \right\} f(\beta_i | \mathbf{b}, \mathbf{B}) \right], \quad (3.4)$$

where

$$f(t_i, \mathbf{w}_i(t_i), \delta_i | Y_i, \mathbf{Z}_i, \gamma, \lambda_0) = \left\{ \lambda_0 \{ \Phi(t, \mathbf{Z}_i, \mathbf{w}_i(t | \gamma_3), \beta_i, \gamma), \gamma) \} \Phi'(t, \mathbf{Z}_i, \mathbf{w}_i(t | \gamma_3), \beta_i, \gamma) \right\}^{\delta_i} \exp \left[- \int_0^{\Phi(t_i, \mathbf{Z}_i, \mathbf{w}_i(t | \gamma_3), \beta_i, \gamma)} \lambda_0(s) ds \right]$$

is the density function of the survival submodel of the proposed joint model, and

$$\begin{aligned} \Phi(t, \mathbf{Z}_i, \mathbf{w}_i(t | \gamma_3), \beta_i, \gamma) &= \int_0^t \phi(s, \mathbf{Z}_i, \mathbf{w}_i(t | \gamma_3), \beta_i, \gamma) ds \\ &= \int_0^t \exp(\gamma_1^T \mathbf{Z}_i + \gamma_2^T \beta_i + \mathbf{w}_i(t | \gamma_3)) ds. \end{aligned}$$

The function $f(Y_{ij} | \mathbf{Z}_i, \beta_i, t_i, \boldsymbol{\alpha}, \sigma^2)$ is the density function of $\text{Normal}(\boldsymbol{\alpha}^T \mathbf{Z}_i + \beta_i^T \xi(t_i), \sigma^2)$, and $f(\beta_i | \mathbf{b}, \mathbf{B})$ is the density function of $\text{Normal}(\mathbf{b}, \mathbf{B})$.

We propose to estimate the parameters in the joint model (3.1) by using the Monte Carlo EM algorithm⁶⁷. The EM-algorithm⁷² is an iterative procedure with two steps: the expectation (E) step and the maximization (M) step. In the E-step, we compute the expectation of joint log-likelihood function over the latent variable β_i using the observations and parameter estimates obtained so far. In the M-step, we maximize the expected joint log-likelihood over the parameters.

3.3.1 E-step

At the t -th iteration of the E-step, the expectation of the log-likelihood function w.r.t the latent variable β_i can be expressed in the following form

$$\begin{aligned} Q(\Theta | \Theta^{(t)}) &= E_{\beta} [\log L(\Theta | t, \mathbf{w}(t), S, \delta, \mathbf{Z}, Y(t)) | \Theta^{(t)}] \\ &= \sum_{i=1}^n \int \left[\log f(t_i, \mathbf{w}_i(t_i), S_i, \delta_i | \beta_i, \gamma, \lambda_0) + \sum_{j=1}^{m_i} \log f(Y_{ij} | \beta_i, \boldsymbol{\alpha}, \sigma^2) \right. \\ &\quad \left. + \log f(\beta_i | \mathbf{B}, \mathbf{B}) \right] f(\beta_i | t, \mathbf{w}_i(t), S_i, \delta_i, \mathbf{Z}_i, Y_i(t), \Theta^{(t)}) d\beta_i, \end{aligned} \quad (3.5)$$

where $f(\beta_i | t_i, \mathbf{w}_i(t), S_i, \delta_i, \mathbf{Z}_i, Y_i(t), \Theta^{(t)}) = \frac{f(\beta_i | \mathbf{Z}_i, Y_i(t), \Theta^{(t)}) f(t_i, \mathbf{w}_i(t), S_i, \delta_i | \beta_i, \Theta^{(t)})}{f(t_i, \mathbf{w}_i(t), S_i, \delta_i | \mathbf{Z}_i, Y_i(t), \Theta^{(t)})}$,

$$f(\beta_i | \mathbf{Z}_i, Y_i(t), \Theta^{(t)}) \sim MVN \left(\mathbf{A}_i \left[\frac{\boldsymbol{\xi}_i^T(t) (Y_i(t) - \mathbf{Z}_i^T \boldsymbol{\alpha})}{\sigma^2} \right], \mathbf{A}_i \right),$$

$\mathbf{A}_i = \left[\frac{\boldsymbol{\xi}_i^T(t)\boldsymbol{\xi}_i(t)}{\sigma^2} + \mathbf{B}^{-1} \right]^{-1}$. The integral in the above equation is intractable because of the intractability of normalizing constant $f(t_i, \mathbf{w}_i(t), S_i, \delta_i | \mathbf{Z}_i, Y_i(t), \Theta^{(t)})$. An alternative is to use the importance sampling to approximate the integral in E-step.

- Draw N samples $\beta_i^{(1)}, \dots, \beta_i^{(N)}$ from $f(\beta_i | \mathbf{Z}_i, Y_i(t), \Theta^{(t)})$ based on the current parameter estimates $\Theta^{(t)}$, and compute the normalized weights $w_i^{(s)} \propto f(t_i, \mathbf{w}_i(t), S_i, \delta_i | \beta_i^{(s)}, \Theta^{(t)})$.
- Calculate $\hat{Q}(\Theta | \Theta^{(t)}) = \sum_{i=1}^n \sum_{s=1}^N w_i^{(s)} \cdot l_i^{(s)}(\Theta | t_i, \mathbf{w}_i(t), S_i, \delta_i, \mathbf{Z}_i, Y_i(t), \Theta^{(t)})$, where $l_i^{(s)} = \log f(t_i, \mathbf{w}(t_i), S_i, \delta_i | \beta_i^{(s)}, \Theta^{(t)}) + \sum_{j=1}^{m_i} \log f(Y_{ij} | \beta_i^{(s)}, \Theta^{(t)}) + \log f(\beta_i^{(s)} | \Theta^{(t)})$.

3.3.2 M-step

After computing the expectation of the log-likelihood function in Equation (3.5), in M-step we estimate each parameter of Θ by maximizing $\hat{Q}(\Theta | \Theta^{(t)})$. The MLEs of $\hat{\mathbf{b}}, \hat{\mathbf{B}}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}^2$, the baseline hazard function $\hat{\lambda}_0(t)$ are derived in the supplementary document. The MLE of γ has no closed-form, hence we could use the numeric optimization algorithm⁶⁸ to optimize this parameter. We repeat the E-step and M-step until convergence achieved. The convergence criterion for MCEM in our numerical study is

$$\max \left\{ \frac{|\Theta^{(t)} - \Theta^{(t-1)}|}{|\Theta^{(t)}| + \epsilon_2} \right\} < \epsilon_1,$$

where we set $\epsilon_1 = 0.002$ and $\epsilon_2 = 0.001$. The standard error of $\hat{\Theta}$ is computed using the bootstrap method⁶⁶.

3.4 Application to Clinical Transplant Data

The clinical transplant data resource is from the United Network for Organ Sharing. As mentioned in the introduction, all patients ($N = 13,635$) have both an end-stage renal disease (ESRD) and a diabetic disease. In this data, all patients already have a kidney transplantation from a living or deceased donor, and all of them are on the waiting list for the pancreas transplant. A part of patients ($N = 2,776$) may have a pancreas transplant at any time during the followed-up period. We apply the proposed joint model to this clinical transplant data in this section. The main result from the proposed joint model is shown in Section 3.4.1, and the effect of the side event of pancreas transplant on the all-cause graft loss is shown in Section 3.4.2.

3.4.1 Main Results from the Joint Model

In order to demonstrate the feasibility of the proposed joint model (3.1), we apply it to some clinical transplant data in this section. As shown in Figure 3.1, the baseline of GFR and the

slope of GFR are related to the time-to-event outcome as mentioned before. So we choose two latent factors β_{i1} and β_{i2} , and the joint model can be specified as in the following:

$$\begin{cases} Y_i(t) = \boldsymbol{\alpha}^T \mathbf{Z}_i + \beta_i^T \boldsymbol{\xi}(t)_i, i = 1, \dots, n, \\ \lambda(t|\mathbf{Z}_i, \mathbf{w}_i(t), \boldsymbol{\beta}_i, \boldsymbol{\gamma}) = \lambda_0 \left\{ \int_0^t \phi(s, \mathbf{Z}_i, \mathbf{w}_i(s), \boldsymbol{\beta}_i, \boldsymbol{\gamma}) ds \right\} \phi(t, \mathbf{Z}_i, \mathbf{w}_i(t), \boldsymbol{\beta}_i, \boldsymbol{\gamma}), \end{cases} \quad (3.6)$$

where $\phi(t, \mathbf{Z}_i, \mathbf{w}_i(t), \boldsymbol{\beta}_i, \boldsymbol{\gamma}) = \exp[\boldsymbol{\gamma}_1^T \mathbf{Z}_i + \boldsymbol{\gamma}_2^T \boldsymbol{\beta}_i + \mathbf{w}_i(t|\boldsymbol{\gamma}_3)]$, and $Y_i(t)$ is the GFR value at various time points post kidney transplant. The GFR value is calculated according to the formula in the paper⁷⁵: $GFR = 141 \times \min(\text{Scr}/d, 1)^e \times \max(\text{Scr}/d, 1)^{-1.209} \times 0.993^{\text{Age}} \times (1.018 \text{ if female}) \times (1.159 \text{ if black})$, where Scr is the measured serum creatinine in mg/dL, and the serum creatinine is a chemical waste product from the muscle metabolism and blood. The parameter $d = 0.7$ if female or 0.9 if male, and the parameter $e = -0.329$ if female or -0.411 if male. Let W_i be the time from the kidney transplant to the pancreas transplant or the time on the waiting-list for patients without a pancreas transplant. We specify the piecewise linear function $\mathbf{w}_i(t)$ in $\phi(t_i, \mathbf{Z}_i, \mathbf{w}_i(t), \boldsymbol{\beta}_i, \boldsymbol{\gamma})$ in (3.6) as follows:

$$\mathbf{w}_i(t|\boldsymbol{\gamma}_3) = \begin{cases} 0 & \text{if } t \leq W_i \\ \gamma_{30} & \text{if } W_i < t \leq W + \frac{14}{365} \\ \gamma_{31}(t - W_i - \frac{14}{365}) & \text{if } W_i + \frac{14}{365} < t \leq W_i + \frac{45}{365} \\ \gamma_{32}(t - W_i - \frac{45}{365}) & \text{if } W_i + \frac{45}{365} < t \leq W_i + \frac{90}{365} \\ \gamma_{33}(t - W_i - \frac{90}{365}) & \text{if } W_i + \frac{90}{365} < t \leq W_i + \frac{180}{365} \\ \gamma_{34}(t - W_i - \frac{180}{365}) & \text{if } W_i + \frac{180}{365} < t \leq W_i + \frac{365}{365} \\ \gamma_{35}(t - W_i - \frac{365}{365}) & \text{if } W_i + \frac{365}{365} < t \leq W_i + \frac{730}{365} \\ \gamma_{36} & \text{if } t > W_i + \frac{730}{365}. \end{cases} \quad (3.7)$$

Table 5.5.2 displays the coefficients and standard errors of all parameters in the longitudinal sub-model and the AFT survival sub-model. The baseline term (β_1) in the mixed-effects submodel is 48.94, which indicates that most patients have a good kidney function at the baseline. The slope term ($\beta_2 = -1.36$) of GFR is negative and statistically significant, which means that the kidney function progression decreases during the followed-up period time. The estimates for other coefficients in the mixed-effect submodel are also reasonable. For example, the value of GFR decreases by 0.17 as the age of patients increases by 1. In other words, the kidney function of older patients is worse than young patients. The average GFR value of patients with a deceased donor is 1.15 less than patients with a living donor.

In the survival sub-model, the coefficients (γ_{21} and γ_{22}) of the random intercept and slope (β_1 and β_2) of GFR are negative, and they are also statistically significant. These results indicate that the latent baseline level and the latent slope of GFR are related to the time-to-event outcome ACG. In other words, the failure rate of ACG increases as

Table 3.1: Estimates for parameters in Model (5.1). The standard errors of the estimates are given in brackets.

Parameters	The longitudinal submodel		The survival submodel	
	Coef.(SE)	<i>P</i> value	Coef.(SE)	<i>P</i> value
Age (per year)	-0.17(0.05)	< 0.001	0.02(0.01)	0.044
Female	5.39(0.73)	< 0.001	-0.23(0.03)	0.029
Black	-3.69(1.34)	< 0.001	0.17(0.08)	0.020
Other	-6.45(1.08)	< 0.001	0.23(0.07)	< 0.001
TX era 1993 – 1997	7.56(1.15)	< 0.001	-0.29(0.04)	0.080
TX era 1998 – 2002	10.75(1.16)	< 0.001	-0.76(0.03)	< 0.001
TX era 2003 – 2007	16.52(1.30)	< 0.001	-0.95(0.03)	< 0.001
PKPRA 1 – 29	-0.93(0.17)	0.024	0.06(0.01)	0.030
PKPRA 30 – 100	-2.35(0.11)	0.159	0.27(0.13)	0.049
HLA mismatch 1 – 6	-1.54(0.61)	0.045	0.24(0.05)	0.004
Dialysis time 0.1 – 1 years	-0.27(0.17)	0.689	0.07(0.01)	0.005
Dialysis time 1.1 – 2 years	-0.52(0.42)	0.166	0.08(0.01)	0.007
Dialysis time 2.1 – 3 years	-0.73(1.01)	0.142	0.33(0.03)	0.028
Dialysis time > 3 years	-0.96(0.11)	0.029	0.38(0.05)	< 0.001
Decreased Donor	-1.15(0.18)	0.015	0.14(0.05)	0.024
β_1	48.94(2.34)	< 0.001		
β_2	-1.36(0.12)	< 0.001		
γ_{21}			-0.07(0.01)	< 0.001
γ_{22}			-0.21(0.04)	< 0.001
γ_{30}			1.22(0.01)	< 0.001
γ_{31}			-9.42(0.04)	0.035
γ_{32}			-2.51(0.05)	0.041
γ_{33}			-0.65(0.45)	0.542
γ_{34}			-0.35(0.21)	0.251
γ_{35}			-0.06(0.01)	0.045
γ_{36}			-0.28(0.04)	< 0.001
Random-effect parameters		Value(Std.Error)	Correlation	
SD(β_1)		14.91(2.30)	-0.40	
SD(β_2)		2.89(0.12)		

the value of GFR decreases during the followed-up time period, and patients in the higher baseline of GFR are less likely to have the time-to-event outcome ACGL. The coefficients of female patients are larger than male patients. It is reasonable that patient age is significantly related to all-cause graft loss, which indicates that patients are more likely to have an ACGL with the hazard ratio (1.02) as patient age increases per year. Compared with the white patients, the black patients are more likely to have the time-to-event outcome ACGL. The transplantation era and the dialysis duration before transplant are also related to the time-to-event outcome ACGL. For example, patients have a longer dialysis duration before kidney transplant, and the more likely patients have all-cause graft failure. Compared with patients who have a deceased donor, patients who have a living donor transplant are less likely to have the time-to-event outcome ACGL. The dynamic effect of the pancreas transplant on all-cause graft loss is presented in the next subsection.

3.4.2 Effect of Pancreas Transplant on Allograft

In order to evaluate the average and time-varying relative risk of the pancreas transplant on all-cause graft loss, we can set the piecewise linear function $\mathbf{w}_i(t|\boldsymbol{\gamma}_3)$ in Equation (3.3) in two separate forms. In order to evaluate the average relative risk of the pancreas transplant on all-cause graft loss, we set $\mathbf{w}_i(t|\boldsymbol{\gamma}_3)$ as:

$$\mathbf{w}_i(t|\boldsymbol{\gamma}_3) = \begin{cases} 0 & \text{if } t \leq W_i, \\ \gamma_{31} & \text{if } t > W_i. \end{cases}$$

Then the coefficient vector $\boldsymbol{\gamma}_3$ in the piecewise function has only one element γ_{31} . In fact, the coefficient γ_{31} represents the average relative risk of the side event on the time-to-event outcome in this case when we set $K = 0$ and $D_0 = 0$ in Equation (3.3). We find that the pancreas transplant has a significantly statistical benefit effect on ACGL because the hazard ratio is $\exp(\gamma_{31}) = \exp(-0.13) = 0.88$ with the p-value 0.045. In other words, the pancreas transplant can reduce the risk of the time-to-event outcome ACGL.

However, the relative risk of this side event on the time-to-event outcome is non-proportional. Therefore, we need to display the relative risk curve at various time points post pancreas transplant. It is also useful to control the potential risk for the clinical practice if we can determine the relative risk at specified time points. In order to evaluate the time-varying relative risk of the pancreas transplant on all-cause graft loss, we set the piecewise linear function $\mathbf{w}_i(t|\boldsymbol{\gamma}_3)$ as the formula (3.7) after specifying the values of D_k as $D_0 = 14$ days, $D_1 = 45$ days, $D_2 = 90$ days, $D_3 = 180$ days, $D_4 = 365$ days, and $D_5 = 730$ days. Then we obtain the coefficient vector $\boldsymbol{\gamma}_3$ at these specified time points from the date of pancreas transplant in comparison with patients without pancreas transplant, which is shown in Table 5.5.2.

For easy comparison, we transform the estimated coefficients into the hazard ratios. Figure 3.4 displays the hazard ratio curve of pancreas transplant at different time points. It shows that the hazard ratio is very high in the beginning because of the clinical surgery or organ acute rejection, then the hazard ratio decreases to 1.00 at 152 days from the date of pancreas transplant, and then becomes less than 1 thereafter. From the time point when the hazard ratio is equal to 1.00, the pancreas transplantation starts to have a survival benefit. It is a good clinical example to demonstrate the hazard ratio curve when the hazard ratios are not proportional.

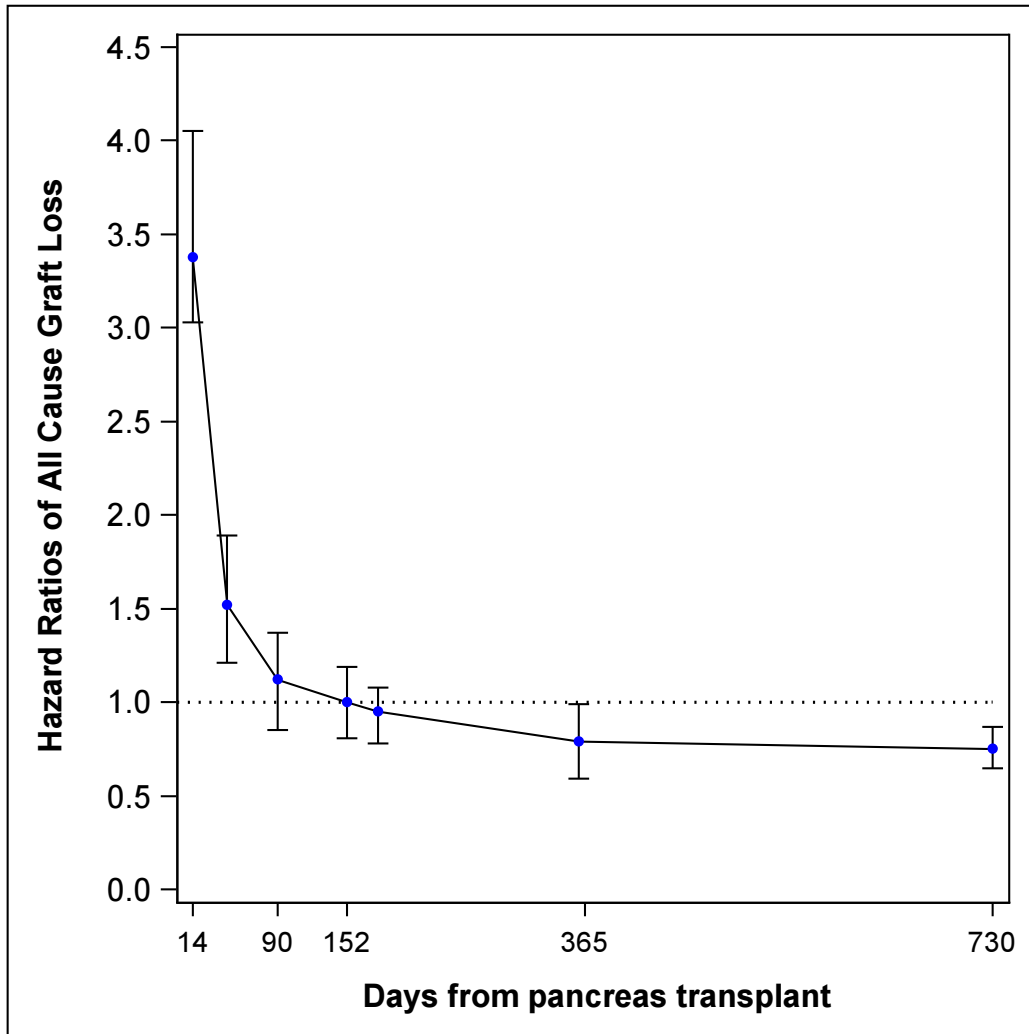


Figure 3.4: The curve of hazard ratios of all-cause graft loss for patients with a pancreas transplant with the 95% confidence intervals at 14, 45, 90, 152, 180, 365, 730 days from the date of pancreas transplant. The reference group are patients without a pancreas transplant. The hazard Ratio curve reaches 1.00 at 152 days from the date of pancreas transplant.

3.5 Simulations

3.5.1 Simulation 1

The first simulation study is implemented to access the finite-sample performance of our proposed MCEM algorithm in Section 3.3. A multivariate mixed-effects model is chosen to simulate the longitudinal trajectories:

$$Y_i(t) = \sum_{p=1}^{15} \alpha_p Z_{ip} + \beta_{i1} + \beta_{i2}t + \epsilon_{ij},$$

where $\beta_{i1} = \beta_1 + b_{i1}$, $b_{i1} \sim \text{Normal}(0, \sigma_1^2)$, $\beta_{i2} = \beta_2 + b_{i2}$, and $b_{i2} \sim \text{Normal}(0, \sigma_2^2)$, $i = 1, \dots, n$. In our application example, the longitudinal outcome is GFR. Here $\alpha_1, \alpha_2, \dots, \alpha_{15}$ are the coefficients for age, gender, and other fixed covariates shown in Table 5.5.2. We set $\beta_1 = 48.94$, $\sigma_1 = 14.91$, $\beta_2 = -1.36$, and $\sigma_2 = 2.89$, which are the estimate from the real data in Table 5.5.2. The measurement error $\epsilon_{ij} \sim \text{Normal}(0, \sigma_3^2)$, where we set $\sigma_3 = 0.85$ and $j = 1, \dots, 12$. The scheduled measurement times of the repeated longitudinal outcome are set at the sequence year $(1, 2, \dots, 12)$ for each subject, but there are no measurements available after death or censoring time. The time-to-event T_i is specified as follows:

$$\text{Log}(T_i) = \sum_{p=1}^{15} \gamma_{1p} Z_{ip} + \gamma_{21} \beta_{i1} + \gamma_{22} \beta_{i2} + w(t|\gamma_3) + \tau_i,$$

where $\gamma_1, \gamma_2, \dots, \gamma_{15}$ are the coefficients for age, gender, and other fixed covariates, γ_{21} and γ_{22} are the coefficient for the random effects β_{1i} and β_{2i} , and the random error $\tau_i \sim \text{Gumbel}(0, 1)$. We set their true values as the estimates from the real data shown in Table 5.5.2. The piecewise linear function $w_i(t|\gamma_3)$ is specified as follows:

$$w_i(t|\gamma_3) = \begin{cases} 0 & \text{if } t \leq W_i \\ \gamma_{31} & \text{if } t > W_i, \end{cases}$$

where $\gamma_{31} = -0.13$. The number of subjects are set as $n = 100$.

We estimate the joint model (5.1) with the Monte Carlo EM algorithm from the simulated data. The simulation procedure is repeated for 100 replicates. Table B.1 shows the parameter estimates, together with biases and root mean square errors (RMSEs). It shows that the means of the parameter estimates by the MCEM algorithm are close to their true values. The average number of iterations till convergence is 12. We notice that the estimate for β_1 and β_1 has large RMSE, which is caused by the setting of our simulated data. In the simulation data, we set $\beta_1 = 48.94$, $\sigma_1 = 14.91$, $\beta_2 = -1.36$, and $\sigma_2 = 2.89$, which are the estimated from the real transplant data. We find that the MCEM algorithm can estimate

Table 3.2: Means, biases, root mean square errors (RMSEs) of the parameter estimates for the joint model (5.1) using our proposed MCEM algorithm in Simulation 1.

Parameters	The longitudinal submodel				The survival submodel			
	True	Mean	Bias	RMSE	True	Mean	Bias	RMSE
Age (per year)	-0.17	-0.17	0.00	0.012	0.02	0.02	-0.00	0.001
Female	5.39	5.41	-0.02	0.236	-0.23	-0.23	-0.00	0.010
Black	-3.69	-3.68	-0.01	0.235	0.17	0.17	0.00	0.009
Other	-6.45	-6.45	0.00	0.220	0.23	0.23	0.00	0.010
TX era 1993 – 1997	7.56	7.57	-0.01	0.300	-0.29	-0.29	0.00	0.010
TX era 1998 – 2002	10.75	10.80	-0.05	0.245	-0.76	-0.76	-0.00	0.010
TX era 2003 – 2007	16.52	16.55	-0.03	0.227	-0.95	-0.95	0.00	0.011
PKPRA 1 – 29	-0.93	-0.93	-0.00	0.114	0.06	0.06	0.00	0.003
PKPRA 30 – 100	-2.35	-2.43	-0.08	0.119	0.27	0.27	0.00	0.011
HLA Mismatch 1 – 6	-1.54	-1.54	0.00	0.132	0.24	0.24	-0.00	0.010
Dialysis time 0.1 – 1 years	-0.27	-0.27	0.00	0.121	0.07	0.07	0.00	0.005
Dialysis time 1.1 – 2 years	-0.52	-0.49	-0.03	0.118	0.08	0.08	-0.00	0.005
Dialysis time 2.1 – 3 years	-0.67	-0.66	-0.01	0.147	0.33	0.33	-0.00	0.004
Dialysis time > 3 years	-0.96	-0.94	-0.02	0.163	0.38	0.38	-0.00	0.005
Decreased Donor	-1.15	-1.14	-0.01	0.109	0.14	0.14	0.00	0.004
β_1	48.94	49.35	-0.41	1.439				
β_2	-1.36	-1.47	0.11	0.505				
γ_{21}					-0.07	-0.07	-0.00	0.002
γ_{22}					-0.21	-0.21	-0.00	0.005
γ_3					-0.13	-0.13	0.00	0.005

parameters accurately in the proposed joint model, which is consistent with the literature such as^{54,89}.

3.5.2 Simulation 2

In order to study the effect of the various correlation construction between the longitudinal submodel and the survival model, we develop two simulation studies in this subsection.

The relationship between the longitudinal submodel and the survival model in our proposed joint model (3.1) is based on latent features. Some other studies treat the longitudinal outcome as a covariable in the survival model as shown in the models (3.8) such as^{48,52,54}.

$$\begin{cases} Y_i(t) = X_i(t) + \epsilon_i, i = 1, \dots, n, \\ \lambda(t|X(t)) = \lambda_0 \left\{ \int_0^t \gamma_1 X(s) ds \right\} \exp(\gamma_1 X(t)). \end{cases} \quad (3.8)$$

The simulation data are generated in two scenarios. In the first scenario, we set $\gamma_1 = \gamma_2 = 1.00$. In the second scenario, we set $\gamma_1 = 1.00$ and $\gamma_2 = -1.00$. We choose these two different scenarios because we want to see the differences between the proposed model (5.1) and the model (3.8) when the effects of the intercept and slope of GFR curves are in

the same or opposite direction. A mixed-effects model is chosen to mimic the longitudinal trajectories:

$$Y_i(t) = \beta_{i1} + \beta_{i2}t + \epsilon_i,$$

where $\beta_{i1} = \beta_1 + b_{i1}$, $b_{i1} \sim \text{Normal}(0, \sigma_1^2)$, $\beta_{i2} = \beta_2 + b_{i2}$, $b_{i2} \sim \text{Normal}(0, \sigma_2^2)$, $i = 1, \dots, n$. Here we set the true values for these parameters as $\beta_1 = 2.50$, $\sigma_1 = 1$, $\beta_2 = -0.20$, and $\sigma_2 = 0.02$. The random measurement error $\epsilon_i \sim \text{Normal}(0, 1)$, and the number of subjects are set as $n = 100$. The preliminary scheduled measurement time of the longitudinal outcome are set at the sequence year $(1, 2, \dots, 12)$ for each subject, but there are no measurements available after death or censoring time. The time-to-event T_i is specified as follows:

$$\text{Log}(T_i) = \gamma_1\beta_{i1} + \gamma_2\beta_{i2} + \tau_i,$$

where the random error $\tau_i \sim \text{Gumbel}(0, 1)$. Note that our proposed joint model (3.1) and the alternative model (3.8) treat $Y_i(t)$ in two different ways. Our proposed joint model (3.1) chooses a mixed-effects submodel for $Y_i(t)$, and shares two random parameters β_{i1} and β_{i2} with the AFT survival submodel. The alternative model (3.8) treats $X_i(t)$ as a covariate in the AFT survival component.

Table 3.3: Means and standard deviations (STD) of the parameter estimates for our proposed joint model (3.1) and the model (3.8) in Simulation 2.

Parameters	Scenario 1		Scenario 2	
	Mean (STD)	Mean (STD)	Mean (STD)	Mean (STD)
True value	γ_1	γ_2	γ_1	γ_2
Fitted value in Model ^(5.1)	1.00	1.00	1.00	-1.00
Fitted value in Model ^(3.8)	0.98(0.05)	1.00(0.05)	1.02(0.05)	-0.98(0.01)
	0.64(0.03)	-	-0.24(0.06)	-

We estimate the joint model (3.1) with the Monte Carlo EM algorithm from the simulated data. The simulation procedure is repeated for 100 replicates. The average number of steps till convergence is 32. Table 3.3 displays the parameter estimates, together with their estimated standard errors. In Scenario 1, when the coefficients ($\gamma_1 = \gamma_2 = 1.00$) of the intercept (β_{i1}) and the slope (β_{i2}) are same, the estimated coefficient $\hat{\gamma}_1$ for the model in⁵⁴ has the same sign as the true value γ_1 , although there is a relatively large gap between them. In Scenario 2 when the coefficients ($\gamma_1 = 1.00$ and $\gamma_2 = -1.00$) are different, the estimated coefficient $\hat{\gamma}_1$ from the model (3.8) is completely different from the true value. In summary, the results from Table 3.3 demonstrate that model (3.8) cannot describe both the relationship between the intercept (β_{i1}) and the slope (β_{i2}) of the longitudinal outcome with the time-to-event outcome by a single parameter γ_1 . Especially in Scenario 2 when the intercept and the slope are in an opposite relationship with the time-to-event outcome ($\gamma_1 = 1.00$ and $\gamma_2 = -1.00$), it is impossible to describe the two relationships by a single

parameter γ_1 . This simulation example shows the advantages in our proposed joint model by using the features from the longitudinal submodel in the survival submodel.

The second advantage of our proposed joint model is that the estimated results offer a straightforward interpretation. For example, in Scenario 2, if patients have a higher baseline, i.e. a larger β_{i1} , then patients are more likely to have the time-to-event outcome. So physicians can tell patients which level they are in and the corresponding risk to have the time-to-event given a baseline value. Similarly, if patients have a larger value of the slope i.e. a larger β_{i2} , then patients are less likely to have a time-to-event outcome. So physicians can tell patients the trend of the longitudinal outcome and the corresponding risk to have a time-to-event given the value of the slope.

3.5.3 Simulation 3

In this subsection, we investigate the effect of misspecification of the distribution of random effects on parameter estimates. A mixed-effects model is chosen to mimic the longitudinal trajectories:

$$Y_i(t) = \beta_{i1} + \beta_{i2}t + \epsilon_i,$$

where the random effects β_{i1} and β_{i2} are sampled in two scenarios. In the first scenario, β_{i1} and β_{i2} are sampled from the normal distribution $\beta_{i1} = \beta_1 + b_{i1}$, $b_{i1} \sim \text{Normal}(0, \sigma_1^2)$, $\beta_{i2} = \beta_2 + b_{i2}$, $b_{i2} \sim \text{Normal}(0, \sigma_2^2)$, $i = 1, \dots, n$. Here we set the true values for these parameters as $\beta_1 = 2.50$, $\sigma_1 = 1$, $\beta_2 = -0.20$, and $\sigma_2 = 0.02$. In the second scenario, β_{i1} and β_{i2} are sampled from a bimodal mixture of normal distributions, where we set $\beta_{i1} \sim 0.55 \cdot N(3, 0.7^2) + 0.45 \cdot N(1, 0.5^2)$ and $\beta_{i2} \sim 0.55 \cdot N(-0.3, 0.03^2) + 0.45 \cdot N(-0.1, 0.01^2)$. The random measurement error $\epsilon_i \sim \text{Normal}(0, 1)$, $i = 1, \dots, n$. The number of subjects are set to be $n = 100$. The preliminary scheduled measurement times of the longitudinal outcome are set at the sequence year $(1, 2, \dots, 12)$ for each subject, but there are no measurements available after the date of time-to-event outcome or censoring time. The time-to-event T_i is specified as follows:

$$\text{Log}(T_i) = \gamma_1 \beta_{i1} + \gamma_2 \beta_{i2} + \tau_i,$$

where $\gamma_1 = 1.00$, $\gamma_2 = 1.00$, and the random error $\tau_i \sim \text{Gumbel}(0, 1)$.

We estimate the joint model (3.1) with the Monte Carlo EM algorithm from the simulated data. Therefore, the first scenario has the correct model assumption and the second scenario has the misspecified model assumption. The simulation procedure is repeated for 100 replicates. The average number of steps till convergence is 30. Table 3.4 displays the summary of the simulation results in the two scenarios. In comparison with the simulation results when the distribution for the random effects are correctly specified, the simulation results are similar when the distribution for the random effects are incorrectly specified as the bimodal mixture of normal distributions.

Table 3.4: The mean, bias, standard deviation (STD), and root mean squared error (RMSE) of the parameter estimates for the joint model (3.1) when the model assumption is correct or misspecified in Simulation 3.

Model Assumption	Correct		Misspecified	
Parameters	γ_1	γ_2	γ_1	γ_2
True value	1.000	1.000	1.000	1.000
Mean	0.984	1.002	0.986	0.996
Bias	-0.016	0.002	-0.014	-0.004
STD	0.053	0.049	0.051	0.053
RMSE	0.055	0.048	0.053	0.053
95%CI	96%	95%	96%	96%

3.6 Conclusions and Discussion

This paper is motivated by a longitudinal and time-to-event transplant data set. Our proposed joint model has a longitudinal submodel and an AFT sub-model, and both submodels share a vector of latent variables with each other. Our proposed joint model has three major advantages. Firstly, as shown in Table 3.3, the model⁵⁴ can not correctly describe the relationships between the time-to-event outcome and the longitudinal process when the intercept and the slope of the longitudinal process are in an opposite relationship with the time-to-event outcome. Secondly, it is one of few joint models with an AFT regression rather than Cox regression. To calculate the dynamic hazard ratio curve when the proportional hazards assumption is not satisfied, this joint model includes a piecewise linear function in the AFT regression. Finally, this model can estimate the parameters of the longitudinal component by incorporating the time-to-event information through censoring, and similarly, the estimation of the time-to-event accommodates the longitudinal data information.

The proposed joint model is demonstrated with a real clinical transplantation application. The estimation results from our proposed joint model provide at least two useful guidelines for the clinical practice. Firstly, it confirms that the latent baseline and slope of GFR trajectories are significantly related to ACGL. The slope of GFR is negatively correlated with ACGL, which means that patients are more likely to have ACGL when GFR decreases. In addition, patients with a lower baseline GFR are more likely to have ACGL. Secondly, the hazard ratio curve of the effect of pancreas transplant on ACGL helps to understand the risk process of the pancreas transplant for clinical physicians. For example, the hazard ratio is very high in the beginning because of the clinical surgery or acute rejection, decreases to 1 at 152 days post pancreas transplant, and then becomes less than 1. From the time point when the hazard ratio is equal to 1, the pancreas transplant starts to have some survival benefit in comparison with no pancreas transplant. Our proposed joint model can also be applied to other areas, although it is motivated by a clinical data of multiple organ transplantations.

Chapter 4

Jointly Modelling Multiple Outcomes by Functional Principal Component Analysis via a Multistate Model

4.1 Introduction

Multiple studies^{75,76} show that the kidney transplantation can prolong the survival of patients with end-stage renal disease. However, how to extend the long-term survival of the kidney graft still remains the main challenge for transplant despite advancements in pharmaceuticals for the acute rejection, because the kidney graft failure significantly adds to the demand for the limited kidney organ resource. If the kidney graft failure rate can be reduced, kidney recipients can have a long-term survival time. Several surrogate markers have been proposed to predict the kidney graft failure. For example, Marcén et al.⁷⁷ and Moranne et al.⁷⁸ proposed to use the slope of GFR trajectories to predict the graft failure in a Cox model. However, it is not enough to answer how to predict the long-term transplant outcomes.

To well understand to predict the long-term transplant outcomes, we consider three questions. The first question is how to fit the longitudinal trajectory of kidney function progression recorded as repeated GFR measurements. Kidney recipients post transplant are in multi-states: alive, the transplant failure, death after the transplant failure, and death without the kidney failure. The kidney transplant failure is a competing-risk event for death. So the second question is how to estimate the hazard rates of multiple events simultaneously. The third question is to identify the possible markers for the multiple time-to-event outcomes.

To address the first question about how to fit the GFR trajectories, several methods have been developed. The first method is using a parametric model. For example, Marcén et al.⁷⁷, Moranne et al.⁷⁸, and Dong et al.⁹⁶ used a mixed model for the GFR trajectories. The second

method is using a non-parameter model. For example, the functional principal component analysis (FPCA) method is used by Dong et al.⁹⁷ to explore the major sources of variation among GFR trajectories. The dimension of the GFR trajectories is reduced from infinity to four. As shown by Dong et al.⁹⁷, the first four functional principal components (FPCs) can account for 99.8% of variation among GFR trajectories. The GFR trajectories can then be represented with the four FPCs as shown in Figure 4.1, where patient longitudinal GFR trajectory are approaching to a lower level kidney function in the four different patterns which are donated by their first, second, third, and fourth FPC scores respectively.

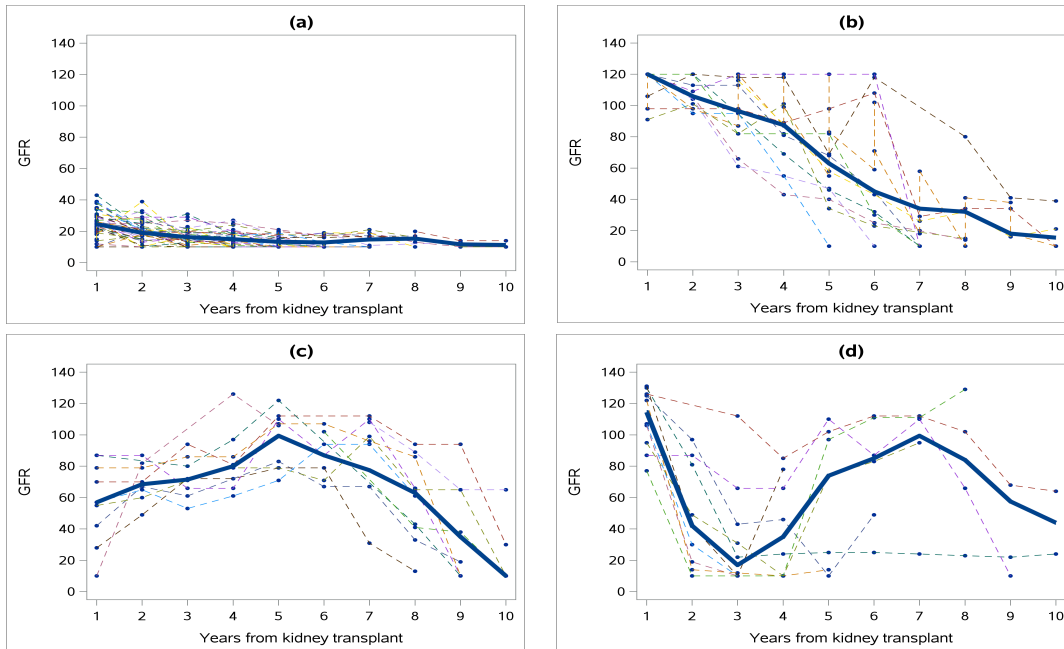


Figure 4.1: GFR curves when their FPC scores are extreme. The thick blue curve in each panel is the average of individual GFR curves in that panel, which represents the common trend in that panel. The four panels are GFR trajectory curves are donated by their first, second, third, and fourth FPC scores respectively.

To address the second question, studies such as Prentice⁸³ and Putter⁸⁴ show that the Kaplan-Meier or Cox method for multiple outcomes may yield unreliable results in the presence of competing risks. The kidney transplant failure is a competing risk for death because the kidney transplant failure increases the probability of death. We model multiple competing risks with a multi-state model in this paper. To address the third question, the longitudinal outcome and the time-to-event outcomes are linked with the shared latent features such as the FPCs after adjusting for other covariates. These latent features can be markers for the multiple time-to-event outcomes. In this paper, we develop a new joint model to address this clinical question about how to predict the long-term transplant outcomes.

Several joint models have been developed for the longitudinal outcome and multiple time-to-event outcomes. However, few joint models are based on a nonparametric approach

for the longitudinal outcome except Yao⁸⁶ and Ding and Wang⁸⁷. Yao⁸⁶ developed a joint model by using FPCA to jointly model the longitudinal outcome and a single time-to-event outcome, where FPCA is used to fit the longitudinal trajectories and then the longitudinal outcome is treated as a covariate in the Cox regression model. Ding and Wang⁸⁷ developed a joint model by treating the longitudinal outcome as a nonparametric multiplicative random effects and jointly modelling a single time-to-event outcomes with a Cox regression model. They mentioned that the first FPC can explain over 71.2% of the total variation in three AIDS studies, the mean functions can mimic the corresponding first FPC, and the longitudinal trajectories of all subjects had different amplitudes but a similar shape. So this paper proposed to treat the longitudinal outcome as a random process proportional to the first FPC.

However, the above two joint models can not be applied to this clinical transplant data directly because these two models can only accommodate a single time-event outcome. Another reason is that we would like to determine the relationships of the dominant variation modes of the GFR trajectories with the multiple time-to-event outcomes. For example, the risk of patients to have the kidney transplant failure or death would be different when their GFR trajectories are flat versus when their GFR trajectories are highly fluctuated. Therefore, we propose a new joint model based on the shared latent features between the longitudinal and survival components, which include the nonparametric approach of functional principal component analysis for the longitudinal outcome and a multi-state submodel for the multiple time-to-event outcomes.

The main contribution of this paper is that we review the clinical question in the kidney transplantation, and tailor a new joint model to address it. To the best of our knowledge, it is the first time to explore the variations of GFR trajectories with multiple time-to-event outcomes based on the latent feature. Our proposed joint models have at least three advantages. Firstly, a multi-status survival model rather than a Cox model can capture the correlation structure of multiple time-to-event outcomes. Secondly, FPCA is an excellent tool for determining the dominant modes of the variations among the longitudinal trajectories, and the estimated dominant modes can be treated as the latent features in the survival model. Lastly, but most importantly, the proposed joint model conditional on the latent features can filter the noises in the longitudinal component in comparison with the method proposed by Yao⁸⁶, in which the longitudinal component is treated as a covariate in the survival model.

The rest of this article is organized as follows. Our proposed joint models are introduced in Section 4.2. We present the estimation method for the proposed joint model in Section 4.3. Section 4.4 demonstrates the application of our joint model in the transplantation clinical data. Section 4.5 presents simulation studies to investigate the finite sample performance of our joint model. Conclusions and discussion are given in Section 4.6.

4.2 A Joint Model

Let $Y_i(t)$ be the longitudinal outcome at the time t for the i -th subject, for example, $Y_i(t)$ is the GFR trajectory in our application. Let T_{mi} be the m -th time-to-event state ($i = 1, \dots, n$, and $m = 1, \dots, M$) for the i -th subject, where M is the number of time-to-event states.

We propose the proposed joint model for the longitudinal outcome $Y_i(t)$ and the multiple time-to-event outcome T_{1i}, \dots, T_{Mi} :

$$\begin{cases} Y_i(t) = \mu(t) + \sum_{k=1}^K \xi_{ik} \phi_k(t) + \boldsymbol{\alpha}^T \mathbf{Z}_i(t) + \epsilon_i, \\ \lambda_{jmi}(t|\boldsymbol{\xi}_i, \mathbf{Z}_i) = \lambda_{jmi}^0 \exp[\overset{T}{j} \boldsymbol{\xi}_i + \overset{T}{j} \mathbf{Z}_i(t)], \end{cases} \quad (4.1)$$

where the first equation is the model for the the longitudinal outcome $Y_i(t)$, $\mu(t)$ is the overall mean of the longitudinal outcome, and $\mathbf{Z}_i(t)$ is a vector of time-dependent/independent covariates. The top K functional principal components (FPCs) $\phi_k(t)$, $k = 1, \dots, K$, explain most variations among the longitudinal outcomes, which will be estimated from the data of the longitudinal outcomes. The FPC scores ξ_{ik} serve as the latent features in the survival model, and they link the longitudinal outcome and the multiple time-to-event outcomes in the joint model. We introduce functional principal component analysis for the longitudinal outcome $Y_i(t)$ with more details in Section 4.2.1.

The second equation is a multi-state survival model. The hazard function $\lambda_{jmi}(t|\boldsymbol{\xi}_i, \mathbf{Z}_i)$ is the hazard rate from the j -th time-to-event state to the m -th time-to-event state. If we have two time-to-event state, and we ignore the transfer from one time-to-event state to the other, it is reduced to the competing-risks survival model. We introduce the multi-state survival model with more details in Section 4.2.2.

As mentioned before, there are several advantages for the proposed new joint model. Firstly the proposed joint models share the latent features from the longitudinal component with the survival sub-model instead of treating the whole longitudinal component as a covariate. The latent features are the FPC scores, which denote the curve pattern. In this way, we can filter the unnecessary noise. Secondly, it is helpful for the clinical practice to determine the relationship between the trajectory patterns of GFR and the time-to-event outcomes by different groups, in which patients have homogeneous GFR trajectory scenarios. As shown in the paper⁹⁷, GFR trajectories can be classified into several homogeneous clusters. Patients in one group who have a very flat GFR trajectory versus patients with a big variation of GFR trajectory, the relationships between the pattern of GFR and the time-to-event outcomes may be different. It is not enough to describe the relationship between GFR and the time-to-event outcomes by using only one parameter γ as shown in the paper by Yao⁸⁶. Thirdly, this clinical dataset has multiple time-to-event outcomes as shown in Figure 4.2, and accordingly multi-state model rather other Cox model is developed. Kidney recipients may have a kidney failure post transplant, some patients might

die before kidney failure, and some patients may die after the kidney failure during the follow-up period. It is useful to determine the proportion of each state. For example, how many patients still have kidney function when death. If the number in this state is large, we waste a lot of kidney organs. We will discuss it in more detail in Section 4.4.

4.2.1 Functional Principal Component Analysis

This section introduce functional principal component analysis. As mentioned in the paper by Dong⁹⁷, functional principal component analysis through the conditional expectation (PACE) is good at determining the dominant modes of the longitudinal data, and the first four leading principal components can account for 99.8% of the longitudinal trajectory variations. So the principal components analysis through the conditional expectation method is used to fit the longitudinal measurement outcome $Y_i(t)$ when it has measurement errors or missing values at some time points for some recipients during the followed-up time.

$$Y_i(t) = X_i(t) + \epsilon_i = \mu(t) + \sum_{k=1}^K \xi_{ik} \phi_k(t) + \boldsymbol{\alpha}^T \mathbf{Z}_i(t) + \epsilon_i, \quad (4.2)$$

where $i = 1, \dots, n$, ϵ_i are identically and independently distributed normal random variables with mean 0 and variance σ^2 . The function $\phi_k(t)$ is the k -th functional principal component, and $\xi_{ik} = \int_{\mathcal{T}} (X_i(t) - \mu(t)) \phi_k(t) dt$ is the associated functional principal component score, where \mathcal{T} is the bounded time-frame range. Then the variance-covariance function $G(s, t)$ can be expressed as:

$$G(s, t) = \text{Cov}(X_i(s) - \mu(s), X_i(t) - \mu(t)) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t),$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$. If the observations of GFR are sparse, the mean function is obtained by smoothing the data from all observations based on the local linear smoother method by Fan⁸². If the mean function and the eigenfunctions are assumed to be smooth, then the expansions of a set of smooth basis functions such as B-splines or regression splines can be used to model the overall mean function and the eigenfunctions as shown in the papers James⁸¹ and Yao⁸⁶.

4.2.2 Multi-state models

We want to develop multi-state models for multiple time-to event outcomes. Patients move among a number of discrete states as shown in Figure 4.2. For example, kidney recipients may move to the transplant failure first, and then move to death from the transplant failure; or some patients directly move to death without the transplant failure.

Several multi-state survival models have been developed. We are focus on the competing risks models and the progressive illness-death models. In the competing risks framework, two

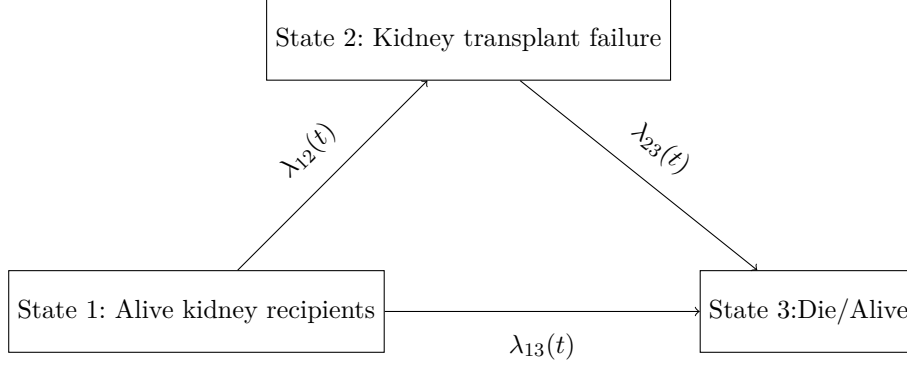


Figure 4.2: The three states of kidney transplant recipients. All patients start from the date of the kidney transplant (state 1), then they may move to state 2 (kidney failure). If not, they directly move to state 3 when die

popular competing risks models are used. One is the cause-specific hazard model proposed by Prentice⁸³ and Putter⁸⁴, and the other is the sub-distribution hazards regression introduced by Fine and Gray⁸⁸. The cause-specific hazard model calculates the occurrence rate of specific event types in subjects who are currently event free. For example, there are 2 types of events in this application example: death with the kidney function from other reasons and death from the kidney transplant failure. The cause-specific hazard of the kidney failure death denotes the instantaneous rate of the kidney failure death in alive subjects who have not yet experienced either event. The sub-distribution hazard model calculates the instantaneous risks of the specific event type in subjects who have not yet experienced this event type. If in the progressive illness-death model framework, then the progressive illness-death model can determine the incidence of kidney transplant failure, the mortality rate for alive patients after kidney transplant, and mortality rate for patients with kidney transplant failure.

If the transition intensities of multi-state models can be specified as $\lambda_{jm}(t) = \lim_{\Delta t \rightarrow 0} \frac{P_{jm}(t, t + \Delta t)}{\Delta t}$, $j \neq m$, and $\lambda_{mm} = - \sum_{j \neq m} \lambda_{jm}(t)$, then the transition intensities can be specified in a matrix. For the convenient notation by setting $M = 3$, the matrix of transition intensities area can be specified as follows:

$$Q(t) = \begin{bmatrix} -(\lambda_{12}(t) + \lambda_{13}(t)) & \lambda_{12}(t) & \lambda_{13}(t) \\ 0 & -\lambda_{23}(t) & \lambda_{23}(t) \\ 0 & 0 & 0 \end{bmatrix}$$

where

$$\lambda_{12} = \lim_{\Delta t \rightarrow 0} \frac{P_{12}(t, t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(\text{state 2 at time } t + \Delta t | \text{state 1 at time } t)}{\Delta t},$$

$$\lambda_{13} = \lim_{\Delta t \rightarrow 0} \frac{P_{13}(t, t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(\text{state 3 at time } t + \Delta t | \text{state 1 at time } t)}{\Delta t},$$

$$\lambda_{23} = \lim_{\Delta t \rightarrow 0} \frac{P_{23}(t, t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(\text{state 3 at time } t + \Delta t | \text{state 2 at time } t)}{\Delta t}.$$

If the probability distribution on the state space of a Markov chain is discrete and the Markov chain is homogeneous, then the Chapman-Kolmogorov equations can be expressed in terms of matrix multiplication:

$$P(s, t) = P(s, u)P(u, t), \quad s < u < t$$

The transition probability $P(s, t)$ is the unique solution of the Kolmogorov forward differential equation:

$$\frac{\partial}{\partial t} P(s, t) = P(s, t)Q(t); \quad P(s, s) = I$$

$P(s, t)$ can be recovered from the transition intensities through product integration

$$P_{11}(s; t) = \exp\left(-\int_s^t (\lambda_{12}(u) + \lambda_{13}(u)) du\right)$$

$$P_{22}(s; t) = \exp\left(-\int_s^t \lambda_{23}(u) du\right)$$

$$P_{23}(s, t) = 1 - P_{22}(s, t)$$

$$P_{12}(s, t) = \int_s^t P_{11}(s, u) \lambda_{12}(u) P_{22}(u, t) du$$

$$P_{13}(s, t) = 1 - P_{11}(s, t) + P_{12}(s, t)$$

$$P_{jm}(s, t) = 0, \quad \text{when } j > m$$

The interpretations of the above transition intensities and probability in the matrix are identical to those in the competing-risks model or the progressive illness-death models. For example, in the competing risks framework, state 1 is being alive after treatment such as kidney transplant, state 2 and state 3 are distinct patient status such as kidney failure and die. In the progressive illness-death models, state 1 is alive, state 2 is that patients have an illness, and state 3 are when patients die without/with an illness.

4.3 The Estimation Method

4.3.1 The joint likelihood functions

In this section, we want to give the inference of the proposed joint model. Let $(t_i, m_i, \delta_{m_i}, \mathbf{Z}_i(t), Y_i(t))$ donate the observations of each subject in the data, where t_i is the observed survival time, m_i is the observed event type, δ_{m_i} is the failure indicator of any event type, $\mathbf{Z}_i(t)$ is observed co-variables, and $Y_i(t)$ are longitudinal outcomes. Let C_i be

a potential censoring time, and T_i be the largest time of all event types. We assume that $S_i = \min(T_i, C_i)$, and $\delta_{m_i} = \mathbb{1}_{T_{m_i} \leq C_i}$. The parameters $\Theta = (\gamma, \beta, \lambda_0, \Lambda, \sigma^2)$ need to be estimated from data.

In order to estimate the parameters, we need to construct the joint likelihood functions. The longitudinal trajectories of $Y_i(t)$ can be determined by the FPC score $\xi_i = (\xi_{i1}, \dots, \xi_{iK})^T$ as shown in Section 4.2.1, so the joint probability density function of $Y_i(t)$ and the time-to-event T_i can be written as the factorization of the density distribution of the FPC score ξ_i and the conditional survival density distribution of T_i on the latent FPC score ξ_i . In other words, the full likelihood of the full set of parameters under independent censoring can be given by:

$$L(\Theta) = \prod_{i=1}^n \prod_{m=1}^M \left\{ \int f(T_i, \delta_{m_i} | \xi_i, \mathbf{Z}_i(t), \gamma, \beta) f(Y_i(t) | X_i(t), \alpha, \sigma) f(\xi_i | \Lambda) d\xi_i \right\}, \quad (4.3)$$

where the density survival function $f(T_i, \delta_{m_i} | \xi_i, \mathbf{Z}_i(t), \gamma, \beta)$ is given by $f(T_i, \delta_{m_i} | \xi_i, \mathbf{Z}_i(t), \gamma, \beta) = \lambda_m(t_i | \xi_i, \mathbf{Z}_i(t_i), \gamma, \beta)^{\delta_{m_i}} S(t_i | \xi_i, \mathbf{Z}_i(t_i), \gamma, \beta)^{1 - \delta_{m_i}}$, $f(Y_i(t) | X_i(t), \alpha, \sigma) = (2\pi\sigma^2)^{-\frac{n_i}{2}} \exp\{-\frac{1}{2\sigma^2}(Y_i - X_i)^T(Y_i - X_i)\}$, and $f(\xi_i | \Lambda) = (2\pi|\Lambda|)^{-\frac{1}{2}} \exp(-\frac{1}{2}\xi_i^T \Lambda^{-1} \xi_i)$.

Given the sub-hazard density function for the sub-survival model as in the following

$$f(T_{m_i}, \delta_{m_i} | X_i(t), \mathbf{Z}_i(t), \gamma, \beta) = \frac{\exp(\gamma_m^T \xi_i + \beta_m^T \mathbf{Z}_i(t))}{\sum_{i=1}^n R_i(t) \exp(\gamma_{m_i}^T \xi_i + \beta_{m_i}^T \mathbf{Z}_i(t))},$$

where $R_i(t) = \mathbb{1}_{[m_i: (C_i \wedge T_{m_i} > T_i) \cup \{(T_{m_i} \leq T_i) \cap (\delta_{m_i} = 0) \cap (C_i \leq T_i)\}]}$, then the likelihood function in the equation (4.3) becomes as following:

$$L(\Theta) = \prod_{i=1}^n \int \prod_{m=1}^M \left[\frac{\exp(\gamma_m^T \xi_i + \beta_m^T \mathbf{Z}_i(t))}{\sum_{i=1}^n R_i(t) \exp(\gamma_{m_i}^T \xi_i + \beta_{m_i}^T \mathbf{Z}_i(t))} \right] f(Y_i(t) | X_i(t), \sigma) f(\xi_i | \Lambda) d\xi_i.$$

According to equation (4.3), the score function is found to be proportional to

$$\begin{aligned} S(\Theta) &\approx \frac{\partial(\sum_{i=1}^n \log\left\{ \prod_{m=1}^M \left\{ \int f(T_i, \delta_{m_i} | \xi_i, \mathbf{Z}_i(t), \gamma, \beta) f(Y_i(t) | X_i(t), \alpha, \sigma) f(\xi_i | \Lambda) d\xi_i \right\} \right\})}{\partial \Theta} \\ &= \sum_{i=1}^n \int \frac{\partial h(\Theta, \xi_i)}{\partial \Theta} f(\xi_i | T_i, \delta_i, Y_i(t), \Theta) d\xi_i \end{aligned} \quad (4.4)$$

where

$$h(\Theta, \xi_i) = \log\left\{ \prod_{m=1}^M f(T_i, \delta_{m_i} | \xi_i, \mathbf{Z}_i(t), \gamma, \beta) f(Y_i(t) | X_i(t), \alpha, \sigma) f(\xi_i | \Lambda) \right\}$$

The observed data score vector in the formula 4.4 is expressed as the expected value of the complete-data score vector with respect to the posterior distribution of the random effects of $\boldsymbol{\xi}$. If the score equations in the formula 4.4 can be solved with respect to Θ , with $p(\boldsymbol{\xi}_i|T_i, \delta_i, Y_i(t), \Theta)$ fixed at the Θ value of the previous iteration, then it is an EM algorithm. However, there are some challenges to estimate parameters in the by the EM algorithm. Therefore, we propose to use a modified two-stage algorithm to estimate the parameters of the proposed joint models.

4.3.2 Parameter estimation

The fully joint log-likelihood function has been given in Section 4.3.1. This section is focus on estimating the parameters. There are the two main challenges to estimate parameters in the joint likelihood functions. One is the requirement for numerical integration of latent variables $\boldsymbol{\xi}_i$ when the dimension of random effects increases. The other is to estimate the density function $f(\boldsymbol{\xi}_i|\Lambda)$ because we don't have a closed-form for FPC function $\phi_k(t)$. Therefore, we propose to use the new two-stage approach, but it is different from the previous two-stage method for joint model in the papers such as Tsiatis⁴⁸. The proposed two-stage algorithm is specified as follows:

- **Stage I**

- Step 1: We estimate all the parameters from the longitudinal process by Functional Principal Component Analysis,
- Step 2: After estimating the mean curve $\hat{\mu}$, the FPC $\hat{\phi}_k$, and the FPC score $\hat{\xi}_{ik}$, and $\hat{\sigma}^2$ from all available GFR data, we can recover any missing or predict future value using the following formula. The vector of random effects $\hat{\xi}_{ik}$ is shared between both longitudinal and survival sub-models. Therefore, we try to reduce the biases from the informative dropout problem for estimating random effects parameters, the missing measurements of the observed longitudinal data are generated for all subjects as in the following:

$$\hat{X}_i(t)|(T_i, \hat{\mu}, \hat{\phi}_k, \hat{\xi}_{ik}, \hat{\sigma}^2) = \hat{\mu}(t) + \sum_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t) + \hat{\boldsymbol{\alpha}}^T \mathbf{Z}_i(t) + \epsilon_i, \quad (4.5)$$

where t can be any past or future time points before patient death. In other words, we have simulated complete longitudinal measurements $Y_i(t)$ in the step 2.

- Step 3: Estimate parameters $\hat{\boldsymbol{\xi}}$, $\hat{\mu}(t)$, $\hat{\boldsymbol{\alpha}}^T$, and $\hat{\sigma}^2$ using complete longitudinal measurements simulated in Step 2. As $\min(n_i) \rightarrow \infty$, the estimated parameters in the submodel will convergence to the estimated parameters obtained from the joint model in probability as shown in the papers by Rizopoulos⁵⁶ and Thu⁵⁸.

- **Stage II**

The proposed joint models become the following by using the fitted values from stage I

$$\begin{cases} \hat{Y}_i(t) = \mu(\hat{t}) + \sum_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t) + \hat{\boldsymbol{\alpha}}^T \mathbf{Z}_i(t) + \epsilon_i, \\ \lambda_{jmi}(t|\hat{\boldsymbol{\xi}}_i, \mathbf{Z}_i) = \lambda_{jmi}^0 \exp[\gamma_{jm}^T \hat{\boldsymbol{\xi}}_i + \beta_{jm}^T \mathbf{Z}_i(t)], \end{cases} \quad (4.6)$$

- Step 1: Now we can approximate the expected function of the complete data likelihood. Instead of using the partial likelihood because of latent variables to estimate the regression coefficients in the joint model.

$$\begin{aligned} L(\Theta) = & \prod_{i=1}^n \int \prod_{m=1}^M \left(\{\lambda_m[t_i|\boldsymbol{\xi}_i, \mathbf{Z}_i(t_i)]\}^{\delta_{m_i}} \exp\{-\int_0^\infty \sum_{i=1}^n R_i(u) \right. \\ & \left. \lambda_m(u|\boldsymbol{\xi}_i, \mathbf{Z}_i(u)) du\} \right) f(Y_i(t)|\boldsymbol{\xi}_i, \hat{\boldsymbol{\alpha}}, \hat{\sigma}) f(\hat{\boldsymbol{\xi}}_i|\hat{\Lambda}) d\boldsymbol{\xi}_i \end{aligned} \quad (4.7)$$

As shown in the paper by Rizopoulos⁵⁶ and the paper by Thu⁵⁸, the expected function of the complete data log-likelihood function can be approximated by the following as $\min(n_i) \rightarrow \infty$:

$$\begin{aligned} E(l(\Theta)) \approx & \sum_{i=1}^n \log\left\{ \prod_{m=1}^M \left(\{\lambda_m[t_i|\hat{\boldsymbol{\xi}}_i, \mathbf{Z}_i(t_i)]\}^{\delta_{m_i}} \exp\{-\int_0^\infty \sum_{i=1}^n R_i(u) \right. \right. \\ & \left. \left. \lambda_m(u|\hat{\boldsymbol{\xi}}_i, \mathbf{Z}_i(u)) du\} \right) f(Y_i(t)|X_i(t), \hat{\boldsymbol{\alpha}}, \hat{\sigma}) f(\hat{\boldsymbol{\xi}}_i|\hat{\Lambda}) \right\} \end{aligned} \quad (4.8)$$

- Step 2: we can estimate all the parameters for survival submodel by maximizing the approximation of the expected function of the complete data log-likelihood as in the formula (4.8).

4.4 The application of the proposed joint model

Total 5654 kidney transplant recipients are included in this study from United Network for Organ Sharing (UNOS), and patient demographics are shown in Table 4.1. According to the clinical research question, we propose the proposed joint model in the following:

$$\begin{cases} Y_i(t) = \mu(t) + \sum_{k=1}^K \xi_{ik} \phi_k(t) + \boldsymbol{\alpha}^T \mathbf{Z}_i(t) + \epsilon_i, i = 1, \dots, n, \\ \lambda_{mi}(t|\boldsymbol{\xi}_i, \mathbf{Z}_i) = \lambda_m^0 \exp[\gamma_m^T \boldsymbol{\xi}_i + \beta_m^T \mathbf{Z}_i(t)], \end{cases} \quad (4.9)$$

where $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iK})^T$ and $\mathbf{Z}_i(t)$ be other independent covariates as in Table 4.1.

Table 4.1: Kidney transplanted recipient characteristics in some kidney transplant data

Characteristics	Percentage (%)
Age	
18-39	48
40-59	49
≥ 60	13
Sex	
Female	41
Male	59
Race	
White	69
Black	31
Other	8
Cause of End-Stage Renal Disease(ESRD)	
Diabetes	32
Hypertension	21
Glomerular Disease	29
Polycystic disease	9
Others	19
Kidney donor type	
Living	22
Deceased	78

4.4.1 Results from Functional Principal Component Analysis

From functional principal components analysis, the four leading functional principal components ($K = 4$) account for 99.8% of the total variability of GFR curves. The first FPC component, second, third, and fourth represents 84.6%, 10.6%, 3.4%, and 1.1% of the total variability of GFR curves comes from the overall weighted mean of GFR curves respectively. Although the first two FPCs account for 95.24% of the total variability, the third and fourth also contain some important patient information as mentioned in the our previous paper⁹⁷. In practice, these eigenfunctions have some straightforward explanation. For example, the first eigenfunction is very flat during the followed-time, which indicates that the largest GFR variation between subjects is from the specific mean curve of GFR curves. In other words, the mean GFR curve captures the largest variation of the data, and most of patients have a stable kidney function trajectory. The third and fourth FPCs represent a small number of patients who have strong fluctuating curves. The third and fourth FPCs can easily identify those abnormal patients with high fluctuate GFR curves, which should be caught more attention by physicians. We will discuss the relationship of the first four FPC scores with time-to event outcomes again in the following Section 4.4.2.

4.4.2 Results from multi-state submodel

This section present the result from the survival sub-model in the proposed model 4.9. Among total 5654 patients, 1,590(28%) patients have a kidney transplant failure and 1,735(31%) patients die. 707(44%) patients die after kidney failure, and 1,028(28%) patients die with kidney function. We use the Akaike Information Criterion (AIC) to select our final model such as the number of principal components, since AIC can consider the joint likelihood of longitudinal and survival models.

The hazard ratios of kidney transplant failure from the joint models are shown in Table 4.2, and the hazard ratios of death from the proposed joint model are shown in Table 4.3. All of the first four FPC scores are statistically significant. The hazard ratios of the first four FPC scores are different. For example, it is a negative relationship between the first FPC score and the time-to event outcome while it is a positive relationship for the second FPC score with the primary time-to event outcome. In fact, we can give a clinical explanation. For example, for the first FPC score, the failure rate of primary increases as patients have a lower level GFR during the followed-up time period, and patients in the higher level of GFR are less likely to have the time-to-event outcome. Similarly the third and fourth FPC scores are also significantly related to time-to event outcome. In other words, these patients with abnormal trajectories should be monitored more closely, and they need to be diagnosed to find out the underlying reason in clinical practice. It is worth mentioning that it isn't enough from our model results if the change slopes of GFR are assumed to be linear. If a

large proportion of trajectories are nonlinear, this simplified linear assumption may cause the long-term time-to event model result to be biased.

More importantly, it is very interesting to see that young patients are more likely to have a kidney transplant failure as in Table 4.2. In other words, the graft life is shorter than patient life. However, old patients are more likely to have a death. In other words, the patients life is shorter than the kidney graft life for these old patients, which indicates that we waste kidney organs when we transplant young better donor to old patients. We should give young donor to young patients so that we can make most use of the scarce organ resources. It is clinically reasonable for the relationship of time-to-event outcomes with other co-variables. Female patients are less likely to have an event compared with male patients. Compared with the white patients, the black patients are more likely to have the time-to-event outcome. Compared with patients who have a deceased donor, patients who have a living donor transplant are less likely to have the time-to-event outcome.

On the other hand, Table 4.2 also displays the estimation results for the Cox model. The Cox model concludes that senior patients are more likely to have a kidney transplant failure, which is contrast to the results from the competing-risks models. In fact, Cox regression have identified factors that affect survival of renal recipients, but these standard models only focus on identifying factors affecting the time of occurrence of the targeted outcome while ignore events that occur for patients during the study which may affect the interest event. In a short, we find that the use of multi-state models are recommended in this paper because the Kaplan-Meier or Cox method for multi-state outcomes may yield unreliable results in the presence of multiple events outcome.

4.5 Simulations

In order to study the difference from the different correlation construction between the longitudinal sub-model and the survival model, this section generate several simulation datasets. our proposed joint models (4.1) construct the correlation of the longitudinal outcome and the time-to-event outcomes through the latent features, which is different from the joint model⁸⁶ that treats the longitudinal outcome as a covariable in the survival model as shown in the model (4.10), we choose several different scenarios to see the differences between the proposed model (4.1) and the model (4.10).

$$\begin{cases} Y_i(t) = \mu(t) + \sum_{k=1}^K \xi_{ik}\phi_k(t) + \epsilon_i, i = 1, \dots, n, \\ \lambda(t|X(t)) = \lambda_0 \exp(\gamma_1 X(t)). \end{cases} \quad (4.10)$$

Similarly, a non-parameter FPCA model is chosen to simulate the longitudinal trajectories:

$$Y_i(t) = X_i(t) + \epsilon_i = \mu(t) + \sum_{k=1}^K \xi_{ik}\phi_k(t) + \epsilon_i,$$

Table 4.2: Estimated hazard ratios of kidney failure post kidney transplant in the joint model with different survival sub-models. 95% confidence interval are given in brackets.

	Joint models			
	Multistate submodel		Cox submodel	
	hazard ratios	p value	hazard ratios	p value
Age				
18-39	1.00		1.00	
40-59	0.68 (0.61,0.76)	< 0.001	1.61(1.35, 1.92)	0.001
≥ 60	0.47 (0.38,0.57)	< 0.001	1.49(1.15, 1.93)	0.001
Sex				
Male	1.00		1.00	
Female	0.89(0.80,0.99)	0.048	0.78(0.66, 0.91)	0.038
Race				
White	1.00		1.00	
Black	1.39(1.23, 1.55)	< 0.001	1.37(1.16, 1.62)	< 0.001
Others	0.82(0.66, 1.02)	0.079	0.63(0.44, 0.89)	< 0.001
Cause of ESRD				
Diabetes	1.00		1.00	
Hypertension	0.89(0.74, 0.97)	0.026	0.74(0.61, 0.91)	< 0.001
Glomerular disease	0.99(0.85, 1.14)	0.834	0.55(0.44, 0.68)	< 0.001
Polycystic disease	0.76(0.60, 0.97)	0.026	0.44(0.31, 0.62)	< 0.001
Others	0.90(0.76, 1.07)	0.221	0.56(0.44, 0.71)	< 0.001
Donor type				
Decreased donor	1.00		1.00	
Living donor	0.90(0.79,0.99)	0.048	0.84(0.68,1.02)	0.084
FPC Scores				
First FPC score	0.981(0.979, 0.982)	< 0.001	0.968(0.965, 0.970)	< 0.001
Second FPC score	1.009(1.005, 1.013)	< 0.001	1.000(0.993, 1.005)	< 0.001
Third FPC score	0.976(0.969, 0.983)	< 0.001	0.964(0.953, 0.975)	< 0.001
Fourth FPC score	0.993(0.974, 1.000)	0.050	0.972(0.946, 0.999)	0.048

Table 4.3: Estimated hazard ratios of death post kidney transplant from different survival sub-models and 95% confidence interval are given in brackets.

	Joint models			
	Multistate submodel		Cox submodel	
	hazard ratios	p value	hazard ratios	p value
Age				
18-39	1.00		1.00	
40-59	1.91(1.69,2.17)	< 0.001	2.51(2.20,2.84)	< 0.001
≥ 60	3.37(2.91,3.90)	< 0.001	4.79(4.13,5.58)	< 0.001
Sex				
Male	1.00		1.00	
Female	0.82(0.74,0.90)	0.001	0.84(0.76,0.93)	0.007
Race				
White	1.00		1.00	
Black	0.97(0.87,1.08)	0.608	0.93(0.83,1.03)	0.170
Other	0.59(0.48,0.73)	< 0.001	0.63(0.51,0.78)	< 0.001
Cause of ESRD				
Diabetes	1.00		1.00	
Hypertension	0.68(0.60,0.77)	< 0.001	0.58(0.51,0.66)	< 0.001
Glomerular disease	0.56(0.49,0.63)	< 0.001	0.46(0.41,0.53)	< 0.001
Polycystic disease	0.43(0.35,0.51)	< 0.001	0.37(0.30,0.45)	< 0.001
Others	0.63(0.54,0.73)	< 0.001	0.55(0.47,0.63)	< 0.001
Donor type				
Decreased donor	1.00		1.00	
Living donor	0.79(0.70,0.89)	0.001	0.74(0.65,0.84)	< 0.001
FPC Scores				
First FPC score	0.992(0.991,0.993)	< 0.001	0.992(0.991,0.993)	< 0.001
Second FPC score	0.991(0.988,0.994)	< 0.001	0.997(0.994,1.000)	0.049
Third FPC score	0.961(0.956,0.966)	< 0.001	0.980(0.974,0.986)	< 0.001
Fourth FPC score	1.008(1.000,1.023)	0.045	0.993(0.977,1.009)	0.376

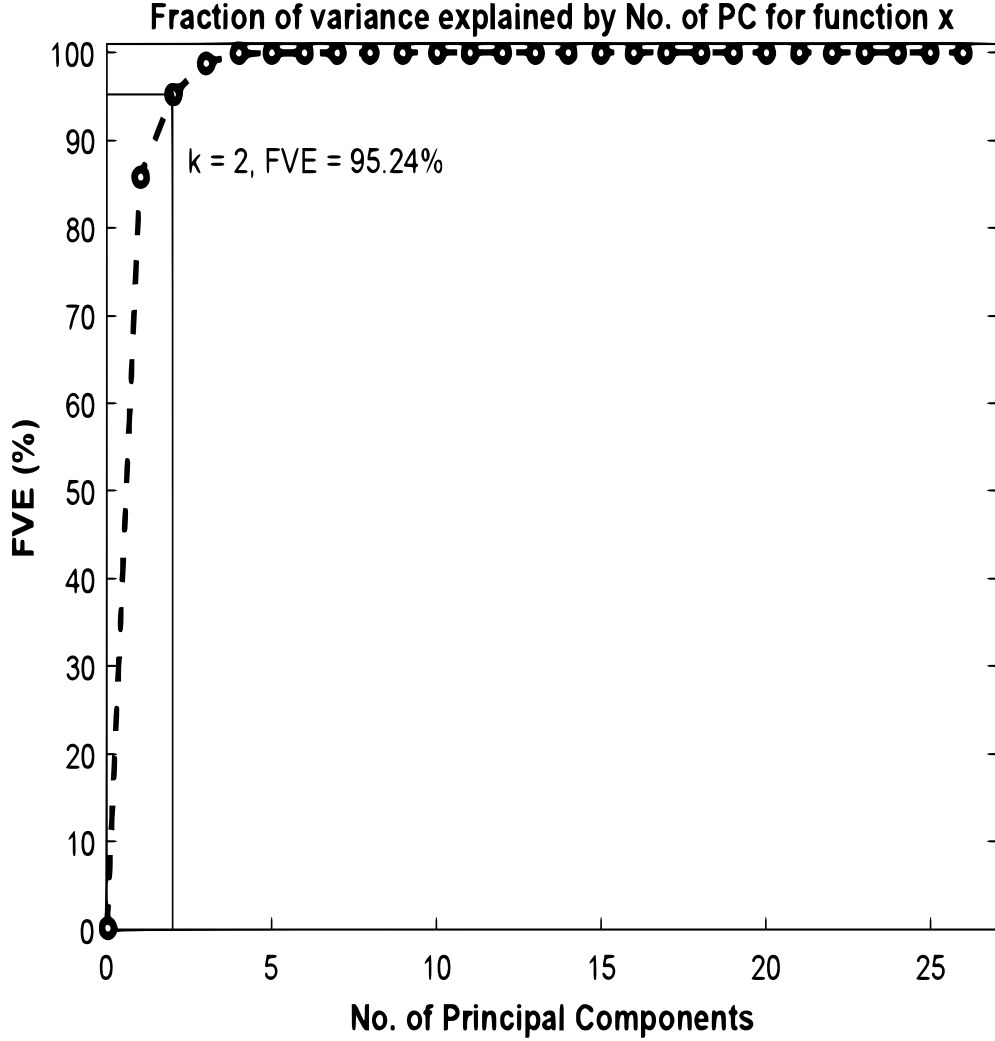


Figure 4.3: The first two leading functional principal components (FPCs) account for 95.24% of the total variability of GFR curves, and the four leading FPCs account for 99.82%

where $\mu(t)$ is the true mean value in the application example, which are the estimate from the real data. The measurement error $\epsilon_i \sim \text{Normal}(0, 0.85)$. The scheduled measurement times of the repeated longitudinal outcome are set at the sequence year $(1, 2, \dots, T_i)$ for each subject, but there are no measurements available after death or censoring time. The time-to-event T_i is specified as in the following:

$$\lambda_m(t|\boldsymbol{\xi}_i) = \lambda_m^0 \exp(\gamma_{m1}\xi_{i1} + \gamma_{m2}\xi_{i2} + \gamma_{m3}\xi_{i3} + \gamma_{m4}\xi_{i4}).$$

Three data cohorts in different scenarios are generated. Each cohort has a maximum follow-up time of 3650 days, the time to the competing-risk event is assumed to follow a Weibull distribution, and the time to the primary event outcome is assumed to follow a log-normal distribution. We assume that the FPC scores $\xi_{ik} \sim \text{Normal}(0, \sigma_k^2)$, where $k = 1, \dots, 4$, and

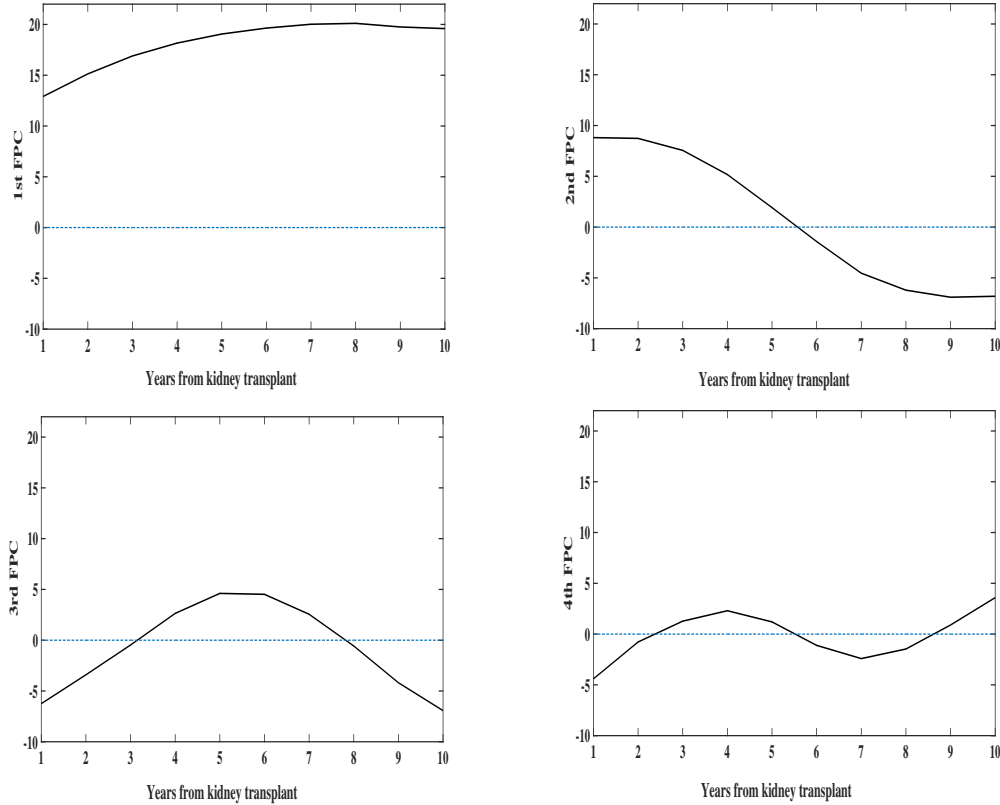


Figure 4.4: The first four leading functional principal components (FPCs) estimated from the GFR curves.

$\sigma_1 = 16$, $\sigma_2 = 8$, $\sigma_3 = 4$, $\sigma_4 = 1$. Our survival model choose the first four FPC scores as covariates. The corresponding γ are designed in the following three different scenarios. In the first scenario, we choose $\gamma_{11} = \gamma_{12} = \gamma_{13} = \gamma_{14} = 1.00$ for the primary event while ξ_{ik} have no effect on the competing-risk event when we set $\gamma_{21} = \gamma_{22} = \gamma_{23} = \gamma_{24} = 0.00$ for the competing-risk event. The simulation result is shown in Table 4.4. In the second scenario, we choose different coefficients $\gamma_{11} = -1.00$, $\gamma_{12} = 1.00$, $\gamma_{13} = -1.00$, and $\gamma_{14} = 1.00$ for ξ_{ik} on the primary event, but we set ξ_{ik} to have same effect on the competing-risk event by choosing $\gamma_{21} = \gamma_{22} = \gamma_{23} = \gamma_{24} = 0.00$. The simulation result is shown in Table 4.4. In the third scenario, we choose different coefficients $\gamma_{11} = -1.00$, $\gamma_{12} = 0.85$, $\gamma_{13} = -0.75$, $\gamma_{14} = 0.50$ for the primary event, and set $\gamma_{21} = 0.50$, $\gamma_{22} = -0.50$, $\gamma_{23} = 0.50$, $\gamma_{24} = -0.50$ for the competing-risk event. The simulation result is shown in Table 4.4

Table 4.4 displays the estimates, together with their estimated standard errors for the first Scenario, where the coefficients ($\gamma_{11} = \gamma_{12} = \gamma_{13} = \gamma_{14} = 1.00$) are same and there is no-competing-risk effect, the estimated coefficient $\hat{\gamma}_1$ for the model in (4.10) has the same as the true value γ_1 . In fact, our proposed joint model become the model (4.10) in this scenario. In other words, the model by (4.10) is a special case of our proposed joint model. In Scenario 2 when the different coefficients γ are different, the coefficient γ_1 from the model

Table 4.4: Means and standard deviations (STD) in three different scenarios. Each scenario has 100 simulation replicates and 100 subjects in each simulation replicate

Scenario 1				
Parameters	γ_1	γ_2	γ_3	γ_4
True value	1.00	1.00	1.00	1.00
Fitted value in Model ^(4.1)	1.01(0.02)	0.96(0.01)	1.02(0.03)	0.97(0.01)
Fitted value in Model ^(4.10)	0.99(0.03)	-	-	-
Scenario 2				
Parameters	γ_1	γ_2	γ_3	γ_4
True value	-1.00	1.00	-1.00	1.00
Fitted value in Model ^(4.1)	-1.02(0.03)	0.96(0.01)	-1.02(0.03)	0.97(0.01)
Fitted value in Model ^(4.10)	-0.96(0.03)	-	-	-
Scenario 3				
Parameters	γ_1	γ_2	γ_3	γ_4
True value	-1.00	0.85	-0.75	0.50
Fitted value in Model ^(4.1)	-0.98(0.03)	0.84(0.01)	-0.72(0.02)	0.47(0.01)
Fitted value in Model ^(4.10)	-0.36(0.03)	-	-	-

(4.10) is completely different from the true value. In summary, the results from Table 4.4 demonstrate that the model (4.10) cannot describe the first four principal components of the longitudinal outcome with the time-to-event outcome by a single parameter γ_1 . Especially in Scenario 3 when there also exists a competing-risk event, it is impossible to describe the two relationships by a single parameter γ_1 . This simulation example shows the advantages in our proposed joint model by using the features from the longitudinal submodel in the survival submodel.

4.6 Conclusions and discussion

This paper is motivated by a longitudinal and time-to-event transplant clinical data. The proposed joint models include a longitudinal FPCA submodel and a multi-state submodel, and both of submodels share some latent variables together. The multi-state survival submodel can calculate the hazard ratios of multiple time-to event outcomes. We have demonstrated the applicability of the proposed joint model to some real transplantation clinical data. The main result from the application data of transplant can answer the clinical transplant question. We find that Cox model may cause bias when exist multiple outcome in the clinical data. For example, different hazard ratios of age categories between multistage model and Cox model in Tables 4.2 and 4.3. The finite sample performance of the proposed method is verified in the simulation study, and some advantages of the proposed model when compared with the joint model (4.10) are shown in the simulation study.

The application results from our proposed joint model can supply some useful references for the clinical practice as in Section 4.4. There are at least two important points. Firstly,

it confirms that the four FPC are significantly related to the event outcome. Also these eigenfunctions have some straightforward explanation. For example, the first eigenfunction is very flat during the followed-time, which indicates that the largest GFR variation between subjects is in the subject specific mean curve of GFR value. In other words, the mean curve captures the largest variation of the data, and the baseline GFR is significantly related to time-to-event outcome.

More importantly, we find that young patients are more likely to have a kidney transplant failure as in Table 4.2 from multistage model, which indicates that the graft life is shorter than patient life. However, old patients are more likely to have a death in Table 4.3, which indicates the patients life is shorter than the kidney graft life for these old patients, which indicates that we may waste kidney organs when we transplant young better donor to old patients. We should give young donor to young patients so that we can make most use of the scarce organ resources. However, these results can't be realized by Cox model.

Chapter 5

Jointly Modelling Multiple Continuous and Discrete Outcomes by a Flexible Class of Generalized Linear Latent Variable Models

5.1 Introduction

We have introduced several joint models for a continuous longitudinal outcome and single/multiple time-to-event outcomes in different scenario in Chapter 3 and 4. However, in many clinical studies, many clustered data have mixed outcomes during the longitudinal followed-up period. For example, how to fit multiple outcomes of transplant recipients in the application example. The continuous longitudinal outcome is the kidney function after kidney transplant recorded as the estimated glomerular filtration rate (eGFR), and the discrete outcome is the repeated kidney transplant status. Most of studies only consider the time frame from transplant to kidney failure by ignoring all available information after kidney failure. It may not complete because it ignores all useful available information after kidney failure such as all eGFR after kidney failure and the kidney retransplant status. Therefore, it is motivated by this clustered multivariate mixed outcomes, and so this chapter want to develop a new joint model with mixed outcomes by considering all available information.

Joint models do have several advantageous compared with separate analysis as mentioned before. For example, jointly modeling these mixed outcomes together can allow these questions to be answered directly since analyses of different outcomes separately do not address directly the questions of interest. More importantly, joint modelling can avoid multiple testing and then lead to global tests so that it result in increased power and better control of Type I error rates⁴⁶ and⁴⁷. According to the multivariate clustered data, the proposed joint models need to account for three level correlations: the correlation among different outcomes, the correlation among repeated continuous measures of the same outcome over

time, and the correlation among repeated discrete outcome, and the third association is between the continuous outcome and the discrete outcome.

A number of joint modeling strategies for mixed outcomes have been studied in the literature such as the papers^{92, 93}, and⁹⁴. The general approach first specifies a model for the joint distribution of mixed outcomes, then fits the model to the current data. After fit the model, an inference for the proposed joint model is made. However, the difficulties in joint modeling for continuous and discrete outcomes are the lack of a natural joint multivariate distribution. This chapter discuss a flexible class method, which is based on the generalized linear latent variables toward a joint model for a continuous and a discrete outcomes. We apply the proposed joint model to some kidney transplant data in the application example.

The rest of this chapter is organized as follows. The proposed joint model for mixed outcomes is formulated in section 5.2. The covariance structure of the latent variables is in section 5.3. The MCEM algorithm steps is in section 5.4. In section 5.5, we analyze the real data set from the kidney study to illustrate the proposed method. We conclude with a discussion in section 5.6.

5.2 Model Specification

Let y_{kij} be the k^{th} response variable of the i^{th} subject at time j , and \mathbf{x}_{kij} be the vector of covariates associated with the response y_{kij} , where $k = 1, 2, \dots, K$, $i = 1, \dots, N$, and $j = 1, \dots, n_i$. For example, two mixed outcomes ($\mathbf{y}_1 = (\mathbf{y}_{11}, \dots, \mathbf{y}_{1N})$ and $\mathbf{y}_2 = (\mathbf{y}_{21}, \dots, \mathbf{y}_{2N})$) denote the sequence of mixed outcomes of continuous and discrete responses from N subjects when $K = 2$. The response outcome $\mathbf{y}_{1i} = (y_{1i1}, y_{1i2}, \dots, y_{1in_i})$ are repeated measurement GFR, and $\mathbf{y}_{2i} = (y_{2i1}, y_{2i2}, \dots, y_{2in_i})$ are repeated transplantation status In the application example.

In order to incorporate those associations of mixed outcomes, we can specify the joint density function $f(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_K)$ for a proposed joint model. However, one major challenge for it is the lack of a suitable multivariate joint distribution. Two approaches are proposed for the multivariate joint distribution. The first approach directly specifies the joint distribution by factorizing it into the conditional distribution of one outcome and a marginal distribution of the other outcome. For instance,[?] parameterized the model such that the joint distribution is factorized as the product of the marginal distribution of a continuous response and the conditional distribution of a discrete response given the continuous response or latent variables. Another case is that the binary response is related to an unobserved continuous latent variable, and the latent variable and the continuous response have a joint Gaussian distribution. For instance,[?] factorized the joint distribution as the product of a marginal Bernoulli distribution for a discrete response, and a conditional Gaussian distribution for a continuous response given the discrete response. Our proposed joint model use the second approach to construct the joint distribution.

1. The unstructured covariance matrix:

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n_i} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n_i} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n_i1} & \sigma_{n_i2} & \cdots & \sigma_{n_i}^2 \end{pmatrix}$$

2. The compound symmetry structure (or exchangeable), with constant variance across occasions and constant correlation coefficients :

$$(\sigma^2) \otimes \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdots & \cdots & \cdots & \cdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

3. The autoregressive structure in time series frame:

$$(\sigma^2) \otimes \begin{pmatrix} 1 & \rho & \cdots & \rho^{n_i-1} \\ \rho & 1 & \cdots & \rho^{n_i-2} \\ \cdots & \cdots & \cdots & \cdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \cdots & 1 \end{pmatrix}$$

4. The exponential correlation structure. For example, an $n_i \times n_i$ matrix B can be defined as

$$\exp(B) = \sum_{j=0}^{\infty} \frac{B^j}{j!}$$

In the clinical studies, the coefficients of the correlation matrix indicate that: (i) the repeated measures are positively/negatively correlated with each other, (ii) the correlations often decrease with increasing time separation. In other words, measures that are taken closer together in time are expected to be more highly correlated than measures that further apart in time. By selecting a covariance structure for the latent variables $\mathbf{u}_{ki} = (u_{ki1}, \dots, u_{kin_i})$, the proposed statistical models can better accommodate the dependence correlation among data.

5.3.1 The autoregressive structure in time series frame

In time series analysis frame, the cross correlation between two time series describes the normalized cross covariance function. For the convenient notation, we discuss two mixed outcomes by setting $K = 2$. If let (u_{1_t}, u_{2_t}) be a pair of stochastic processes that are jointly stationary, then the cross covariance of them is given by:

$$\gamma_{u_1 u_2}(\tau) = \mathbb{E}[(u_{1_t} - \mu_{u_1})(u_{2_{t+\tau}} - \mu_{u_2})],$$

where μ_{u_1} and μ_{u_2} are the means of u_{1_t} and u_{2_t} respectively. The cross correlation function $\rho_{u_1 u_2}$ is the normalized cross-covariance function, $\rho_{u_1 u_2}(\tau) = \frac{\gamma_{u_1 u_2}(\tau)}{\sigma_{u_1} \sigma_{u_2}}$, where σ_{u_1} and σ_{u_2} are the standard deviations of processes u_{1_t} and u_{2_t} , respectively.

Now we suppose the two latent series u_{1_t} and u_{2_t} are satisfying the following relationship

$$u_{2_t} = Au_{1_t} + \omega_t. \quad (5.2)$$

For convenience, we can assume that u_{1_t} and u_{2_t} have zero means, and the noise ω_t is uncorrelated with the u_{1_t} series, the cross-covariance function can be computed as

$$\begin{aligned} \gamma_{u_2 u_1}(\tau) &= \mathbb{E}(u_{2_{t+\tau}} u_{1_t}) \\ &= A \mathbb{E}(u_{1_{t+\tau}} u_{1_t}) + \mathbb{E}(\omega_{t+\tau} u_{1_t}) \\ &= A \gamma_{u_1}(\tau) \\ \gamma_{u_1 u_2}(\tau) &= A \gamma_{u_1}(\tau) \end{aligned}$$

In the context of a joint model with random means, we use the cross correlation between u_1 and u_2 to reflect the correlation among different characteristics,

$$\rho_{u_{1_t} u_{2_{t+\tau}}} = \frac{\gamma_{u_1 u_2}(\tau)}{\sqrt{\gamma_{u_1}(0) \gamma_{u_2}(0)}} \quad (5.3)$$

Based on the above result, the cross correlation function $\rho_{u_{1_t} u_{2_{t+\tau}}}$ is determined if the correlation function for u_{1_t} is selected. We can choose the autoregressive covariance structure of u_{k_t} for illustration. If it is assumed that u_{1_i} and u_{2_i} have the same correlation coefficient ρ , then the resulting covariance matrix for (u_{1_i}, u_{2_i}) is

$$\begin{aligned} \Sigma_i &= \begin{pmatrix} \sigma_1^2 & A\sigma_1^2 \\ A\sigma_1^2 & \sigma_2^2 \end{pmatrix} \otimes \begin{pmatrix} 1 & \rho & \dots & \rho^{n_i-1} \\ \rho & 1 & \dots & \rho^{n_i-2} \\ \dots & \dots & \dots & \dots \\ \rho^{n_i-1} & \rho^{n_i-2} & \dots & 1 \end{pmatrix} \\ &= R \otimes T_i, \end{aligned}$$

where σ_1^2 and σ_2^2 represent the variances of the two series u_{1_i} and u_{2_i} separately. T_i is selected as autoregressive structure, and other types could also be chosen, depending on the property of the data set.

Since we know that $\Sigma_i^{-1} = R^{-1} \otimes T_i^{-1}$ and $|\Sigma_i| = |R|^{n_i} |T_i|^2$, so we can write the joint density function f of $\mathbf{u}_i = (u_{1i}, u_{2i})'$ as

$$f(\mathbf{u}_i) = \frac{1}{(2\pi)^{n_i} (|R|^{n_i} |T_i|^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} \mathbf{u}_i' (R^{-1} \otimes T_i^{-1}) \mathbf{u}_i\right\}$$

Then the log-likelihood function of $\mathbf{u}_i = (u'_{1i}, u'_{2i})'$ as

$$\begin{aligned} -n_i \log(2\pi) & - \frac{n_i}{2} \log|R| - \log|T_i| \\ & - \frac{1}{2} \mathbf{u}_i' (R^{-1} \otimes T_i^{-1}) \mathbf{u}_i. \end{aligned} \quad (5.4)$$

From (5.2) and (5.3), we can compute

$$\begin{aligned} \rho_{u_2, u_1} & = \frac{A\sigma_{u_1}^2}{\sqrt{A^2\sigma_{u_1}^2 + \sigma_\omega^2} \sqrt{\sigma_{u_1}^2}} \\ & = \frac{A}{\sqrt{A^2 + \frac{\sigma_\omega^2}{\sigma_{u_1}^2}}}, \end{aligned}$$

since

$$\sigma_{u_2}^2 = A^2\sigma_{u_1}^2 + \sigma_\omega^2.$$

As $|A|$ approaches infinity, $|\rho_{u_2, u_1}|$ goes to 1. The other property is that there is natural constraint on A . Due to $\sigma_{u_2}^2 = A^2\sigma_{u_1}^2 + \sigma_\omega^2$, we have $A^2 < \frac{\sigma_{u_2}^2}{\sigma_{u_1}^2}$. Therefore we will know that R in (5.4) is definite positive.

5.4 The joint likelihood function

The joint probability for \mathbf{y}_i and \mathbf{u}_i , where $i = 1, \dots, N$, can be expressed as:

$$f(\mathbf{y}_i, \mathbf{u}_i) = f(\mathbf{y}_i | \mathbf{u}_i; \mathbf{b}, \boldsymbol{\phi}) q_i(\mathbf{u}_i; \boldsymbol{\sigma}_i^2), \quad (5.5)$$

$$f(\mathbf{y}_i | \mathbf{u}_i, \boldsymbol{\alpha}, \mathbf{b}, \boldsymbol{\phi}) = \prod_{k=1}^2 \prod_{j=1}^{n_i} f(y_{kij} | u_{kij}; \boldsymbol{\alpha}_k, b_k, \phi_k) \quad (5.6)$$

under the conditional independence assumption. Since the latent variables \mathbf{u}_i are unobserved, inference about the parameters b_k , ϕ_k and $\boldsymbol{\sigma}_i^2$ is based on the marginal likelihood function of the observed data:

$$L(\mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2; y) = \prod_{i=1}^N \int f(\mathbf{y}_i | \mathbf{u}_i; \boldsymbol{\alpha}, \mathbf{b}, \boldsymbol{\phi}) q_i(\mathbf{u}_i; \boldsymbol{\sigma}^2) d\mathbf{u}_i \quad (5.7)$$

The maximum likelihood estimates of $\boldsymbol{\alpha}$, $\mathbf{b} = (b_1, b_2)$, $\boldsymbol{\phi} = (\phi_1, \phi_2)$ and $\boldsymbol{\sigma}^2$ are simply those values of \mathbf{b} , $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}^2$ that maximize this likelihood function. However, the integration (5.7) always involves intractable integrals. Consequently, much work has been focused on approximate techniques that seek to avoid the integration. The Monte Carlo EM (MCEM) algorithm, introduced by Wei and Tanner is an extension of the EM algorithm that estimates the expectation in the E-step with a Monte Carlo approximation. Booth and Hobert 1999 Booth proposed to use rejection sampling and multivariate t importance sampling to generate independent samples to construct Monte Carlo approximations. Because of the hierarchical structure of the model, we can apply Monte Carlo EM algorithm.

5.4.1 Monte Carlo EM

We discuss the likelihood function in a general framework for this proposed joint model with latent variables. In the *EM* algorithm, the *E* step imputes the log-likelihood of the complete data, consisting of the observed data and the latent variables, by the conditional expectation of the complete data log-likelihood given the observed data. In the MCEM algorithm, the conditional expectation of the log-likelihood of the complete data is estimated by averaging the conditional log-likelihoods of simulated sets of complete data.

Let $\boldsymbol{\theta}^T = (\mathbf{b}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\sigma}^2)^T$ denote the complete vector of unknown parameters. Monte Carlo averages of simulated variables is used to estimate

$$E[\log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\theta}) | y; \boldsymbol{\theta}],$$

the expectation is with respect to h , the distribution of \mathbf{u} given \mathbf{y} with parameter value $\boldsymbol{\theta}^{(r)}$,

$$h(\mathbf{u} | y; \boldsymbol{\theta}) \propto f(y | \mathbf{u}; \boldsymbol{\alpha}, \mathbf{b}, \boldsymbol{\phi}) q(\mathbf{u}; \boldsymbol{\sigma}^2).$$

To set up the EM algorithm in the context of the proposed joint model, we consider the latent variables, \mathbf{u} , to be the missing data. Let $f(\mathbf{y}, \mathbf{u}; \boldsymbol{\theta})$ represent the joint density of the complete data, then we have,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^r) = \int \log f(\mathbf{y}, \mathbf{u}; \boldsymbol{\theta}) h(\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^{(r)}) d\mathbf{u}, \quad (5.8)$$

where $h(\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^{(r)})$ is the conditional density function of \mathbf{u} given \mathbf{y} and $\boldsymbol{\theta}^{(r)}$. Specifically, draw a random sample, $\mathbf{u}^1, \dots, \mathbf{u}^L$ from $h(\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^{(r)})$.

A Monte Carlo approximation of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$ is given by

$$\hat{Q}_{r+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)}) = \frac{1}{L} \sum_{l=1}^L \log f(\mathbf{y}, \mathbf{u}^l; \boldsymbol{\theta}) \quad (5.9)$$

The implementation of Monte Carlo EM

In the implementation, due to independence among subjects, one may sample from $h(\mathbf{u}_i | \mathbf{y}_i, \boldsymbol{\theta}^{(r)})$ for the i th individual, for $i = 1, \dots, N$. Because of the introduction of Monte Carlo at the E-step, the incomplete data log-likelihood (5.9) is not guaranteed to increase at every iteration. However, the Monte Carlo EM algorithm still converges to the maximum likelihood estimate under suitable regularity conditions by Chan at 1995.

M-step

The M-step maximizes the approximate Q function obtained in the Monte Carlo E-step, with respect to $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(r+1)}$. The MCEM algorithm iterates between the approximate E-step and the M-step, each time drawing a sample of the unobserved data from the conditional distribution given the observed data at the updated parameter value, and maximizing the approximate Q function obtained from the new sample and the updated parameter to get a new estimate of the parameter. As McCulloch 1997 has pointed out, the Monte Carlo M-step is usually relatively simple in the generalized linear mixed model context. The reason is that $\hat{Q}_{r+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$ is the sums of log-likelihoods from two generalized linear models. The first term involves \mathbf{b} and $\boldsymbol{\phi}$, and the second one involves only $\boldsymbol{\sigma}^2$. The first term can be maximized via iteratively reweighed least squares and, depending on the distribution of the conditional distribution, the maximizer of the second term can sometimes be written in closed form.

E-step

The implementation of the Monte Carlo E-step involves sampling the unobserved \mathbf{u} from the conditional distribution of $h(\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^{(r)})$. This requires us to choose an appropriate Monte Carlo sampler that simulates u from a distribution that is as close as possible to the target distribution $h(\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^{(r)})$. The choice could be rejection sampling, importance sampling, or dependent samples from an invariant target distribution based on Markov chain Monte Carlo methods. Rejection sampling is more efficient when sample sizes are small, whereas importance sampling is better with larger sample sizes. Both of them are useful when direct simulation from h is difficult or impossible but direct simulation from another distribution similar to h is possible. When the acceptance rate for the rejection sampler is very low, it may be more efficient to use the importance sampling. The approximation of the complete

data likelihood based on the importance sampling is

$$\omega_l = \frac{\exp\{\log f(\mathbf{y}, \mathbf{u}^l; \boldsymbol{\theta})\}/q(\mathbf{u}^l)}{\sum_{k=1}^L \exp\{\log f(\mathbf{y}, \mathbf{u}^k; \boldsymbol{\theta})\}/q(\mathbf{u}^k)} \quad (5.10)$$

and (\mathbf{u}^l) are random samples from the importance density q .

The choice of missing data has two advantages. Firstly the \mathbf{y} are independent when the \mathbf{u} known. Secondly, the M-step of the EM algorithm maximizes the complete data likelihood with respect to \mathbf{b} , $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}^2$. The M-step with respect to \mathbf{b} and $\boldsymbol{\phi}$ only needs $f(\mathbf{y}|\mathbf{u})$, so it becomes a standard generalized linear model problem given the values of \mathbf{u} known.

Steps of Monte Carlo EM algorithm

The MCEM algorithm for this proposed joint models is as follows,

1. Choose starting values $\boldsymbol{\theta}^{(0)}$, and initial sample size L .
2. Generate L values, $\mathbf{u}_r^1, \dots, \mathbf{u}_r^L$ from $h(\mathbf{u}|\mathbf{y}, \boldsymbol{\theta}^{(r)})$ using rejection or importance sampling methods.
3. Using the approximation (5.9) or (5.10) to obtain $\boldsymbol{\theta}^{(r+1)}$ by maximizing $\hat{Q}_{r+1}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(r)})$.
4. If convergence is achieved, then declare $\boldsymbol{\theta}^{(r+1)}$ to be the maximum likelihood estimate of $\boldsymbol{\theta}$; otherwise, return to Step 2.

Booth at 1998 proposed a multivariate Student t importance density whose mean and variance match the mode and curvature of h . More specifically, we write $h(\mathbf{u}_i|\mathbf{y}_i, \boldsymbol{\theta}) = a_i \exp\{l(\mathbf{u}_i)\}$, where a_i is the unknown normalizing constant.

$$\begin{aligned} l_i(\mathbf{u}_i) &= \log\{f(\mathbf{y}_i|\mathbf{u}_i; \boldsymbol{\alpha}, \mathbf{b}, \boldsymbol{\phi})\} + \log\{q(\mathbf{u}_i; \boldsymbol{\sigma}^2)\} \\ &= \sum_{k=1}^2 \sum_{j=1}^{n_i} \log\{f(\mathbf{y}_{kij}|\mathbf{u}_i; \boldsymbol{\alpha}_k, \mathbf{b}_k, \boldsymbol{\phi}_k)\} - \frac{1}{2} \log |2\pi \boldsymbol{\Sigma}_i| - \frac{1}{2} \mathbf{u}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{u}_i, \end{aligned}$$

Let $l_i^{(1)}(\mathbf{u}_i)$ and $l_i^{(2)}(\mathbf{u}_i)$ denote the vector of the first derivatives and the hessian matrix of the second derivatives of $l_i(\mathbf{u}_i)$ separately,

$$l_i^{(1)}(\mathbf{u}_i) = \text{vec} \left\{ \text{vec} \left\{ \frac{y_{kij} - \mu_{kij}}{a_{kij}(\boldsymbol{\phi}) b_k''(\boldsymbol{\theta}_{kij}) g'(\mu_{kij})} \right\}_j \right\}_k - \boldsymbol{\Sigma}_i^{-1} \mathbf{u}_i \quad (5.11)$$

$$l_i^{(2)}(\mathbf{u}_i) = -W_i - \boldsymbol{\Sigma}_i^{-1}, \quad (5.12)$$

where W_i is the diagonal matrix of iterative weights, $\frac{1}{a_{kij}(\boldsymbol{\phi}_k) b_k''(\boldsymbol{\theta}_{kij}) g'(\mu_{kij})^2}$, for $j = 1, \dots, n_i$, and $k = 1, 2$.

Suppose that $\tilde{\mathbf{u}}_i$ is the maximizer of $l_i(\mathbf{u})$ satisfying the equation $l_i^{(1)}(\mathbf{u}) = 0$. The Laplace approximation of the mean and variance are $\tilde{\mathbf{u}}_i$ and $-l_i^{(2)}(\tilde{\mathbf{u}}_i)^{-1}$ respectively. The approximations to the conditional mean and variance of \mathbf{u}_i are

$$E(\mathbf{u}_i|\mathbf{y}_i) \approx \tilde{\mathbf{u}}_i$$

$$\text{Var}(\mathbf{u}_i|\mathbf{y}_i) \approx -l_i^{(2)}(\tilde{\mathbf{u}}_i)^{-1}$$

5.4.2 Information Matrix

Denote the MLE from the MCEM algorithm by $\hat{\boldsymbol{\theta}}$. Louis (1982) showed that the observed information matrix is given by

$$-E\left\{\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right\}_{\mathbf{u}=\hat{\mathbf{u}}} - \text{Var}\left\{\frac{\partial l(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u})}{\partial \boldsymbol{\theta}}\right\}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad (5.13)$$

where the expectation and variance are with respect to $h(\mathbf{u}|\mathbf{y}, \hat{\boldsymbol{\theta}})$.

5.5 The application to Clinical Transplant Data

Patients who received a renal transplant were extracted from United Network for Organ Sharing (UNOS), which direct the transplant community to reduce disparity in access to transplant, to allocate organs over as wide of a geographic area as possible, and to ensure organs to be allocated on the basis of medical necessity. The most common measure of kidney function is called the estimated glomerular filtration rate (GFR), which was measured once a year in the study, by the formula "4-variable MDRD" (serum creatinine, age, race, and gender). Kidney function is considered to be normal, when GFR is in the level of 90 or higher. On the other hand, if GFR is less than 15, the kidney function is considered to be a failure. Patients need to have repeated kidney transplants when kidney graft fails.

5.5.1 Model specification in the application example

We model the kidney function status after the first kidney transplant among the following-up periods. Specifically, we treat GFR as the continuous variable and retransplant status as the binary variable. We then combine them together as mixed outcomes to jointly describe how kidney function changes and the disease progress over time after the first transplant. We use the proposed joint model to determine the association between mixed outcomes over time following the first kidney transplant with other covariates such as age and gender included.

Let y_{1ij} and y_{2ik} denote GFR and the retransplant status, where $j = 1, \dots, n_i$; $k = 1, \dots, n_i$; $i = 1, \dots, N$. The binary response y_{2ik} takes 1, if the re-transplant occurred, otherwise 0. We assume y_{1ij} and y_{2ik} are conditionally independent given u_{1i} and u_{2i} according

to $y_{1ij}|u_{1ij} \sim N(\mu_{1ij}, \sigma^2)$, our proposed joint model is presented in the following:

$$\begin{cases} Y_{1ij}|u_{1ij} \sim N(\mu_{1ij}, \sigma^2), \text{ and } \mu_{1ij} = \boldsymbol{\alpha}_1^T \mathbf{X}_i + u_{1ij} \\ Y_{2ij}|u_{2ij} \sim \text{Poisson}(\mu_{2ik}), \text{ and } \text{logit}(\mu_{2ik}) = \boldsymbol{\alpha}_2^T \mathbf{X}_i + u_{2ij}, \end{cases} \quad (5.14)$$

where $\mu_{1ij} = \alpha_{10} + \alpha_{11}\text{time} + \alpha_{12}\text{age} + \alpha_{13}\text{gender} + \alpha_{14}\text{Race} + \alpha_{15}\text{ESRD cause} + u_{1ij}$, and $y_{2ik}|u_{2ik} \sim \text{Poisson}(\mu_{2ik})$, where $\text{logit}(\mu_{2ik}) = \alpha_{20} + \alpha_{21}\text{time} + \alpha_{22}\text{age} + \alpha_{23}\text{gender} + \alpha_{24}\text{Race} + \alpha_{25}\text{ESRD cause} + u_{2ik}$.

5.5.2 Model results

As mentioned in the model specification, it is assumed that (u_{1i}, u_{2i}) is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ as formulated in (5.4). We applied the Monte Carlo EM algorithm based on importance sampling. A multivariate Student t distribution with 45 degrees of freedom was chosen as the initial distribution. We started with $L = 50$, and increased by $L = L + L/10$, until $L = 5000$.

An important issue in implementing the Monte Carlo EM algorithm is to assess the convergence of the algorithm. We used the criteria that when the relative change in the parameter values from successive iterations is small.

$$\max|\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}| < \delta,$$

where δ is predetermined constants. We set $\delta = 0.0001$. The method involves the initial value $\boldsymbol{\theta}^{(0)}$, and the resulting approximation is local in nature. Thus we iterated our procedure a few times by updating $\boldsymbol{\theta}^{(0)}$ to the current estimate of $\boldsymbol{\theta}$. We also evaluated the marginal likelihood at several parameter values.

The main results from the proposed joint model are shown in Table 5.1 The estimate of the time slope of the continuous variable GFR is -8.42 , which is statistically significant, and the negative sign indicates that GFR decreases over time after the first transplant. In addition, the estimate of the time slope of the binary variable retransplant is 3.98 , which is also statistically significant. The positive sign shows that the probability of getting a second transplant is increasing over time. The estimate of autocorrelation coefficient ρ is 0.79 , which shows that there is a positive correlation in the latent processes u_{1i} and u_{2i} .

5.6 Conclusion

We developed a joint model for observations of mixed response variable for the cluster and longitudinal data, and applied the MCEM algorithm to find the MLEs of the generalized linear latent variable models for multivariate responses. With the nice properties of kronecker product, the form of the log-likelihood function of the mixed outcomes is explicitly expressed, especially the inverse and the determination of the covariance of the latent pro-

Table 5.1: Estimates for parameters in Model (5.14). The standard errors of the estimates are given in brackets.

Parameters	The longitudinal submodel		The submodel	
	Coef.(SE)	<i>P</i> value	Coef.(SE)	<i>P</i> value
Intercept	55.93(1.45)	<0.001	-15.94(2.12)	<0.001
Time	-8.42(1.56)	<0.001	3.98(1.24)	<0.001
Age (per year)	-0.17(0.05)	< 0.001	0.02(0.01)	0.044
Female	5.39(0.73)	< 0.001	-0.23(0.03)	0.029
Black	-3.69(1.34)	< 0.001	0.17(0.08)	0.020
Other	-6.45(1.08)	< 0.001	0.23(0.07)	< 0.001
Diabetes	0		1.0	
Hypertension	0.86(0.27)	0.001	-0.41(0.06)	0.042
Glomerular disease	0.76(0.32)	0.002	-0.54(0.08)	0.012
Polycystic disease	1.09(0.29)	< 0.001	-0.78(0.11)	0.021
Others	0.93(0.27)	< 0.001	-0.87(0.10)	0.047
Other parameters				
σ_1^2		857.05(148.25)		
σ_2^2		89.87(15.23)		
σ_ω^2		2.10(0.23)		
ρ		0.79(0.03)		
A		-0.32(0.01)		

cesses. We also connect the modeling of the latent processes with exponential covariance structure to the time series. At each *E*-step, we approximate the *Q* function by using the rejection sampling or the importance sampling approach. For the importance sampling, we use the Laplace approximation to find the instrumental distribution with the mean and covariance coming from the posterior distribution of the latent processes given the responses. We demonstrated the methodology with a kidney study, measuring eGFR and the retransplant status for patients.

Chapter 6

A Predict Model with a Polynomial Effects Covariate in Presence of Measurement Errors

6.1 Motivation

As mentioned in the introduction chapter, there are three strategies to narrow the gap between the supply and demand of kidney organ. The former chapters are focus on the second strategy. My future work is focus on the development of predict statistical models, which are motivated from the third strategy.

It is known that deceased/living kidney donors are routinely shared within seven geographically defined regions, but they are infrequently shared between regions. For example, the donor rate per million population (RPMP) in 2004 varied from 6.0 in Manitoba to 18.0 in Quebec. The reasons for this variability remain unclear. The RPMP does not account for population differences between regions that may impact organ donation, which makes it difficult to determine if regional variation is due to differences in the number potential organ donors, differences in organ procurement practices, or differences in family consent rates for organ donation, community social work, household income, donor race, etc.

A major barrier to understand regional differences in deceased organ donation is lack of an informative metric of donor activity. Understanding why some regions have higher deceased/living donor rates than others will inform health policy to improve kidney donation in all regions. Therefore, it is useful to develop some statistical models to explore the information metric of donor activity in the population.

6.2 Introduction

Let D denote the kidney donor in the population under investigation. The incidence rate for D is attributed to the continuous exposure E as well as to p other covariates $\mathbf{Z} = (Z_1, \dots, Z_p)$. Suppose a average sample unit consists of g subjects whose individual expo-

sure levels are E_1, \dots, E_g , and they are expensive to be measured individually. Or it is already available from other resources such as Census data. For example, we don't individual donor patient social economical information such as the household income, but we can know the average patient zip code level income where they live. According to the design, only a aggregated measurement of exposure in the form of sum, $\sum_{j=1}^g E_j$, or the form of average $\sum_{j=1}^g E_j/g$, is actually available. In the meanwhile, the other p covariates Z_1, \dots, Z_p are available in the individual level. For the ease of exposition, we simply consider the case of aggregated exposure $S = \sum_{j=1}^g E_j/g$ in the rest of this thesis.

A remarkable finding by Weinberg and Umbach (1999) is that under the multiplicative model for the probability of D , the set-based logistic regression model can bypass the aggregation. The resulting model takes the form of

$$\log \left[\frac{P(\text{case unit}|S, Z)}{P(\text{control unit}|S, Z)} \right] = \alpha g + \beta S + \ln(r_g) + \boldsymbol{\gamma}^T \mathbf{Z} \quad (6.1)$$

where α and β are the unknown regression coefficients, and r_g is the ratio of the number of case units of size g over the number of control units of size g . Clearly, β is the parameter of central interest, which represents the linear effect of pooled exposure. Note that in model (6.1), the other covariates are muted just for simplicity, but they would appear in the same aggregation form as that of the S should their linear terms be included in the study.

Unfortunately, this trick works only for linear exposure effects. For example, consider a simple quadratic exposure, E_j^2 . A similar derivation will lead to a logit model with the linear predictor of the form: $\alpha g + \beta_1 S + \beta_2 SS + \ln(r_g)$, with $SS = \sum_{j=1}^g E_j^2/g$ whose measurement is apparently unavailable. What is available is $S^2 = (\sum_{j=1}^g E_j/g)^2$, which is larger than $SS = \sum_{j=1}^g E_j^2/g$. According to our simulation in Section 2, if one naively replaces SS by S^2 , the resulting model, $\alpha g + \beta_1 S + \beta_2^* S^2 + \ln(r_g)$, will exaggerate the true curvature and hence overestimate the quadratic exposure effect.

6.3 Model Specification

To overcome this difficulty arising from the utility of parametric logistic models to assess polynomial effects of pooled exposures, we propose a nonparametric kernel approach to correct measurement error. The feasibility of the nonparametric solution is ensured by the fact that the nonparametric kernel regression can be conducted in a form of local linear fit. Precisely, Weinberg and Umbach's success in bypassing the aggregation for linear exposure effects in model (6.1) would work locally. This local property will propagate ultimately into global nonlinear exposure effect.

Suppose a generalize logistic model for the risk for disease D as follows:

$$\text{logit}\{P(D|E, Z)\} = \alpha + \gamma^T \mathbf{Z} + \theta(E). \quad (6.2)$$

where exposure E is a continuous random variable with density f in the general population. To make α interpretable, we set $\theta(\mu_E) = 0$ with μ_E being the population mean exposure. Similar constraints may be imposed on the other nonparametric functions in the model. So, the intercept term α may be regarded as the base log-odds at zero exposure as well as zero covariates. According to Prentice and Pyke (1979), these nonparametric functions can be estimated in the framework of logistic regression with the utility of back-fitting algorithm (Hastie and Tibshirani (1990)).

6.4 Method

To derive our proposed logit model, we consider randomly pooling the specimens from cases with an equal size of g . The corresponding exposure values from g randomly selected cases are denoted by E_1, \dots, E_g , and only the sum $S = \sum_{j=1}^g E_j$ is measured. Let $E = (E_1, \dots, E_g)$. For the ease of exposition, only the exposure covariate is included in the following derivation. It is easy to see that the density of E conditional on status D is

$$\begin{aligned} h(E|g \text{ cases}) &= \frac{P(g \text{ cases}|E)P(E)}{\{P(D)\}^g} \\ &= \frac{\prod_{j=1}^g P(D|E_j) \prod_{j=1}^g f(E_j)}{\{P(D)\}^g} \end{aligned}$$

where $E_g = S - \sum_{j=1}^{g-1} E_j$. Integrating out the $g - 1$ unobserved exposures E_1, \dots, E_{g-1} , we obtain the conditional density of S given the E_j 's from g randomly sampled cases as follows:

$$\begin{aligned} h(S|g \text{ cases}) &= \frac{1}{\{P(D)\}^g} \int \cdots \int \prod_{j=1}^{g-1} P(D|E_j) f(E_j) P(D|E_g = S - \sum_{j=1}^{g-1} E_j) \times \\ &\quad f(S - \sum_{j=1}^{g-1} E_j) dE_1 \cdots dE_{g-1}. \end{aligned} \quad (6.3)$$

Plugging the nonparametric logit model (6.2) in the density of convolution (6.3), we yeild

$$\begin{aligned} h(S|g \text{ cases}) &= \frac{\exp\{g\alpha + \sum_{j=1}^g \theta(E_j)\}}{\{P(D)\}^g} \times \\ &\quad \int \cdots \int a(E)^{-1} \prod_{j=1}^{g-1} f(E_j) f(S - \sum_{j=1}^{g-1} E_j) dE_1 \cdots dE_{g-1}, \end{aligned} \quad (6.4)$$

where

$$a(E) = \prod_{j=1}^{g-1} [1 + \exp\{\alpha + \theta(E_j)\}] [1 + \exp\{\alpha + \theta(S - \sum_{j=1}^{g-1} E_j)\}].$$

A similar derivation leads to the analogous conditional density of S for controls:

$$\tilde{h}(S|g \text{ controls}) = \frac{1}{\{P(D^c)\}^g} \times \int \cdots \int a(E)^{-1} \prod_{j=1}^{g-1} f(E_j) f(S - \sum_{j=1}^{g-1} E_j) dE_1 \cdots dE_{g-1} \quad (6.5)$$

where D^c denotes the status of no disease.

It follows immediately that the odds for the occurrence of a case set is given by

$$\begin{aligned} \frac{P(\text{case set}|S)}{P(\text{control set}|S)} &= \frac{h(S|\text{case set})P(\text{case set})}{\tilde{h}(S|\text{control set})P(\text{control set})} \\ &= \exp\{\alpha^*g + \sum_{j=1}^g \theta(E_j) + \ln(r_g)\}, \end{aligned}$$

where $\alpha^* = \alpha + \ln P(D^c) - \ln P(D)$ and r_g denotes the ratio of the number of case sets of size g over the number of control sets of size g in the setting of 1 : 1 case-control design. Obviously, $\ln(r_g)$ is an offset of this logistic regression model. Moreover, the set-based nonparametric logistic regression model takes the form

$$\text{logit}\{P(\text{case set}|S)\} = \alpha^*g + \ln(r_g) + \boldsymbol{\gamma}^T \mathbf{Z} + \sum_{j=1}^g \theta(E_j), \quad (6.6)$$

where $\ln(r_g)$ is the offset. The objective is to estimate function $\theta(\cdot)$ based on observed $S = \sum_{j=1}^g E_j$, the sum of exposures from a set of g randomly selected cases or controls.

6.5 Local linear fitting approach

We now develop the local linear fitting approach (Fan and Gijbels, 1996) to estimate of $\theta(\cdot)$. Suppose we observe pooled data $(Y_i, S_i), i = 1, \dots, n$, where $Y_i = 1$ for a case set and $Y_i = 0$ for a control set. Under the 1 : 1 design, there are $n/2$ case sets and $n/2$ control sets.

To proceed, we first fix a target value of E , say E_0 , at which the functional value $\theta(E_0)$ will be estimated. Then, taking a linear Taylor expansion of $\theta(\cdot)$ around the E_0 , we can obtain a local linear approximation:

$$\begin{aligned} \sum_{j=1}^g \theta(E_j) &\approx \beta_0g + \beta_1 \sum_{j=1}^g (E_j - E_0) \\ &= \beta_0g + \beta_1(S - gE_0). \end{aligned}$$

Furthermore, we invoke the kernel smoothing technique based on the local quasi-likelihood given as follows:

$$l_q = \sum_{i=1}^n h(S_i - E_0) [Y_i \{\ln(r_g) + (\alpha^* + \beta_0)g + \beta_1(S_i - gE_0)\} - \log\{1 + \exp(\ln(r_g) + (\alpha^* + \beta_0)g + \beta_1(S_i - gE_0))\}] \quad (6.7)$$

where $h(u) = (u/h)/h$ is a kernel weight function with bandwidth h . We adopt the Gaussian kernel and estimate the bandwidth h by either the cross-validation or the direct plug-in method in this paper. Maximizing l_q with respect to the parameters will output their estimates. If parameter β_0 were estimable, then we would immediately obtain $\hat{\theta}(E_0) = \hat{\beta}_0$. However, we are only able to estimate $\alpha^* + \beta_0$ from this local fitting, and as a matter of fact, parameters α^* and β_0 are not identifiable. Nevertheless, this method allows us to get $\hat{\beta}_1$, which estimates the first derivative of $\theta(\cdot)$ at E_0 , namely, $\hat{\theta}(E_0) = \hat{\beta}_1$.

At the final step, simply convert the estimated first derivative $\hat{\theta}(\cdot)$ into the estimate of the original function $\theta(\cdot)$ by integration:

$$\hat{\theta}(\cdot) = \int \hat{\theta}(E) dE,$$

subject to the condition $\theta(\mu_E) = 0$. Numerically, we implement this conversion as follows:

Step 1. Estimate the mean exposure $\hat{\mu}_E = \frac{1}{ng} \sum_{i=1}^n S_i$. According to the Slutsky's theorem,

$$\theta(\hat{\mu}_E) \xrightarrow{p} \theta(\mu_E) = 0, \text{ as } n \rightarrow \infty.$$

Step 2. Select a sequence of equally spaced dense target values $E^k, k = 1, \dots, K$ allocated on the two sides of $\hat{\mu}_E$, with the distance of two adjacent values equal to δ . Let $k^* = \hat{\mu}_E$. Run the local logistic regression at each of the target values and record $\hat{\beta}_1 = \hat{\beta}_1(E^k)$ from each fit.

Step 3. Set $\hat{\theta}(E^{k^*}) = 0$, and for each target value E^k , assign $\hat{\theta}(E^k) = \hat{\beta}_1(E^k)$. Based on the system of the difference equations,

$$\delta \hat{\theta}(E^k) = \hat{\theta}(E^k) - \hat{\theta}(E^{k-1}), \quad k = 2, \dots, K,$$

we can solve for $\hat{\theta}(\cdot)$ at the target values as follows:

$$\hat{\theta}(E^k) = \begin{cases} -\delta \sum_{l=k+1}^{k^*} \hat{\theta}(E^l), & \text{if } k < k^* \\ \delta \sum_{l=k^*}^k \hat{\theta}(E^l), & \text{if } k > k^*. \end{cases}$$

To find the estimate of β_1 at a given target value E_0 , we apply the New-Raphson algorithm. First Let

$$\eta_i = \ln(r_g) + \beta_0^* + \beta_1(S_i - gE_0).$$

Then, differentiating the local quasi-likelihood (6.7) with respect to β_0^* and β_1 lead to the following estimating equations:

$$\begin{aligned}\frac{\partial l_q}{\partial \beta_0^*} &= \sum_{i=1}^n h(S_i - gE_0)(Y_i - \mu_i) \\ \frac{\partial l_q}{\partial \beta_1} &= \sum_{i=1}^n h(S_i - gE_0)(Y_i - \mu_i)(S_i - gE_0).\end{aligned}$$

So, the quasi-score vector can be expressed as follows:

$$\Psi = \begin{pmatrix} \sum_{i=1}^n h(S_i - gE_0)(Y_i - \mu_i) \\ \sum_{i=1}^n h(S_i - gE_0)(Y_i - \mu_i)(S_i - gE_0) \end{pmatrix} = X'(Y - \mu)$$

where $Y = (Y_1, \dots, Y_n)'$, $\mu = (\mu_1, \dots, \mu_n)'$, $= \{h(S_1 - gE_0), \dots, h(S_n - gE_0)\}$, and

$$X' = \begin{pmatrix} 1 & 1 & \dots & 1 \\ S_1 - gE_0 & S_2 - gE_0 & \dots & S_n - gE_0 \end{pmatrix}.$$

To solve equation $\Psi = 0$, we invoke the Newton-Raphson algorithm

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + (X'\widetilde{W}X)^{-1}X(Y - \mu),$$

where $\widetilde{W} = W$ and $W = [\mu_1(1 - \mu_1), \dots, \mu_n(1 - \mu_n)]$.

6.6 Statistical inference

For the convenient notation, let x is an exposure and Y be the case or control (donor or not), we need to figure out the densities $f(x|Y = 1)$ and $f(x|Y = 0)$.

From Bayes formula,

$$f(x|Y = 1) = \frac{Pr(Y = 1|X)f(x)}{Pr(Y = 1)}$$

where from the assumed model, $\text{logit}[Pr(Y = 1|x)] = \alpha + \beta(x - \mu)^2$, hence

$$\begin{aligned}Pr(Y = 1|x) &= \frac{\exp\{\alpha + \beta(x - \mu)^2\}}{1 + \exp\{\alpha + \beta(x - \mu)^2\}} \\ Pr(Y = 0|x) &= \frac{1}{1 + \exp\{\alpha + \beta(x - \mu)^2\}}\end{aligned}$$

For simplicity, we choose $x \sim N(\mu^*, \sigma^2)$ with $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x-\mu^*)^2}{2\sigma^2}\}$. $Pr(Y = 1)$ and $Pr(Y = 0)$ are constants with respect to x . We denote them as C_1 and C_2 respectively. Therefore,

$$\begin{aligned} f(x|Y = 1) &= C_1 \frac{\exp\{\alpha + \beta(x - \mu)^2\}}{1 + \exp\{\alpha + \beta(x - \mu)^2\}} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x - \mu^*)^2}{2\sigma^2}\} \\ &= C_1^* \frac{\exp\{\alpha + \beta(x - \mu)^2 - \frac{1}{2\sigma^2}(x - \mu^*)^2\}}{1 + \exp\{\alpha + \beta(x - \mu)^2\}} \end{aligned} \quad (6.8)$$

$$\begin{aligned} f(x|Y = 0) &= C_2 \frac{1}{1 + \exp\{\alpha + \beta(x - \mu)^2\}} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(x - \mu^*)^2}{2\sigma^2}\} \\ &= C_2^* \frac{\exp\{-\frac{(x - \mu^*)^2}{2\sigma^2}\}}{1 + \exp\{\alpha + \beta(x - \mu)^2\}} \end{aligned} \quad (6.9)$$

where $C_1^* = \frac{C_1}{\sqrt{2\pi}\sigma}$ and $C_2^* = \frac{C_2}{\sqrt{2\pi}\sigma}$.

To generate $x|Y = 1$ using acceptance-rejection method, we denote the left hand side of Equation 1 as $f_1(x)$, then depending on values of parameters, we have following cases:

Case I, α and β are negative, $e^{\alpha+\beta(x-\mu)^2} < 1$, we put

$$\begin{aligned} g_1(x) &= C_1^* \exp\{\alpha + \beta(x - \mu)^2 - \frac{1}{2\sigma^2}(x - \mu^*)^2\} < f_1(x) \\ &= C_1^* \exp\{\alpha + \mu^2 - \frac{\mu^{*2}}{2\sigma^2} - \frac{(\mu\beta - \frac{\mu^*}{2\sigma^2})^2}{\beta - \frac{1}{2\sigma^2}} - \frac{(x - \frac{\mu\beta - \frac{\mu^*}{2\sigma^2}}{\beta - \frac{1}{2\sigma^2}})^2}{2\sigma^2/(1 - 2\sigma^2\beta)}\} \\ &= C_1^* \sqrt{\frac{2\pi\sigma^2}{1 - 2\sigma^2\beta}} \exp\{\alpha + \mu^2 - \frac{\mu^{*2}}{2\sigma^2} - \frac{(\mu\beta - \frac{\mu^*}{2\sigma^2})^2}{\beta - \frac{1}{2\sigma^2}}\} \frac{1}{\sqrt{\frac{2\pi\sigma^2}{1 - 2\sigma^2\beta}}} \exp\{-\frac{(x - \frac{\mu\beta - \frac{\mu^*}{2\sigma^2}}{\beta - \frac{1}{2\sigma^2}})^2}{2\sigma^2/(1 - 2\sigma^2\beta)}\} \\ &= C_1^* k_1 h_1(x) \end{aligned}$$

where $k_1 = \sqrt{\frac{2\pi\sigma^2}{1 - 2\sigma^2\beta}} \exp\{\alpha + \mu^2 - \frac{\mu^{*2}}{2\sigma^2} - \frac{(\mu\beta - \frac{\mu^*}{2\sigma^2})^2}{\beta - \frac{1}{2\sigma^2}}\}$ and $h_1(x)$ is the density of $N(\frac{\mu\beta - \frac{\mu^*}{2\sigma^2}}{\beta - \frac{1}{2\sigma^2}}, \frac{\sigma^2}{1 - 2\sigma^2\beta})$.

Now the procedure of generating random number from density $f_1(x)$ will be:

- Generate x from $h_1(x)$;
- Generate r.v. U from $U(0, 1)$, calculate $z = C_1 * k_1 U h_1(x)$;
- If $z < f_1(x)$ or equivalently, $k_1 U h_1(x) < \frac{\exp\{\alpha + \beta(x - \mu)^2 - \frac{1}{2\sigma^2}(x - \mu^*)^2\}}{1 + \exp\{\alpha + \beta(x - \mu)^2\}}$, accept x . Otherwise, discard it.

Similarly for $x|Y = 0$, just simply let $k_2 = \sqrt{2\pi}\sigma$ and $h_2(x)$ be the density of $N(\mu^*, \sigma^2)$ and generate r.v. with the same procedure.

Case II, both α and β are positive, $e^{\alpha+\beta(x-\mu)^2} > 1$. Let $k_1 = \sqrt{2\pi}\sigma$ and $g_1(x)$ be the density of $N(\mu^*, \sigma^2)$. Let $k_2 = \sqrt{\frac{2\pi\sigma^2}{2\sigma^2\beta+1}} \exp\{\frac{2\sigma^2\beta+\mu^*}{2\sigma^2\beta+1} - \alpha - \beta\sigma^2 - \frac{\mu^{*2}}{2\sigma^2}\}$ and $g_2(x)$ be the density of $N(\frac{2\sigma^2\beta+\mu^*}{2\sigma^2+1}, \frac{\sigma^2}{2\sigma^2\beta+1})$. The generating process are similar as in Case I.

Case III, one in α and β is positive, the other negative, choose one from procedures of case I or case II that gives more efficiency.

6.7 Simulation

This section gives several simulations, which reveal that the cluster/pooled exposure assessment has good operating characteristics. We use the local linear logistic regression with normal kernel smoothing method, although this non-parameter method may cause remarkable loss of accuracy in estimating the coefficients for a global logistic model.

6.7.1 Simulation setting

We simulated a sample with 4000 cases and 4000 controls, and then we applied various patterns of grouping to each simulated study.

We examined scenarios in which all pooling sets were the same size, with $g = 1, 2, 4$ or 8, the number of assays required under these strategies are 8000, 4000, 2000 and 1000, respectively.

The assumption made in the construction of the simulation are as follow: let X denote the exposure level, it is a continuous random variables follows normal distributions for both case and control group, but with different means, and a common value of 4 for standard deviation for both groups. Then we derive the conditional distribution of X given case or control, which are

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}$$

$$p(X|\bar{Y}) = \frac{p(\bar{Y}|X)p(X)}{p(\bar{Y})}$$

where $p(Y|X) = \frac{\exp(\alpha+\beta(X-\mu)^2)}{1+\exp(\alpha+\beta(X-\mu)^2)}$, $p(\bar{Y}|X) = \frac{1}{1+\exp(\alpha+\beta(X-\mu)^2)}$, we choose X to follow normal distribution, $p(Y)$ and $p(\bar{Y})$ are constants with respect to exposure. We will discuss the detailed derivation in Appendix.

In order to generate random samples from those distribution, we applied Rejection/Acceptation method. The idea is to find a density, which is easy to generate data from. After multiplying a certain constant, it covers the desired density. We accept this data only if it falls below the desired density curve. This function should not be too higher than the desired density to ensure the efficiency of the random number generating process.

6.7.2 Simulation results

Results are reports as follows. In our implementation of the simulated study, we choose $\alpha = 1$, $\beta = -0.8$, $\mu = 10$, the mean of exposure level $\mu^* = 7$, the standard deviation of the exposure level $\sigma = 4$ and bandwidth of normal kernel $h = 1.5$.

Simulation result 1

We first run the logistic model with only linear term for pooling scenarios $g = 1, 2, 4$ and 8 . We found that none of these models achieved a significant fitting. The parameters β_0 and β_1 are listed in the following table. We observe that these values are quite different from each other, which is not surprising.

g	$\hat{\beta}_0$	$\hat{\beta}_1$
1	-3.487613	0.4085145
2	-7.562447	0.8693702
4	-16.007922	1.8087377

Simulation result 2

We run the logistic model with quadratic term, under various pooling sizes. From the results listed in the table, we found that, when $g = 1$, the estimates are fairly close to the underlying true model. But, for $g = 2, 4$ and 8 , the estimates are far from the true values. The bigger the pooling size, the further away. Apparently, the estimates for β_2 get smaller and smaller when the pooling size gets larger. Notice that small value of β_2 implies strong curvature. So, by doing pooling, we are actually exaggerating the curvature of the data.

g	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
1	-69.43038	14.36120	-0.7192559
2	-102.14547	20.84215	-1.0417566
4	-213.44388	43.29466	-2.1683035
8	-555.51642	112.47162	-5.6473384

Simulation result 3

At last, we run the local logistic regression with linear term with normal kernel. Since this is the nonparametric method, we can not get the global estimates as above. Instead, we get the predictions of log odds ratio as well as the probability of having the disease at the sequence of target exposure level. We plot the curves for each pooling strategy to compare with above results. We illustrated in the plots in the following three cases,

1. $g = 1$, the local fitting curve is close to the underlying true model.
2. $g = 2$, the local fitting curve which is drawn in black is satisfactorily close to the underlying true model. But, since we do not have the information of the individual data, the underlying true model which is global logistic model based on. The curvature of the local fitting curve is much closer to the true value than that of the global logistic model with quadratic term, which is shown red.

3. $g = 4$, we have the similar result as $g = 2$. But, the logistic model with quadratic term is steeper. There is some difference between the local logistic curves for $g = 2$ and $g = 4$. Still, the local logistic model performs better than its global counterpart.
4. $g = 8$, we observe the local logistic curve is almost identical to the global logistic model with only linear term.

The logistic model with quadratic term is further away from the true model. In the top left plot, we depict the density curves and the corresponding dominant functions that we use to generate case and control exposure levels by Acceptance/Rejection method. The top right plot shows the differences of those global logistic models with linear term for $g = 1, 2, 4$ and 8 . None of them is close to the true model which is also shown in the plot. The bottom left plot shows the shrinking effect of the global logistic models with quadratic term as the pooling size goes bigger, i.e., the discrepancies of those models from the underlying true model is also getting more obvious. The bottom right plot shows that the fitted curves by local logistic model with linear term are all reasonable except for $g = 8$. Our proposed nonparametric method successfully captured the curvature of the sample after pooling. The failure to capture the curvature when $g = 8$ is mainly caused by the concentration of the exposure levels for both control and case groups to their own centers. Figure 3 shows the comparison of log odds ratio instead of probability. The change is synchronous as in the previous figure.

6.8 The application to kidney transplant data

6.8.1 Data Resource

Simulation study is done to show the feasibility of the nonparametric approach. It will be more convincing to apply the proposed approach to a real data analysis. We applied our method to the association between the probability of kidney transplant with affecting factors by using US kidney patient data set. We linked to United State Census to get patient socioeconomic status data with a reported residential zip code.

there was an inconsistent association between adjusted Hazard ratio and socioeconomic status estimated from US census (Axelrod David, 2010). ESRD patients with high socioeconomic status were less likely to transplant (10%) than those in low- and middle- socioeconomic, where socioeconomic status was estimated from US census data. It isn't appropriate to treat the variable household income from US Census as an individual variable, since it is the total income of all persons who lived in that zip code area.

6.8.2 Model specification

In our study, it is more appropriate to treat the variable household income as a pooled variable instead of an individual variable to avoid the inconsistent misleading result after

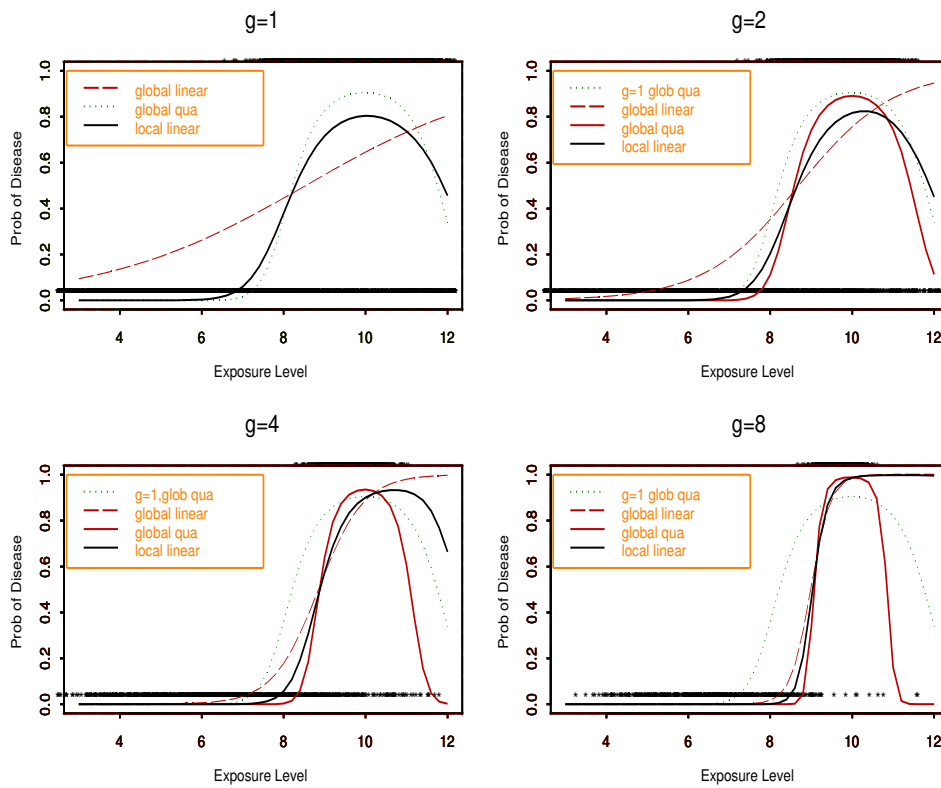


Figure 6.1: Comparison of Models for Different Pooling Scenarios

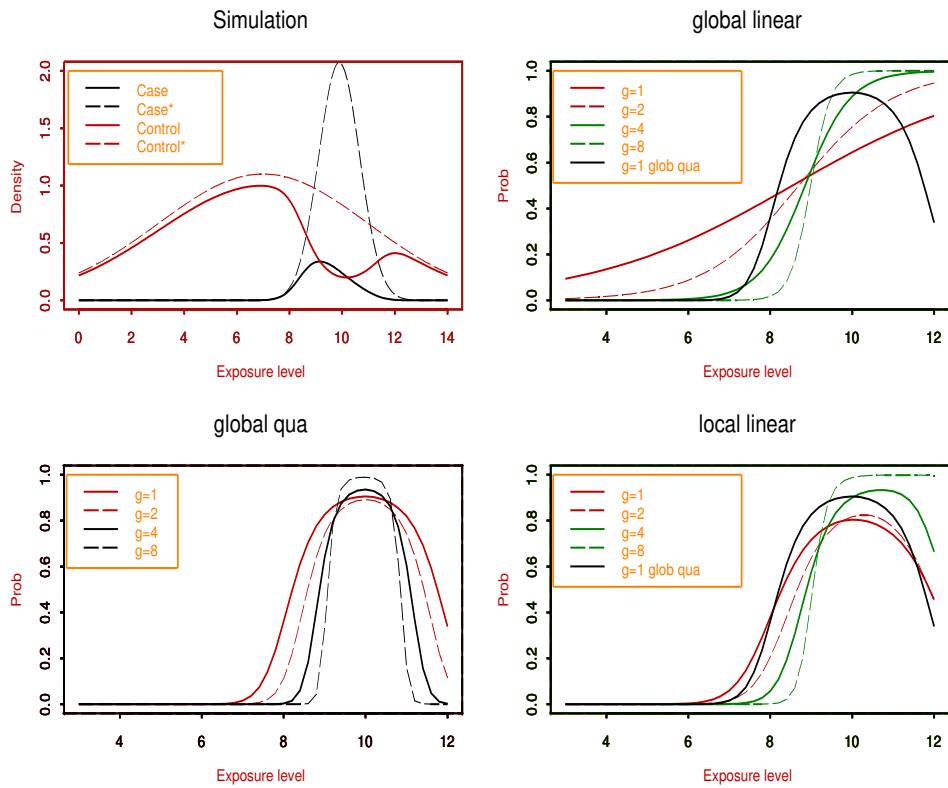


Figure 6.2: Evolution Effect of Different Pooling Scenarios on Models

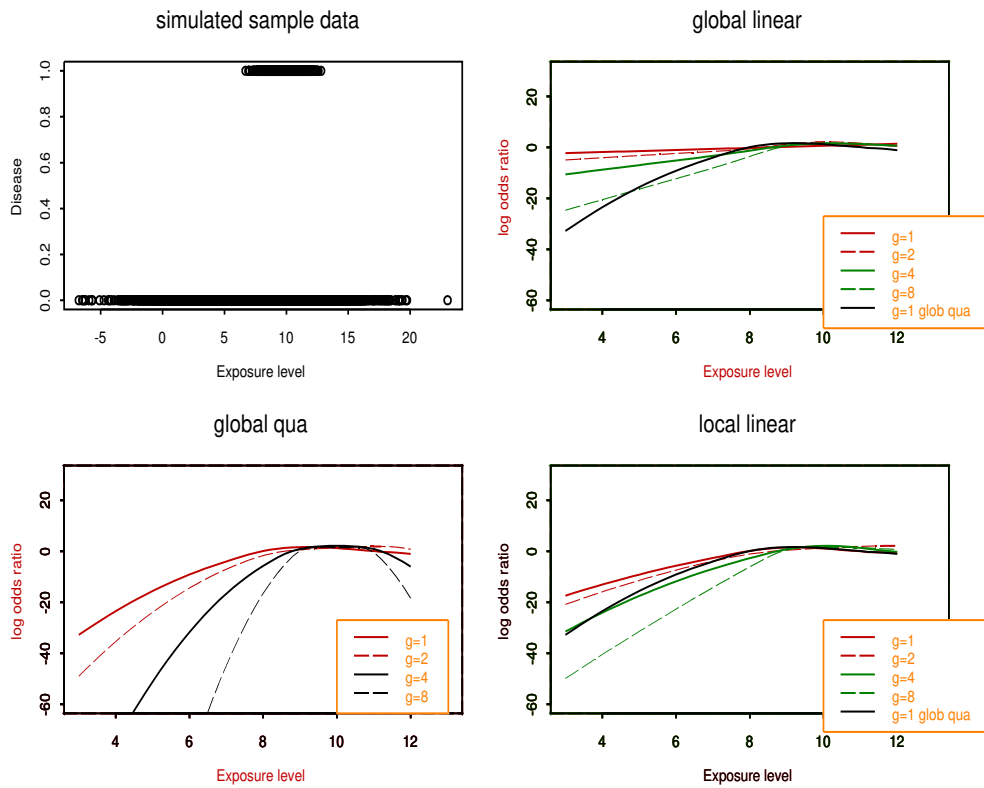


Figure 6.3: Evolution Effect of Different Pooling Scenarios on Models

adjusting other covariates, such as age, gender, waiting time, blood type, PRA, a categorical indicator for the level of resistance and the annual income are considered as potential factors. In order to properly investigate the effect of the pooled income, cases and controls are chosen in a way that the two types of patients are well separated. Case groups are the areas, indicated by zipcode, of which most of patients in the waiting list obtained the transplant. Control groups are chosen that most of its patients in the waiting list did not get transplanted. Deceased Donor Transplant are counted as cases.

We treat household income in zip code level as a pooled measurement from US CENSUS. We propose to fit the donor rate with regular logistic models, which is given by:

$$g(p) = \alpha + \boldsymbol{\gamma}^T \mathbf{Z} + \theta(E),$$

where $\boldsymbol{\gamma}^T = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7)^T$, $\mathbf{Z} = (\text{Time, Gender, PRA (0, 1-29, and 30-100), Blood type (A, AB, B, O)})$, and $\theta(E) = \beta_1 \log \text{Income}$.

Another logistic model 2 with quadratic effect of logarithm of income ($\beta_9 \log^2 \text{Income}$) is also fitted and the quadratic term is found to be significant, which is given by:

$$g(p) = \alpha + \boldsymbol{\gamma}^T \mathbf{Z} + \theta(E),$$

where $\boldsymbol{\gamma}^T = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7)^T$, $\mathbf{Z} = (\text{Time, Gender, PRA (0, 1-29, and 30-100), Blood type (A, AB, B, O)})$, and $\theta(E) = \beta_1 \log \text{Income} + \beta_2 \log^2 \text{Income}$.

6.8.3 Model results

The results from model 1 and model 2 are summarized in Table 6.8.3.

Para.	Model 1			Model 2		
	Est.	Std. Error	<i>t</i> value	Est.	Std. Error	<i>t</i> value
α	-21.043	1.042	-20.187	-129.843	18.223	-7.125
γ_1	1.747	0.052	33.508	1.755	0.052	33.529
γ_2	0.142	0.063	2.258	0.147	0.063	2.328
γ_3	-0.118	0.083	-1.422	-0.117	0.083	-1.412
γ_4	-1.330	0.083	-16.110	-1.338	0.083	-16.172
γ_5	0.366	0.173	2.113	0.407	0.174	2.342
γ_6	-0.895	0.092	-9.772	-0.893	0.092	-9.719
γ_7	-0.535	0.068	-7.825	-0.513	0.069	-7.471
β_1	0.871	0.084	10.427	21.151	3.386	6.246
β_2				-1.945	0.158	-5.997

Apparently the quadratic term in the model is statistically significant and is negative, which indicates that the probability of transplant will increase first and then decrease as the income increases, which was very similar to the Figure 2 by Axelrod David. This result is hard to explain for the income effect on the transplant. It may come because of the impact

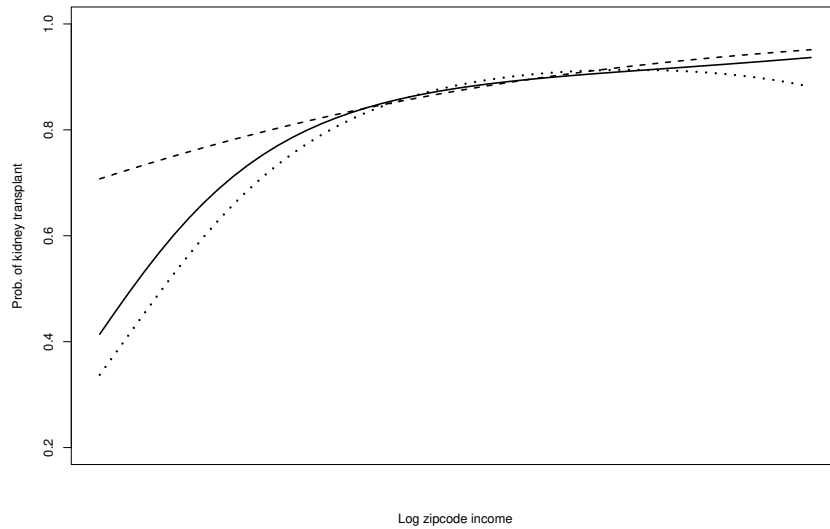


Figure 6.4: Probability of kidney transplant versus logarithm of income. Solid line represents the curve with local logistic regression, dashed line is the curve of regular logistic regression and the dotted line is the fitted logistic regression with quadratic term.

of pooling of covariate. In fitting the local regression model, we treat the effect combining all but intercept, logarithm of income and the squared log-income as an “offset” in the model, by observing the estimates of β_1 to β_7 have not much changes between the two models. This method not only makes the estimation of local fitting plausible but also makes the interpretation clearer. When our pooling approach as discussed in this paper was used to fit this data, the quadratic term in model 2 is totally removed, as shown in Figure 6.8.3.

From the result of the local logistic regression, the probability of kidney transplant decreased from low income to middle income and continued to decrease as the income increases. The parabolic curvature at the upper part from the graph of the logistic regression with the quadratic term of log-income disappears in the local fitting. The bottom part of the local fitting is similar to that of the logistic regression model with quadratic term. In other world, the probability of kidney transplant may increase quadratically when the average income increases from low to middle level. When the average income increases from middle level to high level, instead of decreasing, the probability of transplant increases linearly. We approximated the quadratic function based on individual probabilities obtained from the local logistic regression. The coefficients were given by -16.468 , 3.105 and -0.138 for the intercept, log-income and the quadratic term of log-income. Comparing with those of the regular logistic regression model with the quadratic term, the square term was apparently much smaller in magnitude, showing an insignificant quadratic effect.

6.9 Conclusion

The nonparametric method can provide a solution to unit measurement for complicated data sets. The main advantage of this proposed method is that it can overcome the difficulty in which Weinberg & Umbach's method encountered when dealing with data with higher order polynomial (non-linear, etc.) structures. Consequently, the exposures for case and control in their simulated data are extremely apart from each other with huge blank area between the two small ranges having observations as shown in the simulation study. It brings hardness for model fitting. We instead do not stick on the scenario. We generate data from an assumed model with a quadratic term.

We observe that the size of the pooled set will affect the resulting fitted nonparametric curve. The more data pooled into the set, the less data points left in the fitting data. Besides, we observe the pooled exposures for case and control are further away. As we increase the size of the pooled set, convergence of the algorithm may become a problem.

The nonparametric method is driven by the practical data. So it can be misleading if we can not decide which prediction values to choose for given exposure level. You might obtain quite different results from fitting with different sample. Although we do not consider any interactions in our current study, it will be our future work. Consider adding more covariates. If they are simply linear we can apply Weinberg & Umbach's method. But if there are more higher order terms or interactions we need to figure out the counterparting nonparametric solution, generalized additive model can be a choice.

Chapter 7

Future works

7.1 Motivation

This dissertation is mainly focus on the development of the novel and powerful statistical methodology to solve the problem in the complex real clinical data. All proposed statistical models are motivated and illustrated by the clinical transplant data. As mentioned in the introduction, there are three strategies to address the problem about the demand supply of organs is not sufficient to meet the increasing demand:

1. Decrease the incidence of ESRD
2. Increase the number of deceased and living organ donors
3. Maximize the utility of the available organ supply

How to realize these specific aims in each strategy motivate us develop the new models in this thesis. Then the computational and applied skills for these new models motivate me to continue the theoretically challenging research. The research undertaken in my thesis, together with future research plans, are described below.

7.2 Current work and future research

7.2.1 Functional data analysis (FDA)

Functional data analysis (FDA) has recently become a very hot topic in statistical research, as recent technological progress in measuring devices now allows one to observe spatiotemporal phenomena on arbitrarily.

As my senior supervisor told me that FDA remains distinct due to its contribution to climatology, medicine, meteorology, economics, etc. Characterizing nonlinear variation in FDA is a challenging problem. It provides the opportunity for statisticians to develop new methodologies to address it. In particular, when random curves are observed on regular dense grids, the existing literature on FDA focused on estimation and inference. This,

however, is not enough to provide understanding of the variability of the estimator of the whole regression curve and its derivative, nor can it be used to correctly answer questions about the curve shape.

In chapter 2, the proposed functional principal component analysis through conditional expectation by Dong⁹⁷ have explored the major source of variations of GFR curves. We find that the estimated functional principal component (FPC) scores can be used to cluster GFR curves. Ordering FPC scores can detect abnormal GFR curves. Finally, FPCA can effectively estimate missing GFR values and predict future GFR values. Chapter 4 by Dong⁹⁵ have developed a joint model, which uses functional principal component analysis (FPCA) to fit the longitudinal outcome and proposes the multi-state model to describe multiple time-to-event outcomes together. The FPCA method is efficient in reducing the dimension of the longitudinal trajectories. Multistate submodel can be used to describe the dynamic process of multiple time-to-event outcomes. The longitudinal trajectories and the multiple time-to-event outcomes is linked with the shared latent features. In our application example, the estimated functional principal components are found efficient with fitting the GFR curves, and the latent FPC scores are significantly related to the multiple time-to-event outcomes.

In the future, I will continue to develop new FPCA approach in terms of recovering trajectories from noisy functional data and dimension reduction in regression models of functional data. Then we can compare the current results from FPCA through conditional expectation with some new method.

7.2.2 Joint modeling

In this thesis, we have developed three joint models:

1. **Joint model 1 in Chapter 3:** The accelerated failure time submodel used in our proposed joint model. On the other hand, the proposed joint model is different from some traditional joint models, which treats the longitudinal component as a covariate in the survival analysis. In our proposed joint model, instead of using the whole longitudinal component as a covariate, we propose to use some latent features of the longitudinal component in the survival submodel. Finally, our proposed joint models has considered a method to obtain the dynamical non-proportional hazard ratio curve of a side event when hazard ratios are non-proportional during the followed-up time period. we propose a new joint model with a longitudinal submodel and an accelerated failure time (AFT) submodel, which are linked by some latent variables. The AFT submodel is used to determine the relationship of the time-to-event outcome with all predictors. In addition, the piecewise linear function in the survival submodel is used to calculate the dynamic hazard ratio curve of a time-dependent side event, because the effect of the side event on the time-to-event outcome is non-proportional. The model parameters are estimated with a Monte Carlo EM algorithm.

2. **Joint model 2 in Chapter 4:** This paper develops a joint model, which uses functional principal component analysis (FPCA) to fit the longitudinal outcome and proposes the multi-state model to describe multiple time-to event outcomes together. The FPCA method is efficient in reducing the dimension of the longitudinal trajectories. Multistate submodel can be used to describe the dynamic process of multiple time-to-event outcomes. The longitudinal trajectories and the multiple time-to-event outcomes is linked with the shared latent features. In our application example, the estimated functional principal components are found efficient with fitting the GFR curves, and the latent FPC scores are significantly related to the multiple time-to-event outcomes
3. **Joint model 3 in Chapter 5:** Our third proposed joint model use a flexible class of generalized linear latent variable models for multivariate responses, which has an underlying Gaussian latent processes. The model accommodates any mixture of outcomes from the exponential family. Monte Carlo EM algorithm is proposed for parameter estimation and estimates of the variance components of the latent processes

In the future, I will continue to develop new joint approaches.

7.2.3 Measurement error models

In this thesis, we have considered missing values/measurement errors in several models:

1. **The proposed FPCA in Chapter 2:** The proposed FPCA through conditional expectation for GFR in Chapter 1.
2. **The proposed nonparametric kernel approach in Chapter 6:** I have used this nonparametric kernel method to investigate the predict models with a high-order effect co-variable in present of measurement errors in the Chapter 5.
3. It is worthwhile to mention that we have develop a Bayesian approach to a calibration problem with one interested covariate subject to multiplicative measurement errors¹⁰⁶ before, where a stem cell study with the objective of establishing the recommended minimum doses for stem cell engraftment after a blood transplant. When determining a safe stem cell dose based on the prefreeze samples, the postcryopreservation recovery rate enters in the model as a multiplicative measurement error term, as shown in the model. We examine the impact of ignoring measurement errors in terms of asymptotic bias in the regression coefficient. According to the general structure of data available in practice, we propose a two-stage Bayesian method to perform model estimation via R2WinBUGS. We illustrate this method by the aforementioned motivating example. The results of this study allow routine peripheral blood stem cell processing laboratories to establish recommended minimum stem cell doses for transplant and develop a systematic approach for further deciding whether the postthaw analysis is warranted.

In the future, I will continue to develop new approaches in Measurement error models and Joint models with covariates in present of measurement errors.

7.2.4 Cost-effectiveness analysis

In our previous work^{102,107}, we constructed a Markov model in accordance with existing guidelines for economic evaluation. Base case analyses were performed with Markov cohort simulation with transitions modelled on an annual basis, though we used first order Monte Carlo simulation to determine the incidence of end stage renal disease over time. We considered several health states in both the screening and no screening strategies, including people without chronic kidney disease, those with non-dialysis chronic kidney disease, patients receiving dialysis, and patients with a functioning transplant.

In the future, I will continue to do the cost-effective analysis for modeling the following outcomes of: 1) combined multi-organ transplantation 2) non-renal organ transplantation +/- later kidney transplantation; or 3) kidney-alone transplantation. Three cost-effectiveness ratio per quality-adjusted life-year will be calculated when compared to non-transplantation treatment quality-adjusted life-year.

Bibliography

- [1] Zeger S.L, Karim M.R. Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association* 1991; **86**(413):79–86.
- [2] Schall R. Estimation in generalized linear models with random effects. *Biometrika* 1991; **78**(4):719–727.
- [3] Breslow N. E. , Clayton D. G. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**(421):9–25.
- [4] Berk K. N., Lachenbruch P. A. Repeated measures with zeros. *Statistical Methods in Medical Research* 2002; **11**:303–316.
- [5] Levey A., Bosch J., Lewis J., Greene T., Rogers N., Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine : a new prediction equation. *Annals of Internal Medicine* 1999; **130**(6):461–470.
- [6] Salvadori M., Rosati A., Bock A., Chapman J., Dussol B., Fritsche L., Kliem V., Lebranchu Y, Oppenheimer F, Pohanka E, Tufveson G, Bertoni E. Estimated one-year glomerular filtration rate is the best predictor of long-term graft function following renal transplant. *Transplantation* 2006; **81**(2):202–206.
- [7] Locatelli F., Vecchio L.D., Pozzoni P. The importance of early detection of chronic kidney disease. *Nephrology Dialysis Transplantation* 2002; **17**(Suppl 11):2–7.
- [8] Klahr S., Levey A.S., Beck G.J., Caggiula A.W., Hunsicker L., Kusek J.W., Striker G. The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease. *The New England Journal of Medicine* 1994; **330**(13):877–884.
- [9] Marcén R., Morales J.M., Fernández-Rodríguez A., Capdevila L., Pallardáó L., Plaza J.J., Cubero J.J., Puig J.M., Sanchez-Fructuoso A., Arias M., Alperovich G., Serón D. Long-term graft function changes in kidney transplant recipients. *Nephrology Dialysis Transplant* 2010; **3**[Suppl 2]:ii2–ii8.
- [10] Rao R.C. Some statistical methods for comparison of growth curves. *Biometrics* 1958; **14**(1):1–17.
- [11] Castro P.E., Lawton W., Sylvestre E. Principal modes of variation for processes with continuous sample curves. *Technometrics* 1986; **28**(4):329–337.
- [12] Dauxois J., Pousse A., Romain Y. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis* 1982; **12**(1):136–154.

- [13] Zhang J.T., Chen J. Statistical inferences for functional data. *The Annals of Statistics* 2007; **35**(3):1052–1079.
- [14] Benko M., Härdle W., Kneip A. Common functional principal components. *The Annals of Statistics* 2009; **37**(1):1–34.
- [15] Hall P., Hosseini-Nasab M. On properties of functional principal components analysis. *Journal of the Royal Statistical Society Series B* 2006; **68**(1):109–126.
- [16] Hall P., Hosseini-Nasab M. Theory for high-order bounds in functional principal components analysis. *Mathematical Proceedings of the Cambridge Philosophical Society* 2009; **146**(1):225–256.
- [17] Rice J.A., Silverman B.W. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society Series B* 1991; **53**:233–243.
- [18] Pezzulli S., Silverman B.W. Some properties of smoothed principal components analysis for functional data. *Computational Statistics* 1993; **8**(1):1–16.
- [19] Silverman B.W. Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* 1996; **24**(1):1–24.
- [20] Boente G., Fraiman R. Kernel-based functional principal components. *Statistics and Probability Letters* 2000; **48**(4):335–345.
- [21] James G.M., Hastie T.J., Sugar C.A. Principal component models for sparse functional data. *Biometrika* 2000; **87**(3):587–602.
- [22] Yao F., Muller H.G., Wang J.L. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005; **100**(470):577–590.
- [23] James G.M., Wang J., Zhu J. Functional linear regression that’s interpretable. *The Annals of Statistics* 2009; **37**(5A):2083–2108.
- [24] Tian T.S., James G.M. Interpretable dimensionality reduction for classification with functional data. *Computational Statistics and Data Analysis* 2013; **57**:282–296.
- [25] Lin Z., Wang L., Cao J. Interpretable functional principal component analysis. *Biometrics* 2016; **72**:846–854.
- [26] Feng C.X., Cao J., Bendell L. Exploring spatial and temporal variations of cadmium concentrations in pacific oysters from British Columbia. *Biometrics* 2011; **67**(3):1142–1152.
- [27] Luo W., Cao J., Gallagher M., Wiles J. Estimating the intensity of ward admission and its effect on emergency department access block. *Statistics in Medicine* 2013; **32**(15):2681–2694.
- [28] Ramsay J.O., Silverman B.W. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2nd ed., 2005.
- [29] MATLAB R2017a. The MathWorks Inc., Natick, MA, 2017.

- [30] Fan J. and Gijbels I. *Local Polynomial Modelling and its Applications*. CRC Press, 1996.
- [31] Shibata R. An optimal selection of regression variables. *Biometrika* 1981; **68**(1):45–54.
- [32] D’Agostino Sr R.B. and Russell H.K. *Encyclopedia of Biostatistics*. John Wiley, 2005.
- [33] Forgy E.W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrika* 1965; **21**(3):768–769.
- [34] Hartigan J., Wong M.A. Algorithm as 136: A k -means clustering algorithm. *Journal of the Royal Statistical Society Series C* 1979; **28**(1):100–108.
- [35] Rousseeum P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987; **20**(20):53–65.
- [36] Filzmoser P., Maronna R., Werner M. Outlier identification in high dimensions. *Computational Statistics and Data Analysis* 2008; **52**(3):1694–1711.
- [37] Fraley C. and Raftery A.E. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 2002; **97**:611–631.
- [38] Vinh X.N., Epps J., and Bailey J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research* 2010; **11**:2837–2854.
- [39] Wolfe, RA, and Ashby, VB, and Milford, EL, and Ojo, AO, and Ettenger, RE, and Agodoa, LY, and Held, PJ, and Port, FK. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of first cadaveric transplant. *New England Journal Medicine* 1999; **341**(23):1725-1730.
- [40] Knoll, G, and Nichol, G. Dialysis, kidney transplantation, or pancreas transplantation for patients with diabetes mellitus and renal failure: a decision analysis of treatment options. *Journal of The American Society of Nephrology* 2003; **14**(2):500-515.
- [41] Orsenigo, E, and Fiorina, P, and Cristallo, M, and Socci, C, and La, Rocca, E, and Maffi, P, and Invernizzi, L, and Zuber, V, and Secchi, A, and Di Carlo, V. Long-term survival after kidney and kidney-pancreas transplantation in diabetic patients. *Transplantation Proceedings* 2004; **36**(4):1072-1075.
- [42] Kleinclauss, F, and Fauda, M, and Sutherland, DER, and Kleinclauss, C, and Gruessner, RW, and Matas, AJ, and Kasiske, BL, and Humar, A, and Kandaswamy, R, and Kaul, S, and Gruessner, AC. Pancreas after living donor kidney transplants in diabetic patients: Impact on long-term Kidney graft function. *Clin Transplant* 2009; **23**:437-446.
- [43] Waki,K, and Terasaki,PI, and Kadowaki,T. Long-Term Pancreas Allograft Survival in Simultaneous Pancreas-Kidney Transplantation by Era. *Diabetes Care* 2010; **33**(8):1789-1791.
- [44] Poommipanit, N, and Sampaio, MS, and Cho, Y, and Young, B, and Shah, T, and Pham, PT, and Wilkinson, A, and Danovitch, G, and Bunnapradist, S. Pancreas after living donor kidney versus simultaneous pancreas-kidney transplant:An analysis of the

- Organ Procurement Transplant Network/United Network of Organ Sharing Database. *Transplantation* 2010. **89**(53):1496-1503.
- [45] Pavlakis, M, and Khwaja, K, and Mandelbrot, D, and Tang, H, and Whiting, JW, and Lorber, MI, and Gautam, A, and Johnson, SR, and Uknis, ME. Renal allograft failure predictors after PAK transplantation: Results from the New England Collaborative Association of Pancreas Program. *Journal of Transplantation* 2010; **89**(11):1347-1353.
- [46] Laird, N, and Ware, J. Random-Effects Models for Longitudinal Data. *Biometrics* 1982; **38**(4):963-974.
- [47] Pocock, S, and Geller, N, and Tsiatis, A. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**(3):487-498.
- [48] Tsiatis, AA, and DeGruttola, V, and Wulfsohn, MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of American Statistical Association* 1995; **90**(429):27-37.
- [49] Wulfsohn, MS, and Tsiatis, AA. A joint model for survival and longitudinal data measured with error. *Biometrics* 1997; **53**(1):330-339.
- [50] Wu L, and Hu X J, and Wu H. Joint inference for nonlinear mixed-effects models and time to event at the presence of missing data. *Biostatistics* 2007; **9**(2): 308-320.
- [51] Gueorguieva, R, and Agresti, A. A correlated probit model for joint modeling of clustered binary and continuous responses. *Journal of the American Statistical Association* 2001. **96**(455):1102–1112.
- [52] Tsiatis, AA, and Davidian, M. Joint modeling of longitudinal and time-to-event data:an overview. *Statistica Sinica* 2004; **14**(3):809-834.
- [53] Wu, L, and Liu, W, and Hu, X.J. Joint Inference on HIV Viral Dynamics and Immune Suppression in Presence of Measurement Errors. *Biometrics* 2010; **66**(2):327-335.
- [54] Tseng, Y, and Hsieh, F, and Wang J. Joint modelling of accelerated failure time and longitudinal data. *Biometrics* 2005; **92**(3):587-603.
- [55] Song, P, and Li, M, and Yuan, Y. Joint regression analysis of correlated data using gaussian copulas. *Biometrics* 2009; **65**(1):60-68.
- [56] Rizopoulos D. fitting of joint models for longitudinal and event time data using a pseudoadaptive Gaussian quadrature rule. *Comput Stat Data Anal* 2011; **56**:2061–2077.
- [57] Fraley C, and Raftery AE Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 2002; **97**:611–631.
- [58] Pham Thi Thu Huong, Darfiana Nur, Hoa Pham and Alan Branford A modified two-stage approach for joint modelling of longitudinal and time-to-event data. *Journal of Statistical Computation and Simulation* 2018; **88**:3379–3398.
- [59] Rizopoulos, D, and Hatfield, L, and Carlin, B, and Takkenberg, J. Combining Dynamic Predictions From Joint Models for Longitudinal and Time-to-Event Data Using Bayesian Model Averaging. *Journal of the American Statistical Association* 2014; **109**(508):1385-1397.

- [60] Cox, D. Regression models and life tables. *Journal of Royal Statistics Society. Series B* 1972; **34** (2):187-220.
- [61] Cox, D, and Cakes, D. Analysis of survival data. *Chapman & Hall/CRC Press, London* 1984.
- [62] James, R. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* 1992; **79**(2):321–334.
- [63] Reid, N. A conversation with Sir David Cox. *Statistical Science* 1994; **9**(3):439–455.
- [64] Kay, R, and Kinnersley, N (2002) On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: A case study in influenza. *Drug Information Journal* 2002; **36**(3):571-579.
- [65] Levey, AS, and Stevens, LA, and Schmid, CH, and Zhang, YL, and Castro, AF, and Feldman, HI, and Kusek, JW, and Eggers, P, and Van, L, and Greene, T, and Coresh, J. A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine* 2009; **150** (9):604-612.
- [66] Efron, B, and Tibshirani, RJ An introduction to the bootstrap. CRC press 1994.
- [67] Wei, G, and Tanner, M A Monte Carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 1990; **85** (411):699-704.
- [68] Mordecai, A Nonlinear Programming: Analysis and Methods. *Dover Publishing* 2003
- [69] Joe, H Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics and Data Analysis* 2008; **52**:5066-5074.
- [70] Rue, H, Martino, S, and Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71**:319-392.
- [71] Luo S. A Bayesian approach to joint analysis of multivariate longitudinal data and parametric accelerated failure time. *Statistics in Medicine* 2014; **33** (4):580-594.
- [72] Dempster, AP, and Laird, NM, and Rubin, DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)* 1977; 1-38.
- [73] Geyer, C, and Thompson, E. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1992; **54**(3):657–699.
- [74] Hogan, JW, and Laird, NW. Model-based approaches to analyzing incomplete longitudinal and failure time data. *Statistics in Medicine* 1997; **11**:239-257.
- [75] Levey AS, and Stevens LA, and Schmid CH, and Zhang YL, and Castro AF, and Feldman HI, and Kusek JW, and Eggers P, and Van Lente, and Greene T, and Coresh J. A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine* 150(2009); (9):604–612.

- [76] Wolfe RA, and Ashby VB, and Milford EL, and Ojo AO, and Ettenger RE, and Agodoa LY, and Held PJ, and Port FK. Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of first cadaveric transplant. *New England Journal of Medicine* 341(1999); (23):1725–1730.
- [77] Marc'n R, and Morales JM, and Fernández-Rodríguez A, and Capdevila L, and Pallardø' L, and Plaza JJ, and Cubero JJ, and Puig JM, and Sanchez-Fructuoso A, and Arias M, and Alperovich G, and Serø'n D. Long-term graft function changes in kidney transplant recipients. *Nephrology Dialysis Transplantation* 3(2010); (9):ii2–ii8.
- [78] Moranne O, and Maillardb N, and Fafina C, and Thibaudinb L, and Alamartineb E, and Mariatb C Rate of Renal Graft Function Decline After One Year Is a Strong Predictor of All-Cause Mortality. *American Journal of Transplantation* (2013); (13):695–706.
- [79] Dong J, and Wang S, and Wang L, and Gill J, and Cao J. Joint modelling for organ transplantation outcomes for patients with diabetes and the end-stage renal disease. *Statistical Methods in Medical Research* (2018); (<https://doi.org/10.1177/0962280218786980>).
- [80] Dong J, and Wang L, and Gill J, and Cao J. Functional Principal Component Analysis of GFR Curves after Kidney Transplant. *Statistical Methods in Medical Research* (2017); (0):1–12.
- [81] James GM., Hastie TJ., Sugar CA. Principal component models for sparse functional data. *Biometrika*(2000);87,3,pp.587–602.
- [82] Fan J. and Gijbels I. *Local Polynomial Modelling and its Applications*. CRC Press, 1996.
- [83] Prentice R, KalbfñČeisch J, Peterson A, Fournoy N, Farewell V, Breslow N *The analysis of failure times in the presence of competing risks*. *Biometrika*(1978);34,pp.541–554.
- [84] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 26(2007); (17):2389–2430.
- [85] Yao F, Muller HG, Wang JL. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005; **100**(470):577–590.
- [86] Yao F. Functional principal component analysis for longitudinal and survival data. *Statistica Sinica* 17(2007); (17):965–983.
- [87] Ding J, Wang J. Modeling Longitudinal Data with Nonparametric Multiplicative Random Effects Jointly with Survival Data. *Biometrics* 2007; **64**:546–556.
- [88] Fine JP., Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association* 1999; **94**(446):496–509.
- [89] Wu L, Liu W, Hu XJ. Joint Inference on HIV Viral Dynamics and Immune Suppression in Presence of Measurement Errors. *Biometrics* 2010; **66**:327–335.
- [90] Logan B, Zhang M, Klein JP. Marginal models for clustered time to event data with competing risks using pseudovalues. *Biometrics* 2011; **67**:1–7.

- [91] Fraley C, Raftery AE Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 2002; **97**:611–631.
- [92] Gueorguieva, Ralitzka, and Agresti, Alan. A correlated probit model for joint modeling of clustered binary and continuous responses, *Journal of the American Statistical Association* 2001; **96**[455]:1102–1112.
- [93] David B, Dunson. Bayesian Latent Variable Models for Clustered Mixed Outcomes, *Journal of the Royal Statistical Society* 2000; **62**[2]: 355–366.
- [94] Mary Dupuis Sammel , and Louise M. Ryan , and Julie M. Legler Latent Variable Models for Mixed Discrete and Continuous Outcomes *Journal of the Royal Statistical Society* 1997;**59**[3]: 667–678.
- [95] **Dong J**, and Wang L, and Gill J, and Cao J. Jointly Modelling Multiple Time-to-Event Outcomes by Functional Principal Component Analysis via a Multistate Model. *Submitted*.
- [96] **Dong J**, and Wang S, and Wang L, and Gill J, and Cao J. Joint modelling for organ transplantation outcomes for patients with diabetes and the end-stage renal disease. *Statistical Methods in Medical Research* (2018); (<https://doi.org/10.1177/0962280218786980>).
- [97] **Dong J**, and Wang L, and Gill J, and Cao J. Functional Principal Component Analysis of GFR Curves after Kidney Transplant. *Statistical Methods in Medical Research* (2017); (0):1–12.
- [98] Gill J, **Dong J**, Rose C, Gill JS. The risk of allograft failure and the survival benefit of kidney transplantation are complicated by delayed graft function. *Kidney Int* (2016); **89**(6):1331–1336.
- [99] Gill J, **Dong J**, Gill JS. Population income and longitudinal trends in living kidney donation in the United States. *J Am Soc Nephrology* (2015); **26**(1):201–207.
- [100] Elizabeth L, Gill J, Landsberg D, **Dong J**, Rose C, Gill JS Willingness of Directed Living Donors and their Recipients to Participate in Kidney Paired Donation Programs. *Transplantation* (2015); **99**(9):1894–1899.
- [101] Gill J, **Dong J**, Eng M, Landsberg D, Gill JS. Pulsatile Perfusion Reduces the Risk of Delayed Graft Function in Deceased Donor Kidney Transplants, Irrespective of Donor Type and Cold Ischemic Time. *Transplantation* (2014); **97**(6):668–674.
- [102] Chui BK, Manns B, Pannu N, **Dong J**, Wiebe N, Tonelli M, and Klarenbach S Health care costs of peritoneal dialysis technique failure and dialysis modality switching. *American Journal Kidney Disease* (2013); **61**(1):104–111.
- [103] Gill J, **Dong J**, Rose C, Johnston O, Landsberg D, and Gill J Race and Income Related Differences in the Likelihood of Living Kidney Donation in the United States. *Journal of the American Society of Nephrology* (2013); **24**(11):1872–1879.
- [104] Luo D, Zhang P and **Dong J** Random Mean Models for Mixed Continuous and Discrete Clustered Outcomes. *Annual International conference* (2012);

- [105] Manns B, Hemmelgarn B, Tonelli M, Au F, Chiasson TC, **Dong J**, and Klarenbach S. The Impact of Pancreas Transplantation on Kidney Allograft Survival. *Transplantation* (2011); (11):1951–1958.
- [106] Zhang P, Liu J, **Dong J**, Holovati JL, Letcher B, and McGann LE A Bayesian adjustment for multiplicative measurement errors for a calibration problem with application to a stem cell study. *Biometrics* (2012); **68**(1):268–2748.
- [107] Manns B, Hemmelgarn B, Tonelli M, Au F, Chiasson TC, **Dong J**, and Klarenbach S. Population-Based Screening for Chronic Kidney Disease: Cost Effectiveness Study. *British Medicine Journal* (2010); (0): .

Appendix A

Supplementary material for Functional Principal Component Analysis of GFR Curves after Kidney Transplant

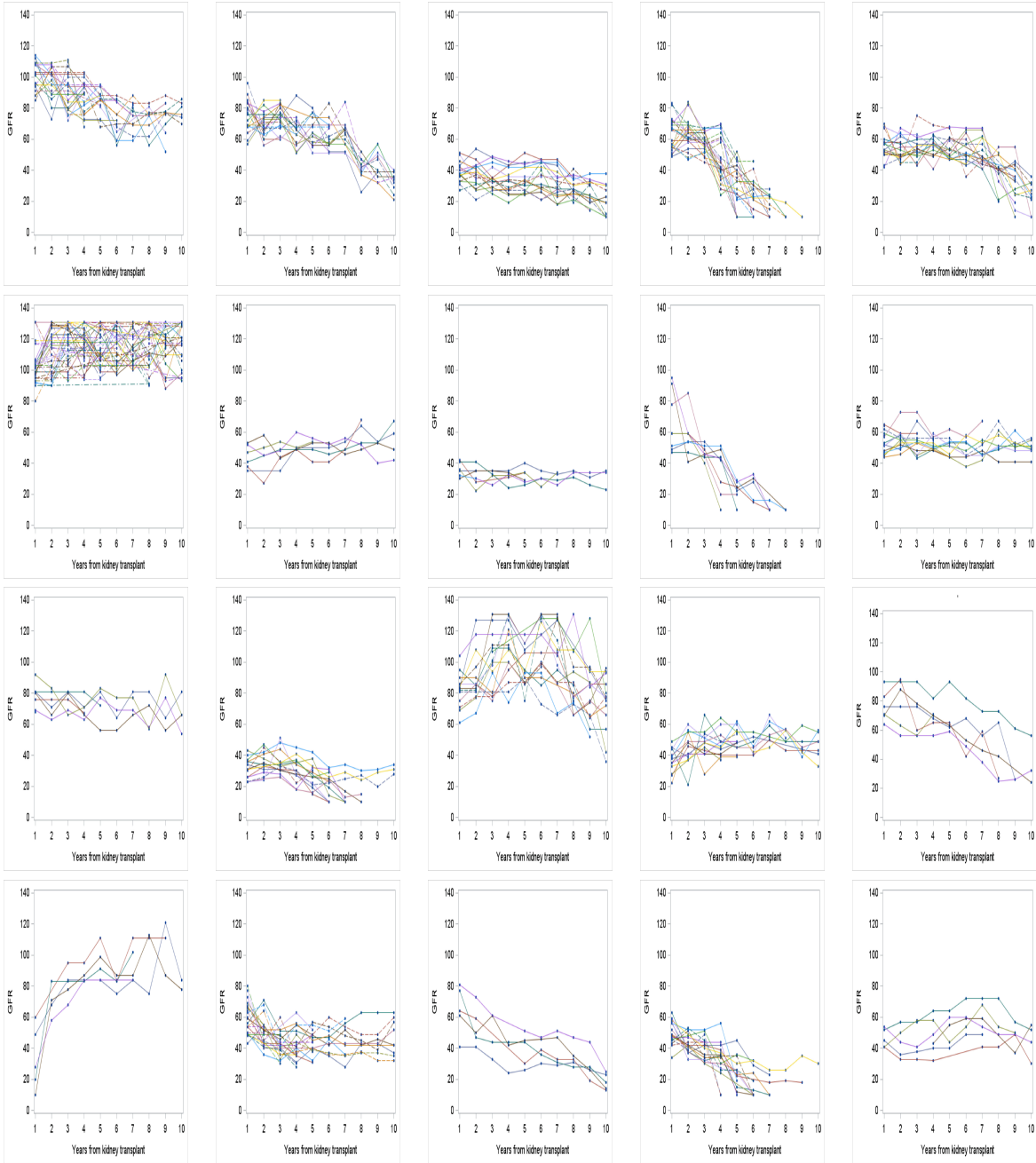


Figure A.1: Part of GFR trajectory curves for the first 20 cluster groups

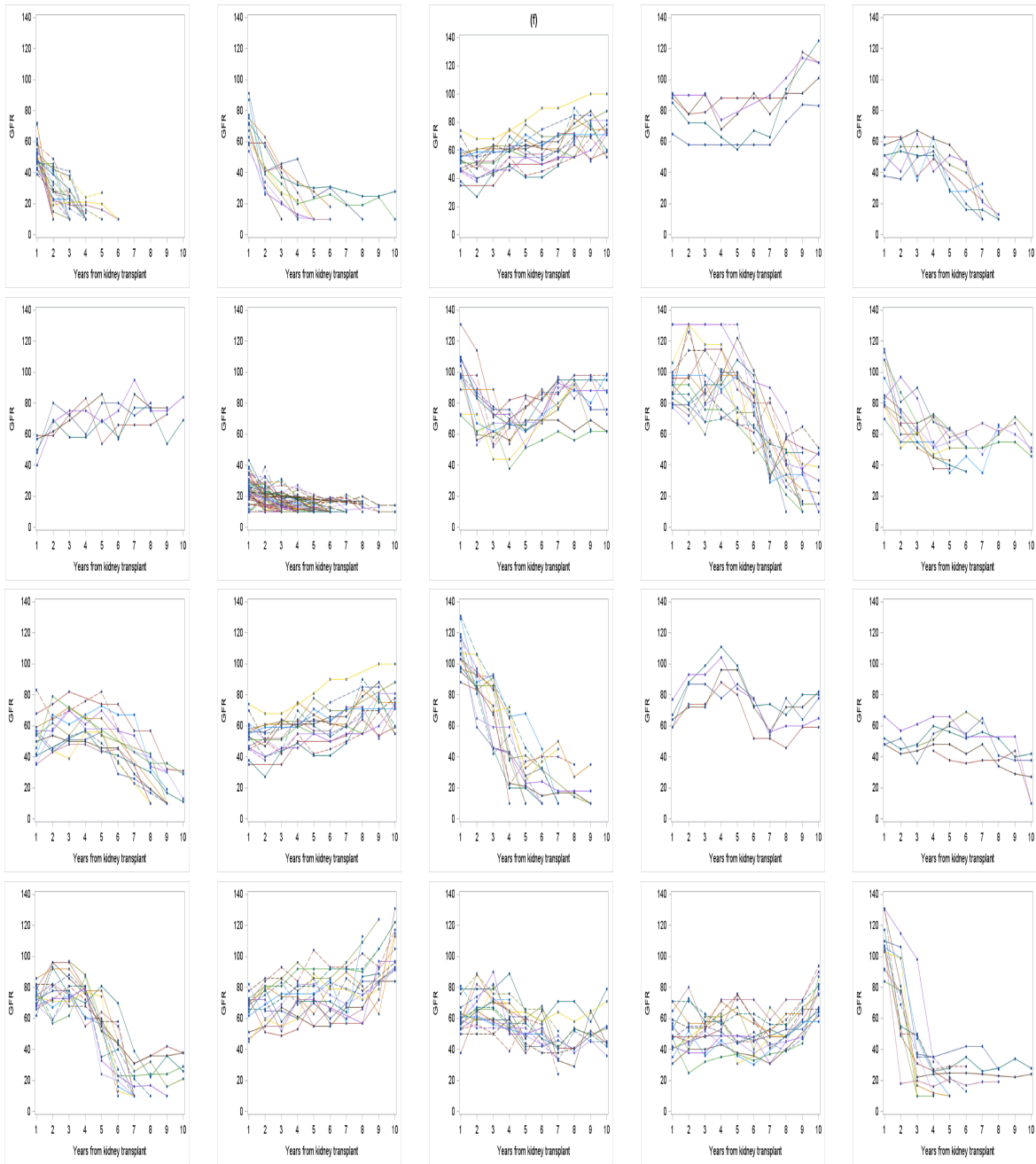


Figure A.2: Part of GFR trajectory curves for the last 20 cluster groups

Appendix B

Supplementary material for A Joint model of a longitudinal and Accelerated Failure Time data and its application to transplant patients with an ESRD and a diabetes

B.1 Monte Carlo EM algorithm

B.1.1 M-step

To make the notation short, let $E^{(t)}(g(\beta_i)) = E[g(\beta_i)|t, \mathbf{w}(t), S, \delta, \mathbf{Z}, Y(t), \Theta^{(t)}]$ be the conditional log likelihood based on the current estimate $\Theta^{(t)}$ for any function $g(\beta_i)$. The MLE of \mathbf{b} , \mathbf{B} , $\boldsymbol{\alpha}$, and σ^2 can be written as

$$\begin{aligned}\hat{\mathbf{b}} &= \sum_{i=1}^n E^{(t)}(\beta_i) \\ \hat{\mathbf{B}} &= \sum_{i=1}^n E^{(t)}(\beta_i - \hat{\mathbf{b}})(\beta_i - \hat{\mathbf{b}}) \\ \hat{\boldsymbol{\alpha}} &= \sum_{i=1}^n E^{(t)}((\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{Y}_i - \beta_i^T \boldsymbol{\xi}(t_i))) \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} E^{(t)}(Y_{ij} - \hat{\boldsymbol{\alpha}}^T \mathbf{Z} - \beta_i^T \boldsymbol{\xi}(t_{ij}))^2}{\sum_{i=1}^n m_i},\end{aligned}$$

where $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im_i})^T$, and $\mathbf{t}_i = (t_{i1}, \dots, t_{im_i})^T$.

We estimate the baseline hazard function λ_0 by a step-function. Let T_1, \dots, T_H be all observed event times, then the baseline failure time is

$$\Phi(T_h, Z_i, \mathbf{w}_i(t_h), \beta_i, \gamma) = \int_0^{T_h} \exp[\gamma_1^T \mathbf{Z}_i + \gamma_2^T \beta_i + \mathbf{w}_i(s|\gamma_3)] ds,$$

where $h = 1, \dots, H$. Let $\mu_h = \Phi(t_h, Z_i, \mathbf{w}_i(t_h), \beta_i, \gamma)$, we estimate μ_h by plugging in the current estimate of β_i and $\gamma_i^T = (\gamma_1^T, \gamma_2^T, \gamma_3^T)$. We get $0 = \hat{\mu}_{(0)} \leq \hat{\mu}_{(1)} \leq \dots \leq \hat{\mu}_{(H)}$ by ordering these estimate in the data. Then the baseline function can be specified as $\lambda_0(\mu) = \sum_{h=1}^H C_h \mathbf{1}_{\{\hat{\mu}_{(h-1)} < \mu \leq \hat{\mu}_{(h)}\}}$. Now let the derivative of $E^{(t)}(l(\Theta))$ w.r.t C_h be equal to zero, then we obtain the maximum likelihood estimate for C_h :

$$\hat{C}_h = \frac{\sum_{i=1}^n E_i^{(t)} [\delta_i \mathbf{1}_{\{\hat{\mu}_{(h-1)} < \mu_i \leq \hat{\mu}_{(h)}\}}]}{\sum_{i=1}^n E_i^{(t)} [\{\hat{\mu}_{(h)} - \hat{\mu}_{(h-1)}\} \mathbf{1}_{\{\hat{\mu}_{(h)} \leq \mu_i\}}]}.$$

If we insert the baseline hazard function $\hat{\lambda}_0(\mu)$ into the conditional log likelihood, then we have

$$\begin{aligned} Q(\Theta|\Theta^{(t)}) &= \sum_{i=1}^n E^{(t)} \left[\delta_i \log \left\{ \sum_{h=1}^H \hat{C}_h \mathbf{1}_{\{\hat{\mu}_{(h-1)} < \mu_i \leq \hat{\mu}_{(h)}\}} \right\} + \delta_i (\gamma_1^T \mathbf{Z} + \gamma_2^T \beta_i + \right. \\ &\quad \left. \mathbf{w}(t|\gamma_3)) - \sum_{h=1}^H \hat{C}_h \{\hat{\mu}_{(h)} - \hat{\mu}_{(h-1)}\} \mathbf{1}_{\{\hat{\mu}_{(h)} \leq \mu_i\}} + \right. \\ &\quad \left. \sum_{j=1}^{m_i} \log f(Y_{ij}|\beta_i, \alpha, \sigma^2) + \log f(\beta_i|\mathbf{b}, \mathbf{B}) \right]. \end{aligned}$$

After we have obtained the estimate for the parameters $\mathbf{b}, \mathbf{B}, \alpha, \sigma^2$, and the baseline hazard function $\hat{\lambda}_0(t)$, the last parameter to estimate is γ . The estimate for γ has no closed-form. So we use the numeric optimization algorithm such as *optim()* in R to estimate γ in the M-step.

B.2 The result from simulation 1 when N=500

Table B.1: Mean, bias, RMSE of the parameter estimates using our proposed MCEM algorithm for Model using 100 simulation replicates in the first simulation study ($N = 500$).

Parameters	The longitudinal submodel				The survival submodel			
	True	Mean	Bias	RMSE	True	Mean	Bias	RMSE
Age (per year)	-0.17	-0.17	-0.00	0.011	0.02	0.02	-0.00	0.001
Female	5.39	5.36	0.03	0.187	-0.23	-0.23	0.00	0.010
Black	-3.69	-3.70	0.01	0.208	0.17	0.17	-0.00	0.009
Other	-6.45	-6.45	0.00	0.195	0.23	0.23	-0.00	0.010
TX era 1993 – 1997	7.56	7.58	-0.02	0.198	-0.29	-0.29	0.00	0.010
TX era 1998 – 2002	10.75	10.72	0.03	0.211	-0.76	-0.76	-0.00	0.010
TX era 2003 – 2007	16.52	16.56	-0.04	0.208	-0.95	-0.95	0.00	0.010
PKPRA 1 – 29	-0.93	-0.93	0.00	0.054	0.06	0.06	-0.00	0.003
PKPRA 30 – 100	-2.35	-2.39	0.04	0.155	0.27	0.27	0.00	0.010
HLA Mismatch 1 – 6	-1.54	-1.54	-0.00	0.056	0.24	0.24	-0.00	0.011
Dialysis time 0.1 – 1 years	-0.27	-0.27	0.00	0.062	0.07	0.07	-0.00	0.005
Dialysis time 1.1 – 2 years	-0.52	-0.52	0.00	0.052	0.08	0.08	0.00	0.005
Dialysis time 2.1 – 3	-0.67	-0.67	0.00	0.059	0.33	0.33	0.00	0.005
Dialysis time > 3 years	-0.96	-0.95	-0.01	0.067	0.38	0.38	-0.00	0.005
Decreased Donor	-1.15	-1.16	0.01	0.056	0.14	0.14	-0.00	0.005
β_1	48.94	49.51	-0.57	1.074				
β_2	-1.36	-1.50	0.14	0.455				
γ_{21}					-0.07	-0.07	-0.00	0.001
γ_{22}					-0.21	-0.21	-0.00	0.005
γ_3					-0.13	-0.13	-0.00	0.005