UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Metabolomic Markers and Functional Data Methods for Characterizing and Predicting Diabetic Kidney Disease Progression

Permalink

https://escholarship.org/uc/item/3rn8w1ww

Author

Kwan, Brian

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Metabolomic Markers and Functional Data Methods for Characterizing and Predicting Diabetic Kidney Disease Progression

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biostatistics

by

Brian Kwan

Committee in charge:

Professor Loki Natarajan, Chair
Professor Cheryl Anderson
Professor Lin Liu
Professor Karen Messer
Professor David Strong

2021

Copyright
Brian Kwan, 2021
All rights reserved.

The dissertation of Brian Kwan is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents Chuck and Jun,

thank you for all your love, and support throughout my life's journey. This work would not have been possible without your encouragement.

EPIGRAPH

Do not go where the path may lead, go instead where there is no path and leave a trail.

—Ralph Waldo Emerson

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am extremely grateful for all the guidance and support from the faculty, staff, students, friends, and family that have shaped my academic, professional, and personal growth throughout my graduate studies. Without them, this dissertation would not have been possible.

I would like to first acknowledge my research advisor, Loki Natarajan. Thank you for believing in my success in the Ph.D. Biostatistics program as part of its first class of students. I feel blessed everyday as a Ph.D. student knowing I get to work with a kind, understanding, and patient advisor on impactful research projects. You set a character example that I follow whenever life throws me curveballs. The committee members of my dissertation have contributed much to furthering my goals and interests. I would like to thank Karen Messer for always willing to provide valuable insight in my research directions and for surrounding me with warm colleagues to work alongside with in the Moores Cancer Center. Thank you, Lin Liu, for your encouraging and frank advice on developing professionally as a junior researcher in the academia setting. Thank you, David Strong, for encouraging my collaboration and communication with public health doctoral students in bringing the two Ph.D. programs closer together. Thank you, Cheryl Anderson, for demonstrating through your leadership the wide variety of impacts that individuals involved in public health could have.

The co-authors of the papers in my dissertation, which include my research advisor and committee members, made this work possible. Thank you to H. Irene Su for helping me foster my research interests and skillset outside of my dissertation research. Thank you to Kumar Sharma, Tobias Fuhrer, and Daniel Montemayor for collaborating with and advising me on my statistical applications to metabolomics and diabetic kidney disease. Thank you to all the CRIC co-authors for your faith and involvement with these papers.

I owe the faculty and students of the Ph.D. Biostatistics program for my development as an academic scholar. Thank you to Armin Schwartzman, Florin Vaida, Sonia Jain, Xin Tu, and Ronghui Xu, Charles Berry, Wesley Thompson, and Steven Edland for their valuable biostatistics instruction and advice in both the classroom and research environments. Thank you to Xinlian Zhang and Jingjing Zou for providing frank and earnest advice on developing as a junior scholar. My PhD studies would not be complete without the companionship of fellow students, particularly those in the first two cohorts: Anya Umlauf, Yuqi Qiu, Lingjing Jiang, Ruifeng Chen, Jinyuan Liu, Kristen Hansen, and Wenyi Lin.

It would be impossible for me to succeed as a Ph.D. student without the administrative and research support of key individuals. Huge thank you to Melody Bazyar, Sarah Dauchez, and Stella Tripp for assisting and advising on my student needs over the years. Thank you to Alan Larson for always kindly assisting with any study environment resources I may need. Thank you, Emily Pittman, Jing Zhang, and Minya Pu, for helping me with my research needs and for making an affable office environment for biostatistics. Thank you, Lisa Wesby, for assisting with my collaboration with members of the CRIC study.

Last, but not least, I would like to acknowledge and thank my family and friends. For without their love and care, I would not be where I am today.

Chapter 2, in part, has been submitted for publication of the material as it may appear in Kwan, Brian; Liu, Lin; Strong, David; Su, H. Irene; Natarajan, Loki. “Linear mixed model vs two-stage methods: Developing prognostic models of diabetic kidney disease progression”. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part is currently being prepared for submission for publication of the material. Kwan, Brian; Fink, Jeff; Fuhrer, Tobias; He, Jiang; Hsu, Chi-yuan; Messer, Karen; Montemayor, Daniel; Nelson, Robert; Pu, Minya; Ricardo, Ana; Rincon-Choles,

Hernan; Shah, Vallabh; Ye, Hongping; Zhang, Jing; Sharma, Kumar; Natarajan, Loki. "Identifying metabolite-pair markers for chronic kidney disease stage classification in diabetic patients: results from applying the top-scoring pairs algorithm to the Chronic Renal Insufficiency Cohort (CRIC) Study". The dissertation author was the primary investigator and author of this material.

Chapter 4, in part is currently being prepared for submission for publication of the material. Kwan, Brian; Anderson, Amanda; Anderson, Cheryl; Chen, Jing; Fuhrer, Tobias; Montemayor, Daniel; Ricardo, Ana; Rosas, Sylvia; Yang, Wei; Zhang, Jing; Natarajan, Loki. "Inference for Functional Principal Components of Kidney Disease Progression in Diabetic Patients Across Albuminuria Groups in the Chronic Renal Insufficiency Cohort (CRIC) Study". The dissertation author was the primary investigator and author of this material.

VITA

2015	Bachelor of Science in Mathematics University of California, Irvine
2021	Doctor of Philosophy in Biostatistics University of California San Diego

PUBLICATIONS

Kim J, Whitcomb BW, **Kwan B**, Zava D, Sluss PM, Dietz A, Shliakhtsitsava K, Romero SAD, Natarajan L, Su HI. Psychosocial stress and ovarian function in adolescent and young adult cancer survivors. *Human Reproduction*. In press. DOI: 10.1093/humrep/deaa313.

Kwan B, Fuhrer T, Zhang J, Darshi M, Van Espen B, Montemayor D, de Boer IH, Dobre M, Hsu C, Kelly TN, Raj DS, Rao PS, Saraf SL, Scialla J, Waikar SS, Sharma K, and Natarajan L; on behalf of the CRIC Study Investigators. Metabolomic Markers of Kidney Function Decline in Patients With Diabetes: Evidence From the Chronic Renal Insufficiency Cohort (CRIC) Study. *American Journal of Kidney Diseases*. 2020; 76(4):511-520. DOI: 10.1053/j.ajkd.2020.01.019.

Su HI, **Kwan B**, Whitcomb BW, Shliakhsitsava K, Dietz AC, Stark SS, Martinez E, Sluss PM, Sammel MD, and Natarajan L. Modeling variation in the reproductive lifespan of female adolescent and young adult cancer survivors using AMH. *The Journal of Clinical Endocrinology & Metabolism*. 2020; 105(8):dgaa172. DOI: 10.1210/clinem/dgaa172.

Gauglitz JM, et al. [System-Wide Mass Spectrometry Course Collaboration, including **Kwan B**] Untargeted Mass Spectrometry-Based Metabolomics Tracks Molecular Changes in Raw and Processed Foods and Beverages. *Food Chemistry*. 2020; 302:125290. DOI: 10.1016/j.foodchem.2019.125290.

Acheson DT, **Kwan B**, Maihofer AX, Risbrough VB, Nievergelt CM, Clark JW, Tu X, Irwin MR, and Baker DG. Sleep Disturbance at Pre-deployment is a Significant Predictor of Post-Deployment Re-Experiencing Symptoms. *European Journal of Psychotraumatology*. 2019; 10(1):1679964. DOI: 10.1080/20008198.2019.1679964.

Su HI, Stark S, **Kwan B**, Boles S, Chingos D, Ehren J, Gorman JR, Krychman M, Romero SAD, Mao JJ, Pierce JP, and Natarajan L. Efficacy of a web-based women's health survivorship care plan for young breast cancer survivors: a randomized controlled trial. *Breast Cancer Research and Treatment*. 2019; 176:579–589. DOI: 10.1007/s10549-019-05260-6.

Stark SS, Natarajan L, Chingos D, Ehren J, Gorman JR, Krychman M, **Kwan B**, Mao JJ, Myers E, Walpole T, Pierce JP, and Su HI. Design of a randomized controlled trial on the efficacy of a reproductive health survivorship care plan in young breast cancer survivors. *Contemporary Clinical Trials*. 2019; 77:27–36. DOI: 10.1016/j.cct.2018.12.002.

Wang B, Wu P, **Kwan B**, Tu XM, and Feng C. Simpson's Paradox: Examples. *Shanghai Archives of Psychiatry*. 2018; 30(2):139–143. DOI: 10.11919/j.issn.1002-0829.218026.

ABSTRACT OF THE DISSERTATION

Metabolomic Markers and Functional Data Methods for Characterizing and Predicting Diabetic Kidney Disease Progression

by

Brian Kwan

Doctor of Philosophy in Biostatistics

University of California San Diego, 2021

Professor Loki Natarajan, Chair

Patients with diabetic kidney disease (DKD) are at high risk for hospitalization, morbidity, and mortality. Early detection of patients with kidney function decline can lead to effective intervention and management of high risk of developing DKD. The human metabolome is a powerful tool for informing the physiological and pathological effects of chronic diseases and could offer direct insights into biochemical pathways potentially linked to kidney dysfunction. Furthermore, functional principal components analysis (FPCA) is a novel approach for modeling and studying the variation of kidney function trajectories for subgroups of diabetic patients, while accounting for complexity in curve estimation. Here, we applied, validated, and extended rigorous statistical approaches that utilize metabolomic markers and functional data methods for uncovering the characteristics of and predicting DKD progression.

In Chapter 1, we give an overview of the background and rationale for our distinct research aims. In Chapter 2, we elucidate the choice between fitting a linear mixed model, with serial estimated glomerular filtration rate (eGFR) outcomes, and two-stage methods, with patient-specific eGFR slopes as outcomes, for modeling DKD progression, with metabolites as predictors. Notably, two-stage models offer a suitable modeling alternative to DKD researchers who can readily implement individual eGFR slopes in standard regression models. In Chapter 3, we apply the top-scoring pair (TSP) algorithm to derive simple, parameter-free decision rules (i.e., pair of metabolites) for binary DKD stage classification. As a methodological contribution, we extended the TSP approach to allow adjustment for clinical variables. In Chapter 4, we implement the FPCA approach, which accounts for nonlinear trajectories via nonparametric smoothing while overcoming sparsity and irregularly spaced data. We examined the longitudinal patterns of kidney function trajectories within clinically defined albuminuria-specific groups and expand the FPCA inferential framework for considering whether separate group-level models to prospectively predict group-specific outcome trajectories are needed. Our findings provided insights into modeling choices for DKD progression, markers for renal dysfunction adjusted for clinical variables, dominant modes of eGFR variation, and varying eGFR patterns between albuminuria groups, which can potentially inform therapeutic targets for personalized DKD treatments.

Chapter 1

Introduction

Quantifying progression and identifying prognostic factors of chronic diseases is a arguably the primary focus of public health and biomedical research. Thus statistical methods for (i) modeling disease progression using longitudinal biomarkers and (ii) building useful and interpretable prognostic models are of keen interest. In this work, we implement, develop and compare statistical approaches focused on these two objectives in the context of diabetic kidney disease.

Diabetes mellitus, or simply diabetes, is a group of diseases characterized by excess levels of glucose in the blood. The pancreas produces the insulin hormone to allow the body to use and store glucose as energy and regulate blood glucose levels. Diabetes is a leading cause of chronic kidney disease (CKD) among the U.S. adult population [1, 2, 3, 4, 5], and more than 90% of all diabetic cases are type 2 diabetes [6]. Patients with diabetic kidney disease (DKD) are at high risk for hospitalization, morbidity, and mortality [7], with severe DKD progression potentially leading to end-stage renal disease (ESRD), otherwise known as kidney failure. At this terminal stage of kidney disease, patients are required to be treated by dialysis or kidney transplant to elongate their life expectancy [1]. In a study of diabetes-related complications from 1990 to 2010, the rate reduction of ESRD cases was lower

compared to the rates of other diabetes-complications among U.S. adults with diagnosed diabetes [8]. Thus, early detection of diabetic patients with rapid kidney function decline can lead to effective intervention and management of DKD progression.

Two clinical markers ubiquitously used for the assessment of kidney function are albuminuria and estimated glomerular filtration rate. Albuminuria is the condition of having an unusually larger amount of albumin in the urine, which is indicative of greater damage to the kidneys. Albuminuria is often stratified into groups of CKD risk, normo-, micro-, and macro-albuminuria, in which patients with micro- or macro-albuminuria warrant monitoring for kidney disease progression. Studies have demonstrated a large proportion of patients could still develop kidney disease progression without micro- or macro-albuminuria [9, 10]. Moreover, early kidney function decline may precede the onset of micro-albuminuria and its development to macro-albuminuria [11]. Estimated glomerular filtration rate (eGFR) is widely used as the standard metric for kidney function with lower levels associated with increased loss of kidney function. Equations have been widely developed and studied to calculate eGFR for assessing kidney function based on notable risk factors, such as serum creatinine, cystatin C, age, race, and sex [12, 13, 14, 15, 16]. In the following chapters of our research aims, we use serial eGFR measures to quantify DKD progression and establish eGFR thresholds to define DKD severity stages.

Metabolomics is the study of metabolites, small molecules that are products of the metabolism and found within cells, tissues, and biofluids. As the furthest downstream product of the genome and its interactions with the biological system, the metabolome provides a direct representation of the molecular phenotype, which makes it a powerful tool for studying the effects of chronic diseases [17, 18]. Recent systematic reviews noted the potential of metabolites for discriminating DKD from controls [19, 20, 21, 22, 23]. Despite their potential, metabolomics data is typically high-dimensional and standard statistical methods are largely not applicable for this setting [24].

Our recent study evaluated the associations of thirteen previously identified metabolites with future DKD progression [25]. We implemented rigorous statistical methods to construct cross-validated multivariate models for kidney function decline that noted several metabolites improving prognostication over and above clinical variables. The following chapters dive into distinct research aims inspired by our study's gaps and limitations. First, our study utilized the "two-stage" approach to modeling kidney function decline via eGFR slope which could result in loss of efficiency in the presence of irregularly spaced time measures and missing data. Second, we worked with an a priori set of thirteen metabolites with notable potential in discriminating DKD from healthy controls [26] and we would like to uncover possibly more biomarkers over a larger pool of metabolites. Third, use of eGFR slope outcome entailed a linearity assumption and we would like to account for nonlinear trajectories. Finally, there may be clinically distinct populations (e.g., albuminuria group) for whom our metabolite signatures may not be optimal.

1.1 Distinct Aspects of Research

Each of the following Chapters 2-4 addresses a distinct research aim in the application of statistical methods for identifying patients at high risk of developing DKD.

In Chapter 2, we compare the use of a linear mixed model and two-stage methods for predicting future disease progression based on clinic entry biomarker data under a set of realistic study design scenarios, e.g., irregularly spaced time measures and missing data in repeated outcome measures, via simulations and analytic calculations. While the linear mixed model is considered the more conventional approach for modeling disease progression, the two-stage methods can be easily implemented using standard statistical methods with slope outcomes, which makes it more accessible for applied researchers. Although our work here is framed in the metabolite-DKD context, our findings are generalizable to

other disease prognostic modeling studies.

In Chapter 3, we implement the top-scoring pair (TSP) and K-TSP binary classification methods, in addition to proposing our residualizing approach that takes into account covariates (e.g., clinical factors) that influence features, for the selection of metabolite-pairs that best discriminate between DKD severity stages. We demonstrated by simulation and application that incorporating our residualizing approach to the existing TSP and K-TSP algorithms could identify novel (residualized) feature-pairs compared to typically using the raw, or unresidualized, features. The residualized metabolites serve to be cleaner features for discriminating DKD severity in which they are largely liberated from much of the extraneous influence of clinical covariates. Furthermore, we compared the classification accuracy of DKD severity stage between TSP, K-TSP, and conventional statistical learning methods, i.e., LASSO and random forests, using both raw and residualized metabolite features.

In Chapter 4, we utilize functional principal components analysis (FPCA) methods to predict long-term eGFR trajectories, uncover for dominant modes of eGFR variation, and investigate for differences in longitudinal eGFR patterns between albuminuria groups in CKD. As a follow-up, we developed a novel goodness-of-fit procedure to elucidate whether fitting a single overall model, trained using data from diabetic patients across all albuminuria groups, is preferred over fitting multiple albuminuria group-specific models, each fitted using data from only diabetic patients of one particular group, for accurately predicting eGFR trajectories for test patients.

Chapter 2

Linear mixed model vs two-stage methods: Developing prognostic models of diabetic kidney disease progression

2.1 Abstract

Identifying prognostic factors for disease progression is a cornerstone of medical research. Repeated assessments of a marker outcome are often used to evaluate disease progression, and the primary research question is to identify factors associated with the longitudinal trajectory of this marker. Our work is motivated by diabetic kidney disease (DKD), where serial measures of estimated glomerular filtration rate (eGFR) are the longitudinal measure of kidney function, and there is notable interest in identifying factors, such as metabolites, that are prognostic for DKD progression. Linear mixed models (LMM) with serial marker outcomes (e.g., eGFR) are a standard approach for prognostic model development, namely by evaluating the time × prognostic factor (e.g., metabolite) interaction. However, two-stage methods that first estimate individual-specific eGFR slopes, and then

use these as outcomes in a regression framework with metabolites as predictors are easy to interpret and implement for applied researchers. Herein, we compared the LMM and two-stage methods, in terms of bias and mean squared error via analytic methods and simulations, allowing for irregularly spaced measures and missingness. Our findings provide novel insights into when two-stage methods are suitable longitudinal prognostic modeling alternatives to the LMM. Notably, our findings generalize to other disease studies.

2.2 Introduction

Repeated longitudinal assessment of a marker of disease occurrence or progression is common in medical studies, e.g., serial measures of prostate specific antigen as a marker of prostate cancer, or repeated hemoglobin A1C for diabetes control [27, 28]. Often, interest lies in identifying baseline factors associated with longitudinal trajectories of these markers, as these factors could provide early insights into actionable guidelines/treatments for the condition in question. Statistical methods for modeling these risk factor-longitudinal marker assessments is the focus of this article, with the specific research question motivated by our prior work in diabetic kidney disease (DKD) [25].

Diabetes is a leading cause of kidney disease and patients with DKD are at high risk of morbidity, hospitalization, and overall mortality [1, 7]. Studies have shown that the human metabolome has considerable potential for characterizing patients with DKD versus healthy controls [19, 20, 21, 22, 23, 26]. By incorporating metabolomic analysis into statistical model development, we could construct prognostic models for early detection of patients at high risk of developing DKD, potentially leading to earlier and more targeted treatments. Estimated glomerular filtration rate (eGFR) is a clinically accepted method for measuring kidney function, with higher eGFR indicating better kidney function [13]; slope of serial eGFR assessments, interpreted as annual eGFR change, are widely used to eval-

uate kidney disease progression. In our previous work [25], we implemented a two-stage approach for identifying metabolomic predictors of DKD progression via, first estimating eGFR slope, and then using this slope as the outcome in a regression model with baseline metabolites as predictors. We used data collected from the Chronic Renal Insufficiency Cohort (CRIC) [29, 30, 31], a racially and ethnically diverse group of adults aged 21 to 74 years with a broad spectrum of renal disease severity, one of the largest in the US, with comprehensive data on clinical and metabolite profiles. However, a more conventional and statistically accepted modeling approach is to fit a single linear mixed model with serial eGFR measures (outcomes) and evaluate the coefficient of the metabolite (biomarker) × time (year) interaction term, also interpreted as annual eGFR change. Nonetheless, two-stage methods offer the advantage of estimating individual slopes, which are by themselves of interest as a marker of disease progression, and can be readily implemented as outcomes in standard regression models by researchers, as evidenced by the plethora of research that uses eGFR slopes as outcomes in DKD research [5, 32, 33, 34, 35, 36]. Given their widespread use by DKD researchers, in this paper, we aim to provide novel insights into when two-stage methods are suitable longitudinal prognostic modeling alternatives to the linear mixed model.

In prior statistical investigations, Sayers et al. [37] conducted a simulation study comparing two-stage methods with individual slope as a predictor (i.e., independent variable) for a dependent outcome by examining the bias and coverage of the association between birth length, linear growth and later blood pressure under several study design scenarios. Our set-up is different in that the slopes are the dependent variable in our models, and we aim to evaluate a variety of two-stage approaches for assessing the prognostic value of a covariate for predicting this slope. In particular, using the framework of our previous work [25], we will consider the baseline metabolite as the predictor for annual eGFR change (slope). In addition, expanding on the statistical approaches of Sayers et al. [37], we

compare via simulations the linear mixed effects model to our two-stage methods under an expanded set of study design scenarios that incorporate irregularly spaced time measures, and missing data and also analytically examine and compare bias and efficiency across methods. More specifically, in Section 2.3, we outline our statistical approaches which include a range of two-stage methods. In Section 2.4, we describe in detail our simulation process, study design scenarios, and comparison performance metrics for our statistical approaches. Section 2.5 showcases analytical derivations for the relationships between our statistical models. Section 2.6 presents the simulation results for our statistical approaches under our set of study design scenarios. Lastly, Section 2.7 discusses the overall findings, current limitations, and future directions for this work. We emphasize that although this paper is motivated by the metabolite-DKD context with the terms metabolite and eGFR serving as predictor and longitudinal outcome in the following sections, this work applies to any predictor-longitudinal disease modeling application.

2.3 Statistical Approaches

2.3.1 Linear Mixed Model (LMM) Approach

The linear mixed effects model [38], ubiquitously used in longitudinal settings, incorporates fixed and random effects to model individual eGFR trajectories over time. Fixed effects are shared between all individuals and model the population mean eGFR trajectory. Random effects are unique to each individual and characterize individual eGFR profiles. Our model, which incorporated fixed effects for metabolite, time, and their interaction as well as random intercept and slope terms, was expressed as

$$y_{ij} = (\beta_0 + b_{0i} + \beta_1 * M_i) + (\beta_2 + b_{1i} + \beta_3 * M_i) * t_{ij} + \epsilon_{ij}$$

for individual i and occasion j, where y_{ij} is the eGFR response, $(\beta_0, \beta_1, \beta_2, \beta_3)$ are fixed effects and (b_{0i}, b_{1i}) are random effects, M_i is individual i's baseline metabolite value, t_{ij} is time in years, and ϵ_{ij} is the within-individual error. We assume the random effects $(b_{0i}, b_{1i}) \sim N(\mathbf{0_2}, \mathbf{\Omega})$, where $\mathbf{\Omega} = \left(\begin{smallmatrix} \omega_0 & \omega_{01} \\ \omega_{10} & \omega_1 \end{smallmatrix}\right)$, are independent of both M_i and ϵ_{ij}. The within-individual error ϵ_{ij} is assumed to be normally distributed with mean zero and variance σ^2. As our investigation primarily focuses on the association between metabolite and annual rate of eGFR change, the β_3 metabolite × time interaction coefficient is our main effect of interest. The coefficient is interpreted as the population-averaged annual rate of eGFR change for a one-unit higher in metabolite value.

An advantage to using a linear mixed effects model is that it can incorporate incomplete and unbalanced longitudinal data among individuals. Therefore, we would be avoiding the bias of using complete-case analysis as well as not requiring an equal number of available eGFR measurements nor need these measurements be at a common set of occasions for each individual. A further, more extensive overview, of the method is given in Chapter 8 of Fitzmaurice et al. [38].

2.3.2 Two-Stage Approaches

Our two-stage methods model the association between metabolite and annual rate of eGFR change in two stages: (1st) estimate individual eGFR slopes and (2nd) regress eGFR slope on metabolite as the sole predictor. The first stage estimates individual eGFR slopes C_i (for individual i) where the method of estimation varies between approaches. The second stage, similar for all approaches, fits a simple linear regression model with eGFR slope C_i, taken from the first stage, as the outcome on metabolite M_i.

$$\widehat{C}_i = \alpha_0 + \alpha_1 * M_i + \epsilon_{i,SS}, \quad \epsilon_{i,SS} \sim N(0, \sigma^2_{SS})$$

The metabolite coefficient α_1 is the association between metabolite and annual rate of eGFR change and is the population-averaged annual rate of eGFR change for a one-unit increase in metabolite value, which is interpreted similarly to the metabolite × time interaction coefficient β_3 from our linear mixed effects model.

Simple Approach

The simple approach to estimating eGFR slope is to take the difference between a subject's last and first observed eGFR measurements and divide by the elapsed time (years) between measurements.

$$\widehat{C}_{i,SIMPLE} = \frac{(y_{iJ} - y_{i1})}{(t_{iJ} - t_{i1})}$$

This method contains a notable loss of in-between measurement information and calculates annual rate of eGFR change using only the latest and first observed eGFR measurements. Due to this loss of measurement information, $\widehat{C}_{i,SIMPLE}$ will generally have greater variance than the true $\widehat{C}_i$.

Ordinary Lease Squares (OLS) Approach

We can also fit a simple linear regression model to the serial eGFR measures of an individual, with time as the predictor, to estimate the individual's eGFR slope.

$$y_{ij} = \gamma_{0i} + \gamma_{1i} * t_{ij} + \epsilon_{i,OLS}, \quad \epsilon_{i,OLS} \sim N(0, \sigma^2_{OLS})$$

The model parameters $\gamma_{0i}, \gamma_{1i}, \sigma^2_{OLS}$ are estimated by OLS. Let $\widehat{\gamma}_{1i} = \widehat{C}_{i,OLS}$ be the eGFR slope for individual i. Since the individual slopes do not all provide equally precise information (differing number of individual eGFR measurements), $\widehat{C}_{i,OLS}$ has greater variance than the true C_i. This approach requires fitting I total models to estimate all of the individ-

ual eGFR slopes.

Best Linear Unbiased Predictor (BLUP) Approach

As opposed to fitting separate simple linear regression models for each individual, we can fit a single linear mixed-effects model to the longitudinal eGFR data of all individuals to estimate all of their eGFR slopes. Our model consisted of a fixed effect for time and random intercept and slope terms.

$$y_{ij} = (\eta_0 + u_{0i}) + (\eta_1 + u_{1i}) * t_{ij} + \epsilon_{ij,BLUP}$$

for individual i and occasion j, where y_{ij} is the eGFR response, (η_0, η_1) are fixed effects and (u_{0i}, u_{1i}) are random effects, t_{ij} is time in years, and $\epsilon_{ij,BLUP}$ is the within-individual error. We assume the random effects $(u_{0i}, u_{1i}) \sim N(\mathbf{0_2}, \mathbf{\Omega_{BLUP}})$, where $\mathbf{\Omega_{BLUP}} = \begin{pmatrix} \omega_{0,BLUP} & \omega_{01,BLUP} \\ \omega_{10,BLUP} & \omega_{1,BLUP} \end{pmatrix}$, are independent of $\epsilon_{ij,BLUP}$. The estimated random effects $(\widehat{u}_{0i}, \widehat{u}_{1i})$ are the best linear unbiased predictors (BLUPs) for the true (u_{0i}, u_{1i}). The within-individual error $\epsilon_{ij,BLUP}$ is assumed to be normally distributed with mean zero and variance σ^2_{BLUP}. Our estimated individual eGFR slopes are obtained by adding the estimated mean eGFR slope $\widehat{\eta}_1$ to the estimated BLUP slopes $\widehat{u}_{1i}$, i.e. let $(\widehat{\eta}_1 + \widehat{u}_{1i}) = \widehat{C}_{i,BLUP}$.

Similarly, the model written in matrix notation is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\eta} + \mathbf{Zu} + \boldsymbol{\epsilon}_{\mathbf{BLUP}}$$

where $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_I})'$ s.t. $\mathbf{Y_i} = (y_{i1}, y_{i2}, \ldots, y_{iJ})'$ and $\mathbf{Y}$ is a vector of serial eGFR response values with length $I \times J$, $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_I})'$ s.t. $\mathbf{X_i}$ is the fixed effects design matrix for subject i and dim($\mathbf{X}$) = $(I \times J) \times 2$, $\boldsymbol{\eta} = (\eta_0, \eta_1)'$, $\mathbf{Z} = \begin{pmatrix} Z_1 & & \\ & \ddots & \\ & & Z_I \end{pmatrix}$ s.t. $\mathbf{Z_i}$ is the random effects design matrix for subject i and dim($\mathbf{Z}$)

= $(I \times J) \times (2 \times J)$, $\mathbf{u} = (\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_I})'$ s.t. $\mathbf{u_i} = (u_{0i}, u_{1i})'$ and $\mathbf{u}$ is a vector of random effects with length $2 \times J$, and $\boldsymbol{\epsilon_{BLUP}} = (\boldsymbol{\epsilon_{1,BLUP}}, \boldsymbol{\epsilon_{2,BLUP}}, \ldots, \boldsymbol{\epsilon_{I,BLUP}})'$ s.t. $\boldsymbol{\epsilon_{i,BLUP}} = (\boldsymbol{\epsilon_{i1,BLUP}}, \boldsymbol{\epsilon_{i2,BLUP}}, \ldots, \boldsymbol{\epsilon_{iJ,BLUP}})'$ and $\boldsymbol{\epsilon}_{\mathbf{BLUP}}$ is a vector of eGFR measurement errors with length $I \times J$. For our setup, we assume $\mathbf{X_i} = \mathbf{Z_i}$ since our model consisted of only a fixed effect for time, while having both random intercept and slope terms. We assume $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G_{BLUP}})$, where $\mathbf{G_{BLUP}} = \begin{pmatrix} \boldsymbol{\Omega}_{BLUP} & & \\ & \ddots & \\ & & \boldsymbol{\Omega}_{BLUP} \end{pmatrix}$ and $\boldsymbol{\epsilon}_{\mathbf{BLUP}} \sim N(\mathbf{0}, \sigma^2_{BLUP}\mathbf{I})$ are independent of each other. We note that the BLUPs $\widehat{\mathbf{u}}$ are a weighted average of the population- and individual-level counterparts, and hence will have lower variance than the true values. We discuss a way to address this in the next section.

Inflated Approach

To address the under-estimation of variances of the BLUP random effects in comparison to its restricted maximum likelihood (REML) estimation for the covariance matrix G_{BLUP}, Carpenter et al. [39] transformed (re-inflated) the random effects so that their crude covariance matrix is more equivalent to G_{BLUP}. The re-inflated random effects are then added to the estimated fixed effects intercept $\widehat{\eta}_0$ and slope $\widehat{\eta}_1$ to give the estimated eGFR baseline value and slope, respectively, for each individual.

Here we briefly state the analytic steps as described by Sayers et al. [37]. The re-inflation process involves multiplying our estimated random effects matrix by an upper triangular matrix of equal order. Hence, we require finding a transformation $\mathbf{A}$ such that

$$\widehat{\mathbf{U}}^* = \widehat{\mathbf{U}}\mathbf{A}$$

where $\widehat{\mathbf{U}}^*$ is the matrix of the inflated random effects and $\widehat{\mathbf{U}}$ is the matrix of our originally estimated random effects, both with I rows and 2 columns. The matrix $\mathbf{A}$ is formed using the lower triangular Cholesky decompositions of the empirical covariance matrix of

the estimated random effects as well as its corresponding REML covariance matrix. The empirical covariance matrix is calculated as

$$\mathbf{S} = \widehat{\mathbf{U}}^T \widehat{\mathbf{U}} \; / \; \mathbf{N}$$

and the REML covariance matrix as

$$\mathbf{R} = \widehat{\mathbf{\Omega}}_{\mathbf{BLUP}}$$

and $\mathbf{S}$ and $\mathbf{R}$ written in terms of their lower triangular Cholesky decompositions are

$$\mathbf{S} = \mathbf{L_S} \mathbf{L_S^T}$$

$$\mathbf{R} = \mathbf{L_R} \mathbf{L_R^T}$$

Finally, $\mathbf{A}$, an upper triangular matrix can be calculated as

$$\mathbf{A} = (\mathbf{L_R} \mathbf{L_S^{-1}})^T$$

The transformed (re-inflated) random effects $\widehat{\mathbf{U}}^*$ now have covariance matrix equivalent to that of the model estimate $\widehat{\mathbf{\Omega}}_{\mathbf{BLUP}}$.

2.4 Simulation study design

We compare our statistical approaches, i.e., linear mixed model vs two-stage methods, under various study design scenarios. Since the linear mixed model is the more conventional method for modeling disease progression, it served as the data generating model for our simulated study. Our model consisted of fixed effects for metabolite, time (i.e.,

year of follow-up) and their interaction as well as random intercept and slope terms. The number of individuals in our study (I) was set to 200, representing a medium sized study cohort, and individuals had eGFR measurements biennially from 0 to 10 years of follow-up ($J = 6$). For scenarios with missing eGFR data, value J will vary by individual ($J_i \leq 6$). Regardless, we use J for our model notation. We compared the bias and efficiency of our 4 two-stage statistical modeling approaches across study design scenarios based on differing (1) choice of spacing between eGFR measures (regularly vs irregularly spaced), (2) amount of missing completely at random (MCAR) eGFR data (complete, 20%, 50%, 80%), and (3) standard deviation (SD) value for the metabolite, random intercept, random slope, and measurement error as well as the correlation value between the random effects in our data generating model. Our chosen values are as follows

(a) σ_M (Metabolite) = (0.79), 2, 7, 10, 15, 20

(b) ω_0 (Random Intercept) = 0.5, 1, 4, 7, (9.87), 12, 16

(c) ω_1 (Random Slope) = 0.5, 1, (2.27), 4, 7, 10

(d) σ_{err} (Error) = 0.5, 1, 3, (5.87), 8, 10, 15

(e) ρ_ω (Random Effects Corr.) = -1, -0.75, -0.5, -0.25, 0, (0.159), 0.25, 0.5, 0.75, 1

When varying a particular simulation parameter (e.g., metabolite SD), the values for the other parameters (i.e., random intercept SD, random slope SD, error SD, and random effects correlation) were held fixed at the value shown in parentheses for that parameter. The parameter values in parentheses for the varying parameters were selected for our data generating model as they were the numerical estimates from the linear mixed model fitted to the analytic cohort of the Chronic Renal Insufficiency (CRIC) Study from our previous work [25]. In addition, again following from our previous work, the fixed effect parameters and the mean of the metabolite were fixed for all simulations as follows:

(a) $\beta_0 = -24.22$

(b) $\beta_1 = 4.34$

(c) $\beta_2 = -5.13$

(d) $\beta_3 = 0.22$

(e) μ_M (Metabolite Mean) = 14.8

In total, we studied 48 different scenarios and generated D = 1000 replications for each of them to assess the performance of our statistical approaches. We compared the performance of the different methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) by examining the (relative) bias and efficiency, i.e. standard deviation (SD), standard error (SE), and root mean square error (MSE), across methods. The bias, relative bias, SD, and SE are defined as the following:

$$\begin{aligned}
\text{Bias} &= \left(\frac{1}{D}\sum_{d=1}^{D}\widehat{\alpha}_{1,d}\right) - \beta_3 \\
\text{Rel. Bias}(\%) &= \frac{\text{Bias}}{\beta_3} \times 100 \\
\text{Standard Deviation} &= \sqrt{\frac{1}{D-1}\sum_{d=1}^{D}\left(\widehat{\alpha}_{1,d} - \frac{1}{D}\sum_{d=1}^{D}\widehat{\alpha}_{1,d}\right)^2} \\
\text{Standard Error} &= \frac{1}{D}\sum_{d=1}^{D}\text{SE}(\widehat{\alpha}_{1,d})
\end{aligned}$$

where D is the total number of replications and root MSE is calculated as $\sqrt{\text{Bias}^2 + \text{SD}^2}$. The notation here uses our estimated association from the two-stage models ($\widehat{\alpha}_{1,d}$), so calculating these statistics of interest for the linear mixed model would require replacing $\widehat{\alpha}_{1,d}$ with $\widehat{\beta}_{3,d}$.

Simulation study design and statistical analysis was conducted using the R (version 3.6.1) programming environment [40].

2.5 Analytical Relationships Between Statistical Models

2.5.1 Unbiased association for the Simple and OLS methods

We prove analytically that the Simple and OLS methods have unbiased association for the study design scenario with regularly spaced measures and complete data. In particular, our general second-stage model was

$$\widehat{C}_i = \alpha_0 + \alpha_1 * M_i + \epsilon_{i,SS}, \quad \epsilon_{i,SS} \sim N(0, \sigma^2_{SS})$$

and we show that $E(\widehat{\alpha}_1) = \beta_3$ with $\widehat{C}_{i,SIMPLE}$ or $\widehat{C}_{i,OLS}$ as the outcome. The coefficient α_1 has estimate

$$\widehat{\alpha}_1 = \frac{\sum_{i=1}^{I}(M_i - \overline{M})\widehat{C}_i}{\sum_{i=1}^{I}(M_i - \overline{M})^2}$$

where $\overline{M} = \frac{1}{I}\sum_{i=1}^{I} M_i$. We can rewrite $\widehat{C}_{i,SIMPLE}$ based on our data generating model and obtain

$$\widehat{C}_{i,SIMPLE} = (\beta_2 + \beta_3 * M_i + b_{1i}) + \frac{(\epsilon_{iJ} - \epsilon_{i1})}{(t_{iJ} - t_{i1})}$$

Our first-stage model in the OLS approach was

$$y_{ij} = \gamma_{0i} + \gamma_{1i} * t_{ij} + \epsilon_{i,OLS}, \quad \epsilon_{i,OLS} \sim N(0, \sigma^2_{OLS})$$

and we let $\widehat{\gamma}_{1i} = \widehat{C}_{i,OLS}$ be the eGFR slope for individual i such that

$$\widehat{C}_{i,OLS} = \frac{\sum_{j=1}^{J}(t_{ij} - \overline{t}_i)(y_{ij} - \overline{y}_i)}{\sum_{j=1}^{J}(t_{ij} - \overline{t}_i)^2}$$

where $\bar{t}_i = \frac{1}{J}\sum_{j=1}^{J} t_{ij}$ and $\overline{y}_i = \frac{1}{J}\sum_{j=1}^{J} y_{ij}$. Similarly, we can write $\widehat{C}_{i,OLS}$ based on our data generating model and obtain

$$\widehat{C}_{i,OLS} = (\beta_2 + \beta_3 * M_i + b_{1i}) + \frac{\sum_{j=1}^{J}(t_{ij} - \bar{t}_i)(\epsilon_{ij} - \bar{\epsilon}_i)}{\sum_{j=1}^{J}(t_{ij} - \bar{t}_i)^2}$$

where $\bar{\epsilon}_i = \frac{1}{J}\sum_{j=1}^{J} \epsilon_{ij}$. We can write $\widehat{C}_{i,OLS}$ as a function of $\widehat{C}_{i,SIMPLE}$

$$\widehat{C}_{i,OLS} = \widehat{C}_{i,SIMPLE} - \frac{(\epsilon_{iJ} - \epsilon_{i1})}{(t_{iJ} - t_{i1})} + \frac{\sum_{j=1}^{J}(t_{ij} - \bar{t}_i)(\epsilon_{ij} - \bar{\epsilon}_i)}{\sum_{j=1}^{J}(t_{ij} - \bar{t}_i)^2}$$

and if $J = 2$, then $\widehat{C}_{i,OLS} = \widehat{C}_{i,SIMPLE}$.

Defining $\widehat{C}_{i,SIMPLE}$ based on our data generating model and having it as the outcome for the second-stage model, the estimated association $\widehat{\alpha}_1$ is

$$\widehat{\alpha}_1 = \frac{\sum_{i=1}^{I}(M_i - \overline{M})\left[(\beta_2 + \beta_3 * M_i + b_{1i}) + \frac{(\epsilon_{iJ} - \epsilon_{i1})}{(t_{iJ} - t_{i1})}\right]}{\sum_{i=1}^{I}(M_i - \overline{M})^2}$$

Taking the expected value, we have

$$E(\widehat{\alpha}_1) = \beta_3 * \frac{\sum_{i=1}^{I}(M_i - \overline{M})M_i}{\sum_{i=1}^{I}(M_i - \overline{M})^2}$$

and by simplifying we have $E(\widehat{\alpha}_1) = \beta_3$ and conclude that using the Simple slopes for our Two-Stage method give an unbiased association between annual rate of eGFR change and metabolite.

Similarly, defining $\widehat{C}_{i,OLS}$ based on our data generating model and having it as the outcome for the second-stage model, the estimated association $\widehat{\alpha}_1$ is

$$\widehat{\alpha}_1 = \frac{\sum_{i=1}^{I}(M_i - \overline{M})\left[(\beta_2 + \beta_3 * M_i + b_{1i}) + \frac{\sum_{j=1}^{J}(t_{ij} - \bar{t}_i)(\epsilon_{ij} - \bar{\epsilon}_i)}{\sum_{j=1}^{J}(t_{ij} - \bar{t}_i)^2}\right]}{\sum_{i=1}^{I}(M_i - \overline{M})^2}$$

Taking the expected value, we have $E(\widehat{\alpha}_1) = \beta_3$ and conclude that using the OLS slopes for our Two-Stage method also give an unbiased association between annual rate of eGFR change and metabolite.

2.5.2 Correction of association bias for the BLUP methods

In contrast, our BLUP method will contain noticeable bias for the association between annual rate of eGFR change and metabolite, assuming that the β_3 metabolite × time interaction coefficient is the true association. We first derive the bias analytically, and then show how to correct for this bias by a transformation matrix for our estimated random effects (intercept & slope). Like before, we assume the study design scenario with regularly spaced measures and complete data.

In order to derive the parameters of interest, recall that our general second-stage model was

$$\widehat{C}_i = \alpha_0 + \alpha_1 * M_i + \epsilon_{i,SS}, \quad \epsilon_{i,SS} \sim N(0, \sigma^2_{SS})$$

and our goal is to estimate $\widehat{\alpha}_1$ with $\widehat{C}_{i,BLUP}$ as the outcome. As noted in Section 2.3.2, individual eGFR (BLUP) slopes are obtained by adding the estimated mean eGFR slope $\widehat{\eta}_1$ to the estimated BLUP slopes $\widehat{u}_{1i}$, i.e. let $(\widehat{\eta}_1 + \widehat{u}_{1i}) = \widehat{C}_{i,BLUP}$. Using standard mixed model theory we know that the $\widehat{\mathbf{U}}$ matrix of our estimated (centered) random effects (intercept & slope) is indexed by I rows and 2 columns, and can be estimated as $\widehat{\mathbf{U}} = \mathbf{G_{BLUP}Z^T V^{-1}}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$, where $\widehat{\boldsymbol{\beta}} = (\mathbf{X^T V^{-1} X})^{-1}\mathbf{X^T V^{-1} Y}$ and $\mathbf{V} = \mathbf{Z G_{BLUP} Z^T} + \sigma^2_{BLUP}\mathbf{I_N}$ [38]. Having estimated the BLUP slopes, the second step of our two-stage method is to regress this BLUP slope on the metabolite predictor, i.e., to estimate $\widehat{\alpha}_1$. However, although our main focus is on the BLUP slope, for ease of theoretical development, we will use matrix notation, and consider the regression problem $E(\widehat{\mathbf{U}}|\mathbf{M})$, i.e., include random intercept and slope, and evaluate the expected value of our estimated

random effects conditioned on the metabolite predictor. Using algebraic manipulations we see that:

$$\begin{aligned} E(\widehat{\mathbf{U}}|\mathbf{M}) &= E(\mathbf{G_{BLUP}}\mathbf{Z^T}\mathbf{V^{-1}}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})|\mathbf{M}) \\ &= E(\mathbf{G_{BLUP}}\mathbf{Z^T}\mathbf{V^{-1}}\mathbf{HZU}|\mathbf{M}), \text{ where } \mathbf{H} = \mathbf{I_N} - \mathbf{X}(\mathbf{X^T}\mathbf{V^{-1}}\mathbf{X})^{-1}\mathbf{X^T}\mathbf{V^{-1}} \\ &= \mathbf{G_{BLUP}}\mathbf{Z^T}\mathbf{V^{-1}}\mathbf{HZ} * E(\mathbf{U}|\mathbf{M}) \end{aligned}$$

We can see that regressing the estimated random effects results in a multiplicative bias matrix of $\mathbf{G_{BLUP}}\mathbf{Z^T}\mathbf{V^{-1}}\mathbf{HZ}$ on the true random effects $\mathbf{U}$. Thus except in the unlikely scenario that this bias matrix is the identity, use of BLUP slopes will result in biased estimates. We could correct for this bias by taking the inverse of this bias as a transformation matrix for our estimated random effects and multiply it to both sides.

$$(\mathbf{G_{BLUP}}\mathbf{Z^T}\mathbf{V^{-1}}\mathbf{HZ})^{-1}E(\widehat{\mathbf{U}}|\mathbf{M}) = E(\mathbf{U}|\mathbf{M})$$

The recalculated random effects $(\mathbf{G_{BLUP}}\mathbf{Z^T}\mathbf{V^{-1}}\mathbf{HZ})^{-1}\widehat{\mathbf{U}}$ will yield both transformed intercepts and slopes for individuals, which when the slope is used as the outcome for the second-stage model gives an unbiased association between annual rate of eGFR change and metabolite, assuming that the β_3 metabolite × time interaction coefficient is the true association. We apply this correction for our BLUP method in the simulation study; however, calculating the inverse of $\mathbf{G_{BLUP}}\mathbf{Z^T}\mathbf{V^{-1}}\mathbf{HZ}$ proved to be unfeasible in study design scenarios with irregularly spaced time measures or MCAR data.

2.6 Simulation Results

We compared the bias and efficiency of the linear mixed model to our two-stage methods under our simulation study design, with the linear mixed model as the data generating model. We organized our simulation results based on varying a certain parameter in our data generating model. The text, table, and figures elaborate on the results for Complete Data and MCAR 50% and we describe the results in the text for MCAR 20% and 80% in relation to Complete Data and MCAR 50%.

2.6.1 No Varying Parameters

Table 2.1 shows the results. There were similar results for having regularly spaced and irregular spaced time measures in Complete Data. The LMM, Simple, and OLS methods have negligible bias supporting our analytic solution (Section 2.5.1) of the Simple and OLS methods having unbiased association. There is notable upward and downward bias for the BLUP and Inflated methods, respectively. However, after correcting for the bias in our BLUP slopes from our proposed analytic solution (Section 2.5.2), the BLUP had minimal bias (0.004) equal to that of the LMM, Simple, and OLS methods.

The BLUP and Inflated methods displayed overall greater efficiency than the other methods in having lower SD, SE, and root MSE. These results also hold for regularly spaced time measures in MCAR 50%. However, for irregularly spaced time measures in MCAR 50%, the Simple and OLS methods have overwhelmingly large bias and worse efficiency while the BLUP and Inflated performed similarly as in the aforementioned scenarios. For Complete Data scenarios or regularly spaced assessments, when comparing bias and efficiency, all of the two-stage methods are well-suited for modeling the association between eGFR slope and metabolite, with a notable bias-variance trade off in the BLUP and Inflated method. However, for irregularly spaced time measures with 50% missingness

Table 2.1: Comparison of simulation results ($D = 1000, N = 200$) for the estimated association between annual rate of eGFR change and metabolite. True association $\beta_3 = 0.223$.

(a) Complete Data

Statistic	LMM	Simple	OLS	BLUP	Inflated
Bias	0.004 (-0.003)	0.004 (-0.003)	0.004 (-0.003)	0.071 (0.073)	-0.01 (-0.016)
Rel. Bias (%)	1.66 (-1.202)	1.66 (-1.507)	1.66 (-1.175)	31.959 (32.666)	-4.279 (-7.387)
SD	0.212 (0.217)	0.216 (0.225)	0.212 (0.219)	0.194 (0.195)	0.203 (0.205)
SE	0.214 (0.217)	0.218 (0.223)	0.214 (0.217)	0.196 (0.193)	0.205 (0.205)
Root MSE	0.212 (0.217)	0.216 (0.225)	0.212 (0.219)	0.207 (0.209)	0.203 (0.206)

(b) MCAR 50%

Statistic	LMM	Simple	OLS	BLUP	Inflated
Bias	-0.003 (0.003)	-0.008 (7.631)	-0.007 (7.64)	0.136 (0.141)	-0.034 (-0.029)
Rel. Bias (%)	-1.43 (1.127)	-3.398 (3421.86)	-3.187 (3426.183)	60.998 (63.097)	-15.124 (-13.145)
SD	0.241 (0.259)	0.275 (243.471)	0.274 (243.472)	0.189 (0.19)	0.208 (0.216)
SE	0.24 (0.249)	0.272 (20.169)	0.271 (20.17)	0.181 (0.177)	0.206 (0.206)
Root MSE	0.242 (0.259)	0.275 (243.591)	0.274 (243.591)	0.233 (0.237)	0.21 (0.218)

Results displayed as: Regularly Spaced case (Irregularly Spaced case).
LMM, Linear Mixed Model; OLS, Ordinary Least Squares;
BLUP, Best Linear Unbiased Predictor;
SD, Standard Deviation; SE, Standard Error; MSE, Mean Squared Error.
Bias = $\left(\frac{1}{D}\sum_{d=1}^{D}\widehat{\alpha}_{1,d}\right) - \beta_3$; Rel. Bias (%) = $\frac{\text{Bias}}{\beta_3} \times 100$;
Standard Deviation = $\sqrt{\frac{1}{D-1}\sum_{d=1}^{D}\left(\widehat{\alpha}_{1,d} - \frac{1}{D}\sum_{d=1}^{D}\widehat{\alpha}_{1,d}\right)^2}$;
Standard Error = $\frac{1}{D}\sum_{d=1}^{D}\text{SE}(\widehat{\alpha}_{1,d})$;
Root MSE = $\sqrt{\text{Bias}^2 + \text{SD}^2}$

under a MCAR mechanism, we do not recommend using the Simple and OLS methods, as these displayed large bias and root MSE.

Results for MCAR 20%, in both the regularly and irregularly spaced cases, and MCAR 80%, in just the regularly spaced case, were similar to those of Complete Data; results for the irregularly spaced case for MCAR 80% were similar to the same case for MCAR 50%.

2.6.2 Vary Metabolite SD

Figure 2.1 shows the results. Similar results hold across the regularly and irregularly spaced time measures for Complete Data and the regularly spaced case for MCAR 50%. The LMM, Simple, and OLS have low relative bias across the spectrum of metabolite SD values, while the BLUP has notable upwards relative bias and the Inflated having downwards relative bias for lower metabolite SD values. However, with increasing metabolite SD values, the relative bias shrinks toward zero for both the BLUP and Inflated methods. Even with the bias in the BLUP and Inflated methods, they remain competitive to the LMM, Simple, and OLS methods in root MSE due to their lower SD values for their estimated metabolite associations (Appendix Figure A.1). In contrast with the scenario of irregularly spaced measures in MCAR 50%, the Simple and OLS methods have overwhelmingly large relative bias and SD, particularly for lower metabolite SD values which is reflected in their root MSE performance.

Results for the regularly spaced case for MCAR 20% and 80% were similar to those of the regularly and irregularly spaced cases for Complete Data and the regularly spaced case for MCAR 50%. The Simple and OLS in the irregularly spaced case for MCAR 20% had larger SD, SE, and root MSE across metabolite SD values than when they were in the aforementioned scenarios, while the irregularly spaced case for MCAR 80% shared similar results to the scenario of irregularly spaced measures in MCAR 50%.

2.6.3 Vary Random Slope SD

Figure 2.2 shows the results. For regularly spaced time measures in Complete Data and MCAR 50%, the LMM, Simple, and OLS methods have lower relative bias across all random slope SD values. The BLUP and Inflated methods both start off with notable downward relative bias with the BLUP spiking and then leveling off on a consistent up-

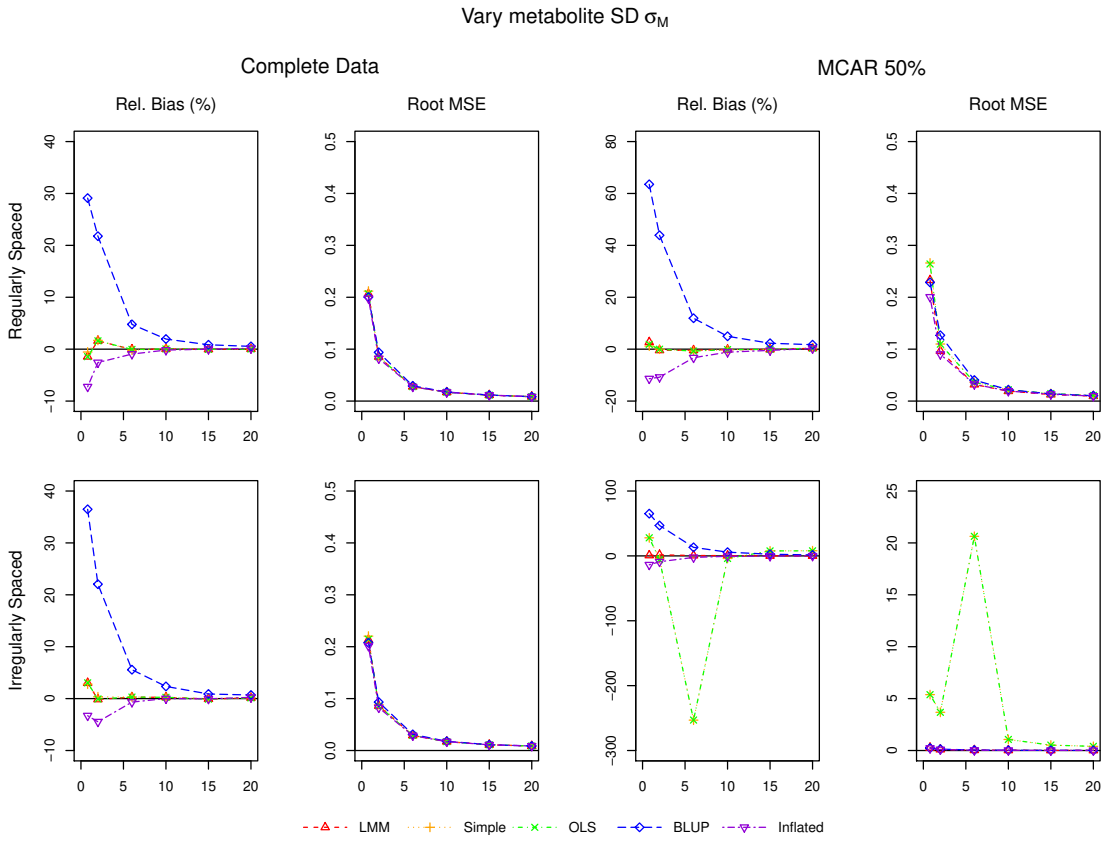

Figure 2.1: Performance in relative bias (%) and root MSE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of metabolite SD for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

wards relative bias trend and the Inflated spiking before proceeding on a decreasing trend. The root MSE was similar across all methods. Similar results were observed for irregularly spaced time measures in Complete Data except for a noticeable spike of increasing relative bias for all methods at our last random slope SD value (15), where the Inflated method had the lowest relative bias. However, for irregularly space time measures in MCAR 50%, the relative bias is noticeably worse for the Simple and OLS methods with the magnitude $>$ 100% for lower random slope SD values. The SD values were many times larger for the Simple and OLS methods (Appendix Figure A.2) resulting in their consistently higher root MSE than the other methods.

Results for the regularly spaced case for MCAR 20% were similar to those of the irregularly spaced case for Complete Data. Both the regularly spaced case for MCAR 80% and irregularly spaced case for MCAR 20% shared similar results to the regularly spaced case for MCAR 50%, but with larger SD, SE, and root MSE across random slope SD values. Finally, the irregularly spaced case for MCAR 80% shared similar results to the scenario of irregularly spaced measures in MCAR 50%.

2.6.4 Vary Correlation between Random Intercept and Slope

Figure 2.3 shows the results. For regularly and irregularly spaced time measures in Complete Data, the LMM, Simple, and OLS methods have small relative bias across correlation values. The Inflated method performs similarly except that for perfect negative or positive correction there is notable upward relative bias. However, the BLUP method has a generally increasing trend in relative bias with greater correlation values. All methods performed competitively in root MSE. Similar results hold for regularly spaced time measures in MCAR 50% but there is more noticeable separation in root MSE performance with higher values for the Simple and OLS methods for non-perfect correlation values.

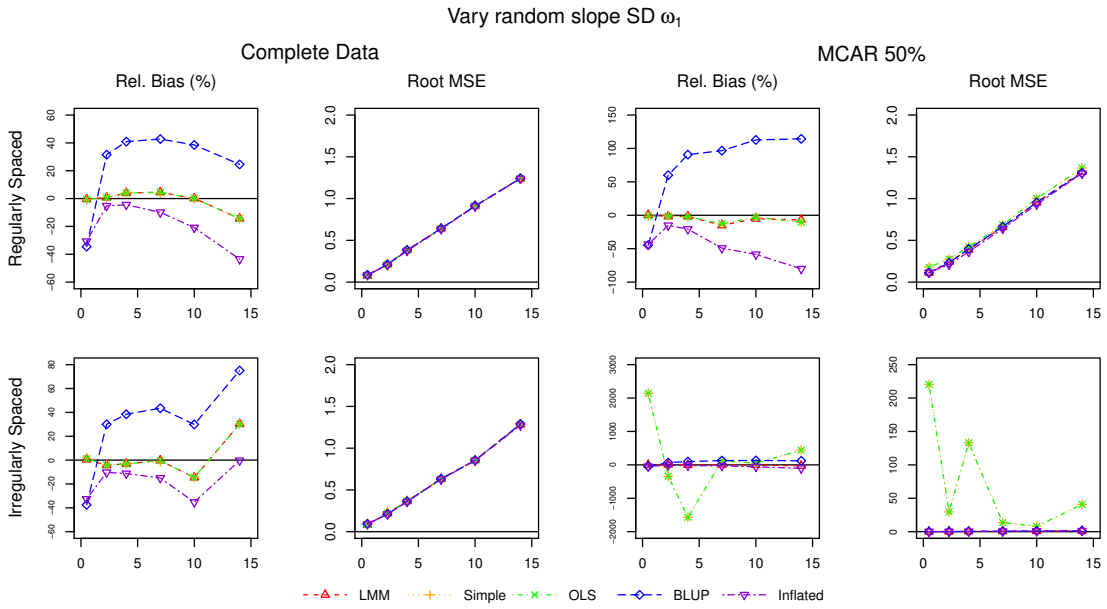

Figure 2.2: Performance in relative bias (%) and root MSE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of random slope SD for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

This is attributed to the noticeable separation in SD performance across correlation values with the Simple and OLS displaying greater variability (Appendix Figure A.3). In contrast, for irregularly spaced time measures in MCAR 50%, the Simple and OLS methods have overwhelmingly large relative bias on much worse efficiency leading to both methods have much larger root MSE across correlation values.

The regularly spaced case for MCAR 20% had similar results to having regularly and irregularly spaced time measures in Complete Data. Similar results were true for the irregularly spaced case for MCAR 20% and the regularly spaced case for MCAR 80%, but with the Simple and OLS having larger SD, SE, and root MSE than the other methods across all correlation values. Finally, the irregularly spaced case for MCAR 80% shared similar results to the scenario of irregularly spaced measures in MCAR 50%.

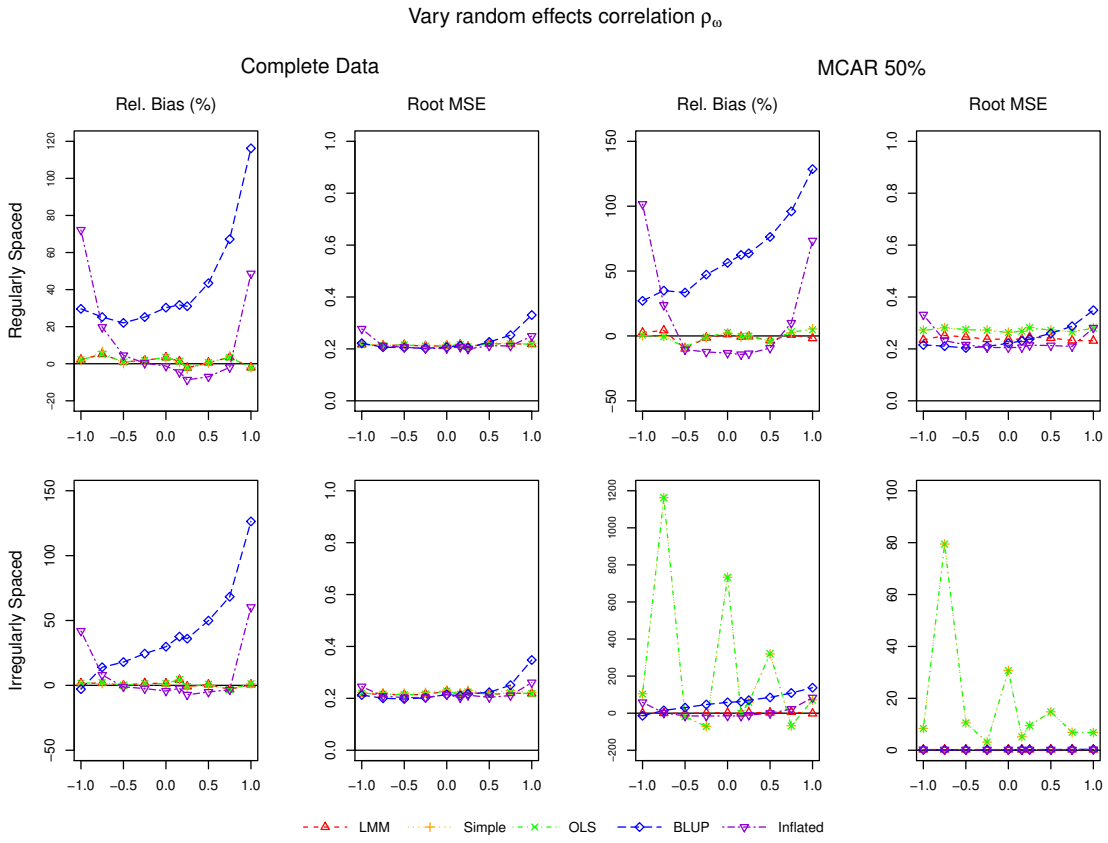

Figure 2.3: Performance in relative bias (%) and root MSE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of the correlation between random intercept and slope for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

2.6.5 Vary Random Intercept SD

The results for varying the random intercept SD are very similar to results from varying metabolite SD with a few exceptions. For regularly and irregularly spaced time measures for Complete Data and just the regularly spaced case for MCAR 50%, the BLUP method has slightly larger root MSE than the other methods for lower random intercept SD values (Appendix Figure A.4). For irregularly spaced time measures in MCAR 50%, the Simple and OLS methods also have noticeably larger root MSE from having larger relative bias and SD than the other methods (Appendix Figure A.5). Results for MCAR 80% are similar to that of MCAR 50%. Furthermore, results of MCAR 20% are similar to that of Complete Data, except that the Simple and OLS methods had overall larger SD, SE, and root MSE than the other methods for the irregularly spaced case compares to the same case for Complete Data.

2.6.6 Vary Error SD

For regularly and irregularly spaced time measures for Complete Data and just the regularly spaced case for MCAR 50%, the LMM, Simple, and OLS methods have low relative bias across all error SD values, while the BLUP and Inflated methods have growing upward and downward relative bias for increasing error SD values, respectively (Appendix Figure A.6). The BLUP and Inflated methods have lower SD values across all error SD values and thus performed similarly or better in root MSE than the other methods (Appendix Figure A.7). For irregularly spaced time measures in MCAR 50%, there was both lack of consistent direction in and relatively larger relative bias for the Simple and OLS methods, particularly for larger error SD values. In addition, the SD and root MSE for both the Simple and OLS methods displayed a monotonically increasing trend for increasing error SD values. Results for MCAR 80% are similar to that of MCAR 50%. Furthermore, results of

MCAR 20% are similar to that of Complete Data, except that the Simple and OLS methods had larger relative bias, SD, SE, and root MSE than the other methods for larger error SD values in the irregularly spaced case of MCAR 20%.

2.7 Discussion

We have uncovered study design scenarios where two-stage methods are well-suited modeling alternatives to the linear mixed model by comparing the association between metabolite and annual eGFR change. For regularly and irregular spaced time measures in Complete Data and just regularly spaced time measures in MCAR 50%, the Simple and OLS methods have lower bias than the BLUP and Inflated methods. However, we have shown that the BLUP method can correct bias and both the BLUP and Inflated methods have greater efficiency with lower SD, SE, and root MSE. This provides credence to our previous work [25] in using the BLUP approach to estimate eGFR slopes. Also, with regularly spaced or complete data, we saw that increasing the SD of metabolite or random intercept is associated with a decreasing trend in the bias for the BLUP and Inflated methods. Furthermore, with regularly spaced or complete data across random slope SD, random effects correlation, and error SD values, the Simple and OLS methods performed much more favorably in bias with the trade-off of slightly worse efficiency compared to the BLUP and Inflated methods. Thus in these scenarios the choice of optimal method will be dictated by the goals of the analysis, namely whether to minimize overall prediction error vs unbiased estimation of associations. Most importantly, throughout our simulation study when varying parameters, the Simple and OLS methods performed noticeably worse in statistical performance with irregularly spaced time measures and MCAR 50% data, scenarios that are not uncommon in observation studies, and so we do not recommend either of these methods when both data criterion are met.

We acknowledge limitations of our work. First, our linear mixed model and two-stage methods assume eGFR has a linear rate of change. Statistical models that account for nonlinear trajectories should be considered; however, despite the rigid linearity assumption, eGFR slopes are an established, clinically useful, and commonly used measure of diabetic kidney disease progression [5, 25, 32, 33, 34, 35, 36]. Second, the chosen number of subjects N=200 for our simulation study design compares our statistical approaches under a medium sized cohort and further analysis with smaller and larger N could provide additional guidance on optimal choice of methods for small and large sized cohorts, respectively. Third, we have only compared our statistical approaches under a single missing data mechanism, MCAR. More complex missing data mechanisms could also arise such as data that is missing at random (MAR) or missing not at random (MNAR). Further investigation of these topics would require additional simulation scenarios and assumptions; we aim to investigate this in future studies.

Our work has elucidated the choice between the linear mixed model vs two-stage methods for predicting possible patient future disease progression based on their clinic entry biomarker data under various study design scenarios. Although the linear mixed model is an optimal approach, there were numerous scenarios where at least one two-stage method was a suitable modeling alternative to mixed models, which opens the doors for clinicians to implement standard statistical methods using slope outcomes. Similar to Sayers et al. [37], we examined a single continuous biomarker predictor and additional studies looking into adjusting for key clinical risk factors and confounders that could further improve prognostication of disease progression, e.g., baseline eGFR [41], will further illuminate our modeling options across various study design scenarios. However, including covariates would require assumptions on joint covariate distributions and additional simulations, and we do not pursue this further here. Importantly, in our single marker setting, we were able to analytically calculate bias (or lack thereof) for the proposed two-stage methods, and

propose a method to mathematically correct this bias.

In summary, in this work via simulations and analytic calculations, we evaluated two-stage methods for estimating marker-DKD progression associations in a longitudinal setting. We examined a range of realistic study designs commonly encountered in medical research (e.g., irregularly spaced measures, missing data), and identified scenarios where two-stage models performed competitively. Of note, for many disease settings (e.g., eGFR trajectory and kidney disease, prostate-specific-antigen change and prostate cancer, rate of decline in FEV and chronic obstructive pulmonary disease) [28, 42, 43], the rate of change (i.e. slope) of the biomarker is of interest in its own right as a marker of disease, and thus is often an outcome of interest. Thus our findings are easily generalizable to other disease prognostic modeling studies.

2.8 Acknowledgements

Chapter 2, in part, has been submitted for publication of the material as it may appear in Kwan, Brian; Liu, Lin; Strong, David; Su, H. Irene; Natarajan, Loki. "Linear mixed model vs two-stage methods: Developing prognostic models of diabetic kidney disease progression". The dissertation author was the primary investigator and author of this paper.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650112. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Chapter 3

Identifying metabolite-pair markers for chronic kidney disease stage classification in diabetic patients: results from applying the top-scoring pairs algorithm to the Chronic Renal Insufficiency Cohort (CRIC) Study

3.1 Abstract

The top-scoring pair (TSP) algorithm has showcased notable potential in deriving biologically interpretable single pair decision rules that are accurate and robust in disease discrimination and classification, particularly in gene-cancer studies. However, existing

TSP-based methods do not take into account covariates that could largely influence feature selection for the top-scoring pair. We proposed our method of residualizing the features on selected covariates to obtain the top-scoring pair liberated from much of their extraneous effects and demonstrated in our simulation study that residualizing features yields a different top-scoring pair when the identified top-scoring pair from raw features was highly correlated with selected covariates. In a cohort of 977 diabetic patients selected for untargeted metabolome profiling in the Chronic Renal Insufficiency Cohort (CRIC) study, we identified raw and resiudalized metabolite ion top-scoring pairs that best discriminated between early and advanced diabetic kidney disease (DKD) stage. In particular, the residualized top-scoring pairs brought us "cleaner" metabolite ion features, uncorrelated from clinical covariates, for discriminating DKD stage, which could motivate follow-up studies on the order reversals in the disease vs non-disease states. Finally, we compared the cross-validated classification accuracy of TSP-based methods to LASSO and random forests, and found TSP-based methods can accurately identify a healthier or less severe disease group.

3.2 Introduction

The ever increasing amount of high-dimensional biomolecular data generated using high-throughput technologies has brought a critical need for decision rules that would strengthen our understanding of clinical diseases and health outcomes [44, 45, 46, 47]. A prominent challenge is deriving decision rules that are not only accurate and robust across a diverse range of settings but also have ease of biological interpretability for a desired future clinic usage. Modern statistical and machine learning methods permeate the literature and frequently achieve superior classification accuracy [48, 49, 50, 51]. However, a key limitation of these methods is the "black box dilemma", namely decision rules that make accurate assessments of patient disease and outcomes often at the expense of using

nonlinear functions of hundreds or even thousands of features, which involves estimating a plethora of model parameters. This leads to the construction of highly complex decision boundaries for distinguishing between different classes of patients, which can be difficult to interpret and characterize in a biologically meaningful manner.

We focus on the parameter-free Top-Scoring Pair (TSP) algorithm [52], that has the advantage of providing simple and biologically interpretable decision rules. As a primer, the TSP algorithm identifies a single pair of features that best discriminates between two classes of interest among all possible feature-pairs – the top-scoring pair – along a pre-defined fixed decision boundary, a 45-degree line passing through the origin in the space defined by the two features. Measure of discrimination of a feature-pair is assessed via a score for which an observed ordering of the two features is prominent in one class and scarce in the other. After the pair of features with the maximal score is identified, classification entails assigning the class for which the observed ordering of the top-scoring pair is most common in a test sample. As the TSP algorithm is concerned with the ordering of features, the method replaces the values of the features with their corresponding ranks within individual profiles prior to identifying the top-scoring pair. Since TSP bases selection in a ranks context, the algorithm is highly robust to data normalization procedures involving monotonic transformation of raw feature values.

The TSP algorithm has been noted for identifying gene-pair markers for cancer classification. For instance, the TSP classifier achieved prediction rates in breast, leukemia, and prostate cancer studies comparable to those of standard classification methods with much fewer genes [52]. The K-TSP classifier, based on the top k gene pairs and a majority voting procedure for classification, had competitive binary and multi-class prediction accuracy for human cancer compared to TSP and standard methods [53]. Moreover, the TSP algorithm identified a robust marker gene-pair for prostate cancer diagnosis through the direct integration of inter-study microarray data [54]. Finally, TSP demonstrated reliable classification

for diverse human diseases, e.g. HIV infection and diabetes, using top-scoring gene pairs [55].

However, the existing TSP methods do not take into account possible covariates that influence the features, e.g., clinical risk factors for genes, in identifying the top-scoring pair and we aim to address this gap. As a novel extension of the TSP algorithm, we propose using the residuals from a regression of features on covariates of interest to select the top-scoring pair, different from the existing practice of using the raw values of features. As a benefit to using residuals, TSP would be selecting from features largely liberated from the extraneous effects of the covariates. Thus, we would be able to capture a top-scoring pair that discriminates the binary-class outcome through its own efforts without much influence from "known" covariates.

Although the TSP algorithm has mainly been applied to gene-cancer studies, we provide an application of TSP to the novel setting of metabolomics and chronic kidney disease (CKD) in diabetic patients of the Chronic Renal Insufficiency Cohort (CRIC) Study. Recent reviews highlighted key metabolites that differentiated cases of diabetic kidney disease (DKD) from healthy controls [19, 20, 21, 22, 23]. In our previous work, we identified several metabolites that were consistently and significantly reduced in patients with diabetes and CKD when compared to patients with diabetes without CKD [26].

In this article, we implement the existing TSP and K-TSP methods [52, 56], as well as our proposed technique of using the features' residuals, for the selection of top-scoring pairs. We conduct a simulation study to illustrate the implications of using the features' residuals as the input to the TSP algorithm. Furthermore, we demonstrate the application of the TSP and K-TSP methods in identifying top-scoring metabolite-pair classifiers that best discriminate between DKD severity in a study sample of the CRIC Study, using both raw and residual values of the patient's metabolites. Finally, since TSP and K-TSP are binary classification methods, we compare their classification accuracy to popular statis-

tical learning methods, i.e., LASSO (least absolute shrinkage and selection operator) and random forests [24].

3.3 Methods

3.3.1 TSP and K-TSP: Brief review

We first provide a review of the TSP and K-TSP methods. Let $X = \{X_1, X_2, \ldots, X_p\}$ denote the p features for an individual profile. The TSP algorithm [52] identifies the top-scoring feature pair Θ^* among the p features with the maximum absolute difference in the probability of $X_i < X_j$ between two classes of individuals, $C = 1, 2$. In particular, we calculate the discriminant score of all possible feature pairs Θ:

$$\widehat{s}_{ij} = |P(X_i < X_j | C = 1) - P(X_i < X_j | C = 2)|$$

and $\widehat{s}_{ij}$ is maximized from the top-scoring feature pair Θ^*, i.e., $\Theta^* = \arg\max_{(i,j)\in\Theta} \widehat{s}_{ij}$. These conditional probability quantities are estimated using maximum likelihood from the observed sample proportions of the ordering $X_i < X_j$ between both classes. Accordingly, it is sufficient to know the ranks of features within individual profiles to obtain the scores for all feature-pairs $\widehat{s}_{ij}, i, j = 1, 2, \ldots, p, i \neq j$. A feature-pair (i, j) achieves perfect discrimination when $\widehat{s}_{ij} = 1$ and no discrimination when $\widehat{s}_{ij} = 0$. If multiple pairs achieve the top score, ties were broken with a secondary rank-score to select a single top-scoring pair [54].

Classification with TSP amounts to observing the ordering of the two features of the top-scoring pair (i, j) for a future test sample. If $P(X_i < X_j | C = 1) < P(X_i < X_j | C = 2)$, then if we observe $X_i < X_j$ TSP classifies the test sample as class $C = 2$ or if instead observed as $X_i \geq X_j$ TSP classifies as class $C = 1$. Otherwise, if $P(X_i < X_j | C = 1) \geq$

$P(X_i < X_j|C = 2)$, then if we observe $X_i < X_j$ TSP classifies the test sample as class $C = 1$ or if instead observed as $X_i \geq X_j$ TSP classifies as class $C = 2$.

In practice, the results of TSP may be sensitive to perturbations in the training data and a more stable alternative, the K-TSP algorithm [53]. K-TSP defines the set of k disjoint features pairs with the highest scores to be $\Theta_k = \{(i_1, j_1), ..., (i_k, j_k)\}$. The set of k disjoint top-scoring pairs is chosen $\Theta_k^* = \{(i_1^*, j_1^*), ..., (i_k^*, j_k^*)\}$ to maximize the score $\widehat{s}_{i_r j_r}$:

$$\Theta_k^* = \arg\max_{\Theta_k} \sum_{r=1}^{k} \widehat{s}_{i_r j_r}$$

for each value k. From this, we now obtain the optimal value k^* from the set of k values that maximizes the following criterion $\widehat{\tau}_{KTSP}$ motivated by the concept of analysis of variance [56]:

$$\widehat{\tau}_{KTSP}(\Theta_k^*) = \frac{\sum_{r=1}^{k} \widehat{s}_{i_r j_r}}{\sqrt{\widehat{Var}(\sum_{r=1}^{k} I(X_{i_r^*} < X_{j_r^*})|C = 1) + \widehat{Var}(\sum_{r=1}^{k} I(X_{i_r^*} < X_{j_r^*})|C = 2)}}$$

Classification with K-TSP amounts to observing the ordering of the k top-scoring pairs $\{(i_1^*, j_1^*), ..., (i_k^*, j_k^*)\}$ and taking a simple majority voting rule for a test sample. That is, the test sample will be assigned the class receiving the most votes.

We implemented the TSP and K-TSP algorithms from the R package switchbox [57] for our statistical analysis.

3.3.2 Residualizing the Features

In most studies of chronic disease, there are clinical risk factors known to be associated with the outcome of interest. The aforementioned TSP classifiers do not take into account the effects of such variables, $Z = \{Z_1, Z_2, \ldots, Z_q\}$, in selecting the top-scoring feature pairs that best distinguish the binary outcome between two classes of individuals.

In particular, the top-scoring feature pairs may be strongly confounded by these variables and we seek to suppress the effects of Z in the top-scoring pair selection as to capture "cleaner" features for discriminating the outcome between two classes. Thus we aim to identify top-scoring pairs conditional on covariate values, Z:

$$\widehat{s}_{ij}|Z = |P((X_i|Z) < (X_j|Z)|C = 1) - P((X_i|Z) < (X_j|Z)|C = 2)|$$

To operationalize this approach, we propose fitting linear regression models with features X as outcomes and Z as covariates, and use the model residuals as opposed to feature values for selecting the top-scoring pairs. We refer to this process as residualizing which largely decorrelates features, X, from the individual covariates, Z. In particular, the fitted regression model for feature $X_i = \{X_{i1}, X_{i2}, \ldots, X_{iN}\}$, such that $i = 1, 2, \ldots, p$, is defined as

$$\widehat{X}_{ik} = \widehat{\beta}_0 + Z_{k1}\widehat{\beta}_{i1} + Z_{k2}\widehat{\beta}_{i2} + \cdots + Z_{kq}\widehat{\beta}_{iq}$$

for the kth individual, $k = 1, 2, \ldots, N$. We define the residual of the ith feature for the kth individual to be

$$e_{ik} = X_{ik} - \widehat{X}_{ik}$$

in which $e_i = \{e_{i1}, e_{i2}, ..., e_{iN}\}$ is the set of residuals of the ith feature.

Thus, two types of TSP-based methods were developed and compared in our application setting for feature selection and classification accuracy: (1) non-residualized, trained from the raw features data X_i, and (2) residualized, trained from the residuals of features e_i.

3.4 Simulations

3.4.1 Simulation Setup

To describe our residualizing process and its effect on the selection of the top-scoring pair, we first conducted a simulation study. In particular, we generate feature pairs with strong vs no correlation with a "clinical" feature, which itself is highly predictive of the outcome. Sample size for our simulation study was set to $N = 200$ and we defined our "disease" outcome (Y) as binary with values 0 or 1. We consider the simple case of a single "clinical" covariate (Z) also binary with values 0 or 1. Total sample was evenly split between values $Z = 0$ and $Z = 1$, and we defined the high association between Z and Y as the probabilistic relationship $P(Y = 1|Z) = |Z - 0.05|$. Two different sets of bivariate data were generated to illustrate how residualizing changes the scores of these feature-pairs which, in turn, affects whether a set is likely to still be a selected as a top-scoring pair.

The first set generated was bivariate normal (X_1, X_2) conditional on Z:

$$\left(\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right)|_{Z=0} \sim N[\left(\begin{smallmatrix} 0 \\ 5 \end{smallmatrix}\right), \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)]; \quad \left(\begin{smallmatrix} X_1 \\ X_2 \end{smallmatrix}\right)|_{Z=1} \sim N[\left(\begin{smallmatrix} 5 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)]$$

The second set also bivariate normal (X_3, X_4) but conditional on Y and independent of Z:

$$\left(\begin{smallmatrix} X_3 \\ X_4 \end{smallmatrix}\right)|_{Y=0} \sim N[\left(\begin{smallmatrix} 0 \\ 2.5 \end{smallmatrix}\right), \left(\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right)]; \quad \left(\begin{smallmatrix} X_3 \\ X_4 \end{smallmatrix}\right)|_{Y=1} \sim N[\left(\begin{smallmatrix} 2.5 \\ 0 \end{smallmatrix}\right), \left(\begin{smallmatrix} 3 & 0 \\ 0 & 3 \end{smallmatrix}\right)]$$

With features (X_1, X_2, X_3, X_4) and covariate data Z, we calculated the posterior probabilities of Y as $P(Y|X_1, X_2, X_3, X_4, Z)$ using Bayes' theorem, although since (X_1, X_2) is conditionally independent of Y:

$$P(Y|X_1, X_2, X_3, X_4, Z) = P(Y|X_3, X_4, Z) = \frac{P(X_3, X_4|Y, Z) \times P(Y|Z)}{P(X_3, X_4)}$$

These posterior probabilities constituted Bernoulli probabilities, and were used to generate the binary outcome class indicator Y. Finally, we calculate the TSP scores for the raw and residualized (on Z) variants of (X_1, X_2) and (X_3, X_4).

3.4.2 Simulation Results

The results are plotted in Figure 3.1 with the rows corresponding to (X_1, X_2) and (X_3, X_4) and columns to their raw and residualized features. In total, 96 samples had class $Y = 0$ (48%) and 104 samples had class $Y = 1$ (52%). Both classes of Y from the raw (X_1, X_2) data were mostly well separated by the TSP's fixed decision boundary and this raw pair had a score of 0.64. However, the residualized (X_1, X_2) did not achieve nearly the same level of discrimination in the classes of Y by having a much lower score of 0.07. We can attribute this to the raw (X_1, X_2) generated conditional on Z which is strongly associated with Y. Residualizing the raw (X_1, X_2) decorrelates the pair from Z, which substantially decreases the capability of (X_1, X_2) to discriminate between the two classes of Y along TSP's decision boundary. In practice, if the raw values of (X_1, X_2) are identified as the top-scoring pair, and we know this pair to be highly dependent on Z, then residualizing (X_1, X_2) with Z would most likely drop its candidacy as a top-scoring pair and opens the door for another feature-pair to be selected by the TSP algorithm for best discriminating between the classes of Y.

In contrast, residualizing the raw (X_3, X_4) with Z did not drastically affect the feature-pair's capability to distinguish between both classes of Y along the decision boundary based on the small drop in score from 0.38 to 0.33. We can attribute this to the raw (X_3, X_4) generated conditional on Y and independent on Z, which would likely retain (X_3, X_4) as the top-scoring pair even after residualizing with Z. In summary, residualizing captures "cleaner" top-scoring pairs liberated from much of the extraneous influence of

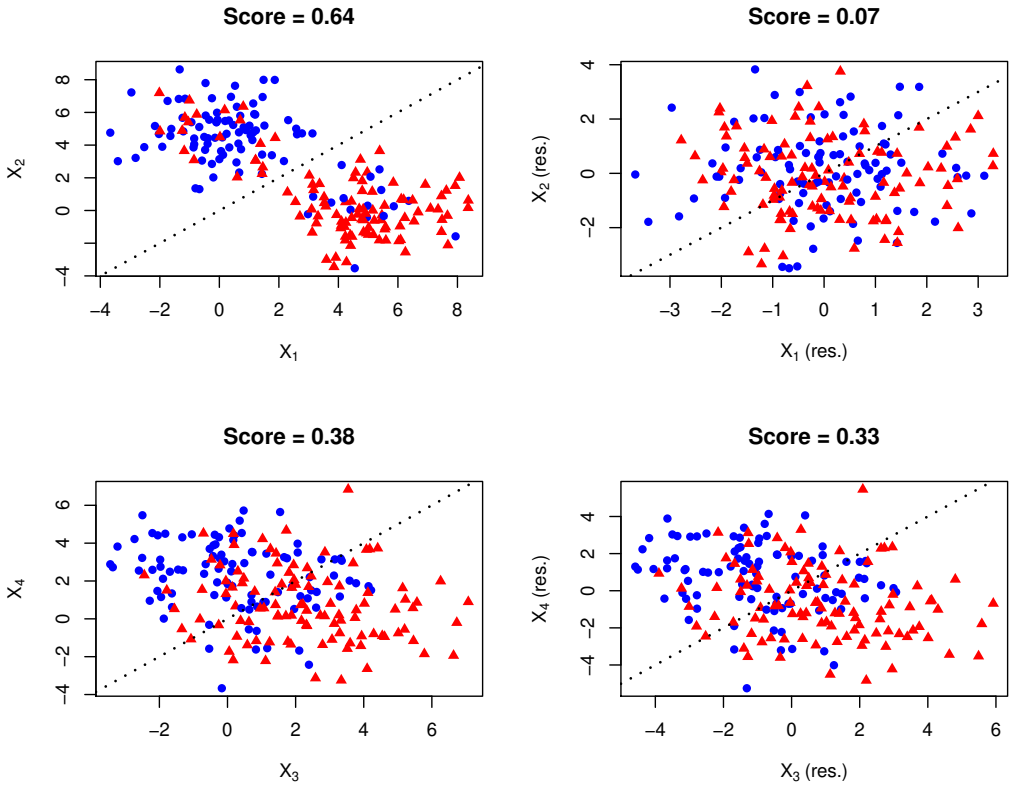

Figure 3.1: Left column: Scatter plots of generated feature pairs from our simulation study ($N = 200$) conditional on our single "clinical" covariate, (X_1, X_2), and independent of our single "clinical" covariate, (X_3, X_4). Right column: Scatter plots of the residualized feature pairs. The two evenly split classes are represented as red and blue and TSP's decision boundary is overlayed on the plots.

covariates, and thus identifies potentially novel markers of outcome.

3.5 Application

3.5.1 CRIC Study description with outcome

Our study cohort comprised a sample of 977 diabetic patients selected for untargeted metabolome profiling in the Chronic Renal Insufficiency Cohort (CRIC) Study. Details on the rationale and design of the parent CRIC Study and this metabolomics sub-

study have been previously published [25, 29, 30, 31]. The study included a racially and ethnically diverse group of adults aged 21 to 74 years with a broad spectrum of kidney disease severity, assessed via estimated glomerular filtration rate (eGFR) between 20 and 70 ml/min/1.73m^2. Sociodemographic information, medical history, medications used in the previous 30 days, anthropometric measures (weight, height), and resting blood pressure were collected from CRIC participants. In addition, blood specimens and 24-h urine samples were also obtained. For our cohort, 45% were white, 44% were female, 57% smoked $>$ 100 cigarettes in lifetime, and 81% reported using angiotensin-converting enzyme (ACE) inhibitors or angiotensin-receptor blockers (ARB). Patients have mean (SD) age 59.9 (9.4) years, body mass index (BMI) 34.2 (7.9) kg/m^2, hemoglobin A1c (HbA1c) 7.6 (1.6)%, mean arterial pressure 89.9 (13.3) mmHg, urine albumin 0.9 (1.8) mg/day, serum creatinine 1.9 (0.6) mg/dL, and estimated glomerular filtration rate (eGFR) 40.6 (11.2) ml/min/1.73m^2. Furthermore, they were sampled across CKD stages G2 (eGFR 60-70), G3a (eGFR 45-60), G3b (eGFR 30-45) and G4 (eGFR 20-30).

Our outcome of interest for the current analysis is a binary indicator of early-stage DKD (stage G2-3b, N=777) versus advanced-stage DKD (stage G4, N=200). Patients with early-stage DKD had mean (SD) eGFR 44.68 (8.53) ml/min/1.73m^2, while those with advanced-stage DKD had mean (SD) eGFR 24.82 (3.6) ml/min/1.73m^2.

3.5.2 Metabolomics

Untargeted metabolome profiling in urine was performed for our 977 CRIC samples. Assay procedures have been described previously [25, 58], but we briefly recapitulate key points here for completeness. Aliquots of urine samples stored at -80 °C and limited to less than 3 free thaw cycles were used in this study. Quantification of relative ion abundance was carried out with an MPS3xt autosampler (Gerstel) coupled to an Agilent 6550

Q-TOF mass spectrometer (Agilent Technologies) by non-targeted flow injection analysis as described previously [58]. Profile mass spectra were recorded in 4Ghz acquisition mode from 50 to 1000 m/z in negative ionization mode. Raw mass spectrometry data was normalized based on creatinine ion abundances. Final annotation of approximately 15k ions common to all datasets were done based on accurate mass comparison using 1 mDa mass tolerance against Human Metabolome Database HMDBv4.0 assuming single deprotonation. A single ion could annotate multiple metabolites resulting in ambiguities in the assignments. Therefore, we shall refer to our features as metabolite ions for our study.

Following prior protocols, we used stringent statistical criteria, based on Pearson and Spearman correlation, intraclass correlation, and coefficients of variation applied to technical replicates, to eliminate noisy metabolite ions, and identified a final set of 698 annotated metabolite ions for our analysis.

3.5.3 TSP and K-TSP results on CRIC Study with and without residualizing

As an application demonstration, we apply the TSP and K-TSP algorithms to our CRIC study sample to identify metabolite-pairs that best discriminate between DKD stages, with and without residualizing the metabolite ions. The residuals are obtained from the fitted linear regression models with metabolitie ions as outcomes and the demographic and clinical variables age, race, sex, smoked $>$ 100 cigarettes in lifetime, BMI, HbA1c, mean arterial pressure, urine albumin, serum creatinine, and ACE Inhibitor or ARB use as covariates. These covariates are known to be associated with DKD severity. The results of TSP and K-TSP algorithms applied to our CRIC samples are displayed in Figures 3.2 and 3.3.

From among the raw metabolite ions, the TSP algorithm identified the metabolite-

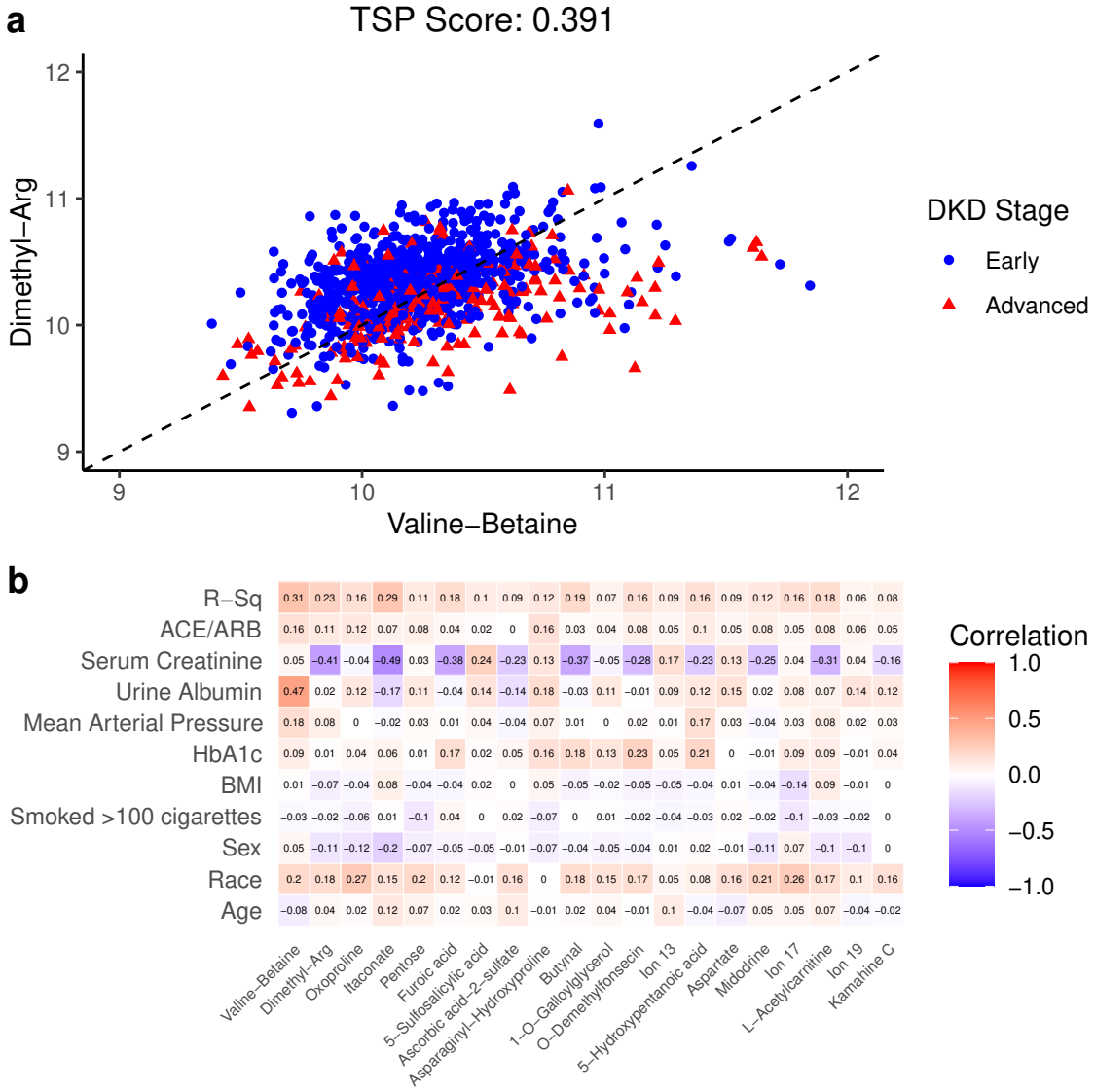

Figure 3.2: (a) Scatter plot for the top pair of raw metabolite ions selected by the TSP algorithm along with TSP's decision boundary. The axes are metabolite ion abundances that were creatinine normalized and natural log transformed. Patients had either early-stage DKD (stage G2-3b, $N = 777$) or advanced-stage DKD (stage G4, $N = 200$). (b) Heatmap correlation matrix of clinical variables vs raw metabolite ions selected by the K-TSP algorithm. Single ion can annotate to multiple metabolites, which resulted in ambiguity in assignments.

Ion 13: 3,6-Dihydro-4-(4-methyl-3-pentenyl)-1,2-dithiin

Ion 17: 13,14,15-trihydroxy-9-oxo-8,17-dioxatetracyclo[8.7.0.0^{2},^{7} .0^{11},^{16}]heptadeca-1(10),2(7),3,5,11,13,15-heptaen-5-yl acetate

Ion 19: 2-Phenylethyl beta-D-glucopyranoside

pair (Valine-Betaine, Dimethyl-Arg) to be the top-scoring pair (score: 0.391) in Figure 3.2a. As mentioned earlier, a single ion could be annotated as multiple metabolites; the selected top-scoring pair contained an ion that could be annotated as Valine or Betaine. Hence, to note this ambiguity we will refer to this feature as Valine-Betaine. Here, the TSP's decision rule is that if a test patient's observed metabolite ion ordering is Valine-Betaine $<$ Dimethyl-Arg then the patient will be classified as having early-stage DKD and the reversed metabolite ion ordering for advanced-stage DKD. Applying the K-TSP algorithm gave us a total of 10 metabolite-pairs, including (Valine-Betaine, Dimethyl-Arg), with score range 0.259-0.391. These 20 metabolite ions are listed in the correlation heatmap with the clinical variables used for the residualizing process in Figure 3.2b. The metabolite ions Valine-Betaine and Dimethyl-Arg have the largest variation explained by the clinical variables with R^2 values 0.31 and 0.23, respectively. Notably, Valine-Betaine has a relatively high correlation with urine albumin and likewise for Dimethyl-Arg and serum creatinine, which likely indicates that in this application, TSP selected metabolites ions with a moderate-high correlation with known clinical markers of kidney disease.

Our simulation study demonstrated that residualizing a raw top-scoring pair that is highly correlated with covariates significantly drops its score; therefore, yielding another candidate to be the top-scoring pair. Hence, we residualize our raw metabolomics data, and the TSP algorithm instead identified the metabolite-pair (Pipazethate, Octaethylene glycol) to be the top-scoring pair (score: 0.25). This TSP's decision rule is that if a test patient's observed metabolite ion ordering is Pipazethate $<$ Octaethylene glycol then the patient will be classified as having early-stage DKD and the reversed metabolite ion ordering for advanced-stage DKD in Figure 3.3a. Applying the K-TSP algorithm gave us a total of 9 metabolite-pairs, including (Pipazethate, Octaethylene glycol), with score range 0.158-0.25. These 18 metabolite ions are listed in the correlation heatmap with the clinical variables used for the residualizing process in Figure 3.3b. The R^2 values for these

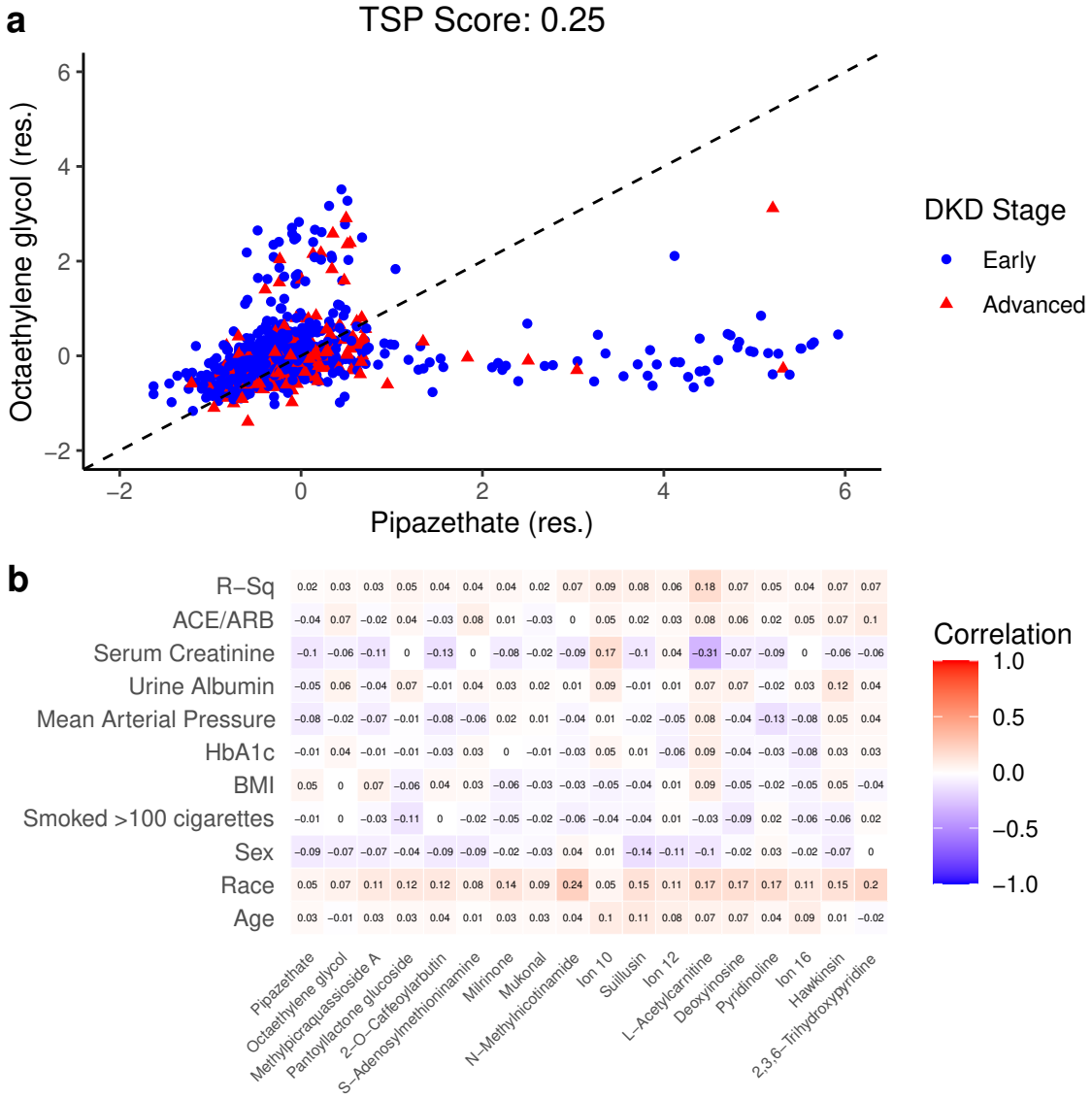

Figure 3.3: (a) Scatter plot for the top pair of residualized metabolite ions selected by the TSP algorithm along with TSP's decision boundary. The axes are residuals of metabolite ion abundances that were creatinine normalized and natural log transformed. Patients had either early-stage DKD (stage G2-3b, $N = 777$) or advanced-stage DKD (stage G4, $N = 200$). (b) Heatmap correlation matrix of clinical variables vs the raw values of residualized metabolite ions selected by the K-TSP algorithm. Single ion can annotate to multiple metabolites, which resulted in ambiguity in assignments.

Ion 10: 3,6-Dihydro-4-(4-methyl-3-pentenyl)-1,2-dithiin

Ion 12: [4-(5-hydroxy-7-methoxy-8-methyl-4-oxo-4H-chromen-3-yl)-2-methoxyphenyl]oxidanesulfonic acid

Ion 16: alpha-L-Rhamnopyranosyl-(1 → 3)-alpha-D-galactopyranosyl-(1 → 3)-L-fucose

18 metabolite ions are much smaller than those of the 20 metabolite ions selected under the raw metabolomics setting. Here, Pipazethate and Octaethylene glycol do not have relatively high correlation values with the clinical variables indicating that these metabolite ions are not serving as proxies for clinical markers of kidney disease.

Thus, in the metabolite-DKD context, if we obtained a top-scoring pair from the raw metabolomics data and the features are highly correlated with our clinical variables, then we would very likely acquire a different top-scoring pair after residualizing.

3.5.4 Comparison to other methods: LASSO and random forests

Methods

To evaluate the relative prediction performance of the TSP and K-TSP algorithms for DKD severity, we compared their classification accuracy to that of LASSO and random forests. The LASSO model was tuned to the regularization parameter that minimizes mean 10-fold cross-validated misclassification error for feature selection among the 698 metabolite ions [59]. The random forests model was fitted using Breiman's algorithm, growing 500 trees and randomly sampling 26 metabolite ions as candidates at each split [60]. We implemented the LASSO and random forest methods using the R packages glmnet and randomForest, respectively.

Several classification accuracy measures were used for comparing TSP, K-TSP, LASSO, and random forests: (1) overall accuracy, i.e., overall proportion correctly classified, (2) sensitivity, i.e., proportion correctly classified among those with advanced-stage DKD, (3) specificity, i.e., proportion correctly classified among those with early-stage DKD, (4) balanced accuracy, i.e., average of sensitivity and specificity, (5) positive predictive value (PPV), i.e. proportion that truly have advanced-stage DKD among those classified with advanced-stage DKD, and (6) negative predictive value (NPV), i.e. proportion

that truly have early-stage DKD among those classified with early-stage DKD. To estimate their variabilities, we conducted one-hundred iterations of 5-fold cross-validation for each of these accuracy measures. One fold is held out as a test set and our models are trained on the remaining four folds with our accuracy measures calculated on the test set. The four folds in the training data will also each serve as a test set, which would result in accuracy measures from all five folds. The averages of these accuracy measures across all five folds are our 5-fold cross validated estimates. Since the partition of the five folds varies for each iteration, TSP, K-TSP, LASSO, and random forests may select different metabolite predictors at each iteration.

Results

Results for overall accuracy, sensitivity, specificity, and balanced accuracy of early-stage DKD vs advanced-stage DKD for our statistical methods with metabolite ion predictors are displayed in Figure 3.4. TSP and K-TSP did not have higher median cross-validated overall accuracy (0.649-0.728) compared to that of LASSO and random forests (0.793-0.823) with either raw or residualized metabolite ion predictors. However, both LASSO and random forests displayed extremely poor sensitivity and high specificity, which indicates that the overall accuracy for these two methods is driven by classifying an overwhelmingly large number of patients as having early-stage DKD, regardless of their observed DKD stage. In contrast, TSP and K-TSP achieved a more balanced tradeoff of sensitivity and specificity and had values closer to their overall accuracy. Notably, we also have imbalanced classes with 79.5% of our patients with early-stage DKD and we examine balanced accuracy for our methods which is preferred over overall accuracy for class imbalance data. Here, TSP and K-TSP did have higher median cross-validated balanced accuracy (0.566-0.689) compared to that of LASSO and random forests (0.5-0.652) for the raw or residualized metabolite ion predictors cases. For positive predictive value, TSP and K-TSP

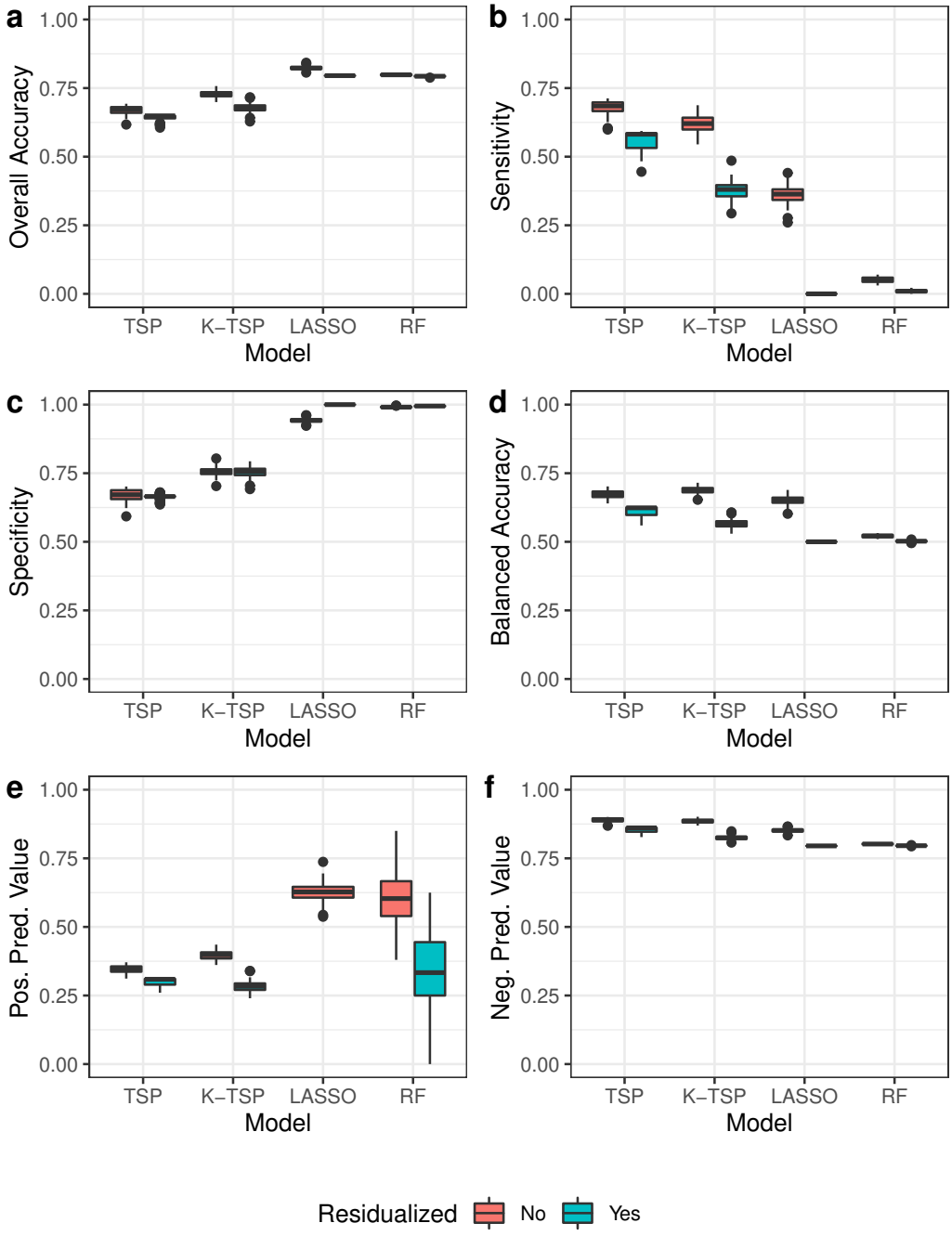

Figure 3.4: Box plots of model prediction performance for DKD stage: 100 repeats of 5-fold cross-validated (a) overall accuracy, (b) sensitivity, (c) specificity, (d) balanced accuracy, (e) positive predictive value, and (f) negative predictive value.
(K-)TSP: (K) Top-Scoring Pair(s)
LASSO: Least Absolute Shrinkage and Selection Operator
RF: Random Forests

did not perform better than LASSO or random forests when using raw metabolite ions; however, random forests displayed relatively high variability in its cross-validated values, and when using residualized metabolite ions, LASSO did not predict advanced-stage DKD even once for any patient in all 100 iterations of 5-fold cross-validation (hence the absence of its boxplot in Figure 3.4e). Finally, for negative predictive value, all methods performed reasonably well with TSP and K-TSP taking the lead.

In essence, TSP and K-TSP displayed comparatively good classification for early-stage DKD patients from their specificity and NPV performances. Furthermore, residualized metabolite ions are a valid option as predictors for our statistical methods in classification for early-stage DKD patients. In particular, specificity and NPV did not show a notable decrease in performance going from using raw metabolite ions to their residualized variants while taking into account that the residualized metabolite ions are features liberated from much of the known important clinical variables.

3.6 Conclusion and discussion

The large influx of high-dimensional biomolecular data brought about by cutting-edge high-throughput technologies, particular in 'omics research, exhibits inordinate potential for improving our understanding of the link between biological profiles and clinical diseases. There is a critical need for accurate and robust decision rules based on these biological data that are easily interpretable for translation to future clinic use. We focused on the TSP (or K-TSP) algorithm that identifies a single pair (or set of pairs) of features that best discriminates between two classes of interest among all possible feature-pairs. TSP identifies the feature-pair for which an observed ordering of the two features is very common in one class and rare in the other, which allows the top-scoring pair to be interpreted as a "biological switch" from one class to another based on feature ordering. Thus, the TSP

approach by construction aims to offer insights into underlying mechanisms of disease, a salient advantage over other statistical and machine learning methods.

A major potential of 'omics studies is the possibility to discover "new" insights into disease mechanisms and discrimination; hence, interest usually lies in identifying markers that are associated with disease status after adjusting for known clinical factors. Previous studies utilizing TSP methods have not accounted for possible covariates in the selection of the top-scoring pairs. We provided a novel extension of the TSP algorithm for removing much of the extraneous effects that covariates could have on the features, so as to capture a top-scoring pair largely independent of covariates. We implemented a residualizing process, and demonstrate via simulation and application that using the residuals from a regression of features on covariates known to be highly associated with the outcome, and then applying the TSP algorithm to these residuals, could identify potentially novel pairs compared to simply using the raw (unresidualized) features. In fact in our data application, the top-scoring pairs using the raw features were valine (or betaine) and dimethyl-arginine, known amino acids linked to albuminuria [26, 61], a potent risk factor for CKD. Thus, this pair could simply reflect a known underlying CKD marker, as seen in Figure 3.2b. On the other hand, the top-scoring pair from the residualized analysis were pipezethate [62], a non-narcotic antitussive agent, and octaethylene glycol [63, 64, 65, 66], a member of the class of polyethylene glycols, found in osmotic laxatives. Thus these residualized metabolite ions, are potentially new markers, and could offer insights into drug metabolism and CKD. Notably, the idea of using residuals to adjust for covariates has been considered in classical discriminant analysis [67, 68, 69, 70], and more recently for decision trees [71], but to our knowledge, our use of residuals for the TSP algorithm is novel.

TSP and K-TSP are classification methods, hence we evaluated and compared their classification accuracy of DKD stage using metabolite ion predictors to more conventional statistical learning methods, i.e., LASSO and random forests. While TSP and K-TSP did

not perform better than LASSO and random forests by the overall accuracy metric, we have imbalanced classes with over three-fourths of our samples with early-stage DKD. By the balanced accuracy metric, which accounts for imbalanced classes, TSP and K-TSP performed better than the others. Both TSP and K-TSP performed moderately and well in specificity and negative predictive value, respectively, which suggests that these methods can accurately identify a healthier or less severe disease group. Furthermore, residualized metabolite ions yielded similar specificity and negative predictive value results for TSP and K-TSP.

We acknowledge limitations and future directions of our work. First, we looked at a binary class outcome since the TSP and K-TSP algorithms were developed as binary classification methods. However, methods exist for multi-class classification [72, 73], which could be easily extended to our setting. In particular, for our DKD setting, multi-class would allow us to use residualized features to discriminate between different levels of kidney (dys)function among diabetic patients; given the relatively small cell sizes we leave this to future work using a larger cohort. Second, our stringent statistical criteria eliminated noisy compounds and resulted in 698 annotated metabolite ions as candidates for the identification of the top-scoring pairs which could have missed biomarkers significant to characterizing DKD stage. Thus, a future aim is to include metabolite ions highly associated with DKD and part of pathways informative of therapeutic targets for DKD to administer biological understanding to top-scoring pair selection. Third, our residualizing process involves the use of linear regression to obtain the residuals of the features and more complex statistical models could be fitted to obtain the residuals. However, the linear regression models have simple implementation and have the residuals orthogonal to the covariates, which is beneficial for capturing cleaner features.

In summary, in this work we extended the TSP-algorithms to account for clinical covariates, via a simple, easy to implement residualizing process. The TSP and K-TSP

algorithms have the advantage of deriving parameter-free decision rules that best discriminate the class outcome of interest by examining just the ordering of feature-pairs. Thus, they yield parsimonious classifiers that are biological interpretable in the 'omics setting. We demonstrated the utility of our residualizing approach for TSP via simulation and real application to the novel metabolite-DKD context. The residualized metabolite ions bought us "cleaner" top-scoring pairs, uncorrelated to clinical covariates, which we could identify as biomarkers classifying disease stage outcome on their own merit. These metabolite ions could serve to motivate hypotheses for future studies, for instance, laboratory studies could further examine the selected pairs to confirm or refute the order reversals in the disease vs non-disease states.

3.7 Acknowledgements

Chapter 3, in part is currently being prepared for submission for publication of the material. Kwan, Brian; Fink, Jeff; Fuhrer, Tobias; He, Jiang; Hsu, Chi-yuan; Messer, Karen; Montemayor, Daniel; Nelson, Robert; Pu, Minya; Ricardo, Ana; Rincon-Choles, Hernan; Shah, Vallabh; Ye, Hongping; Zhang, Jing; Sharma, Kumar; Natarajan, Loki. "Identifying metabolite-pair markers for chronic kidney disease stage classification in diabetic patients: results from applying the top-scoring pairs algorithm to the Chronic Renal Insufficiency Cohort (CRIC) Study". The dissertation author was the primary investigator and author of this material.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650112. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Chapter 4

Inference and Prediction using Functional Principal Components Analysis: Application to Diabetic Kidney Disease Progression in the Chronic Renal Insufficiency Cohort (CRIC) Study

4.1 Abstract

Patients with diabetic kidney disease (DKD) are at high risk for kidney failure and estimated glomerular filtration rate (eGFR) trajectories are markers for DKD progression. We propose applying the functional principal components analysis (FPCA) to modeling

eGFR trajectories while overcoming nonlinear, sparse, and irregularly spaced time series data. Furthermore, FPCA is a novel approach to exploring dominant modes of eGFR variation among diabetic patients of different albuminuria groups, a clinical subgroup of interest in the renal disease context. In a cohort of 2641 participants with diabetes and up to 15 years of annual follow-up from the Chronic Renal Insufficiency Cohort (CRIC) study, we detected novel and different dominant modes of variation and patterns of DKD progression among albuminuria groups. To determine whether fitting a single overall model or fitting separate albuminuria group-specific models is more optimal for modeling eGFR trajectories, we developed a goodness-of-fit procedure based on cross-validation prediction error. Our findings indicate that both model approaches have their advantages in different settings that is mainly linked to the trade-off between parsimony and complexity.

4.2 Introduction

Diabetes mellitus is the leading cause of chronic kidney disease (CKD) [1, 2, 3, 4, 5] and patients with diabetic kidney disease (DKD) progression are at increased risk for kidney failure, which would require treatment by kidney transplant or dialysis. Estimated glomerular filtration rate (eGFR) is a ubiquitous marker for kidney function and previous studies examined change in eGFR for assessing change in kidney function [74, 75], which make eGFR trajectories natural markers for DKD progression. Linear mixed effect models are often used to estimate change in eGFR; however, nonlinear trends may exist. Estimation is further complicated by observing sparse or irregular spaced eGFR time series data as depicted in Figure 4.1.

We propose the functional principal components analysis (FPCA) approach to predicting long-term trajectories while accounting for complexity in curve estimation, i.e., nonlinearity, sparsity, and irregularity. FPCA translates the omnipresent dimension reduc-

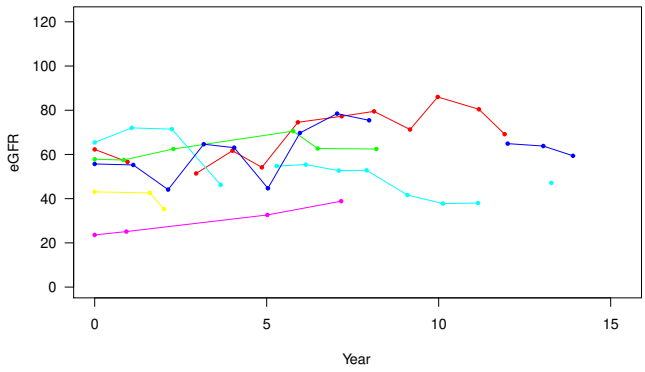

Figure 4.1: Various patterns of observed eGFR trajectories for patients with diabetes in the Chronic Renal Insufficiency Cohort (CRIC) study, including sparse or irregular spaced data.

tion method, principal components analysis (PCA) [76] from the multivariate data setting to the functional data setting, thereby allowing it to investigate for dominant modes of variation and project infinite-dimensional curves into finite-dimensional vector scores. A brief introduction to FPCA, as well as the general area of functional data analysis, is described in Ramsay and Silverman [77]. Moreover, a thorough literature review on the development of FPCA is elaborated in detail in Wang et al. [78]. Worthy of note is that the application of FPCA to the sparse functional data setting has been subjected to much investigation and development since the nineties [79, 80, 81, 82, 83, 84, 85, 86].

In a recent study by Dong et al. [87], FPCA was used to uncover major sources of variations of eGFR trajectories of kidney transplant recipients. Our work is different in that we applied FPCA to predict the long-term eGFR trajectories and detect the dominant modes of eGFR variation of diabetic patients. Furthermore, a question of interest is if longitudinal eGFR patterns, i.e., mean and correlation functions, vary between key clinical subgroups. Albuminuria, excess of albumin in the urine, is an established biomarker for renal disease and, in conjunction with eGFR, is often used in the classification of patient

CKD stage [88]. As such, we investigated for differences in longitudinal eGFR patterns between different albuminuria groups in CKD. Furthermore, a follow-up question is whether an overall model, trained using data from all diabetic patients irrespective of albuminuria groups, is sufficient for accurately predicting the long-term eGFR trajectory within specific albuminuria groups. If not, we consider multiple albuminuria group-specific models, each fitted using data from only patients of one particular group, to prospectively predict eGFR trajectories for new subjects of the same group. To decide whether this group-level approach is needed, we developed a procedure to compare the goodness-of-fit between the overall and group-specific models via prediction error.

The breakdown of the organization of this paper is in the following sections. Section 4.3 presents our methods. This includes the FPCA approach, tests of equality for the mean and correlation functions, comparison of goodness-of-fit between models, our CRIC study cohort, assignment of albuminuria groups, and computational tools. Section 4.4 describes our statistical analysis results in detail. Lastly, Section 4.5 is a discussion of our findings, limitations, and future directions.

4.3 Methods

4.3.1 Functional Principal Components Analysis (FPCA)

Let $Y_{ij} = X_i(t_{ij}) + \epsilon_{ij}$ be the observed outcome at time t_{ij}, where $X_i(\cdot)$ is the measurement-free outcome for subject i at time t_{ij} and ϵ_{ij} are measurement errors assumed to be identically and independently distributed normal with mean zero and variance σ^2 such that $i = 1, 2, \ldots, N$ and $j = 1, \ldots, m_i$.

We model the individual trajectories as a smooth random function $X(t)$ with unknown mean function $\mu(t)$ and covariance function $\Sigma(s,t)$, where $s, t \in T$ and T is a

bounded and closed time interval. Let $X_i(t)$ be the outcome trajectory for the ith individual and t be years of follow-up. Under the Karhunen-Loeve expansion, the ith individual's trajectory can be expressed as

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

where $\phi_k(t)$ is the kth functional principal component (FPC) and ξ_{ik} is the associated kth FPC score for the ith individual. The individual scores ξ_{ik} are uncorrelated random variables with mean zero and variance λ_k, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\sum_k \lambda_k < \infty$. The covariance function $\Sigma(s,t)$ can be defined as

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

We briefly recapitulate the workflow of the PACE algorithm by Yao et al. [83] to estimate these model components. The mean function $\widehat{\mu}(t)$ is estimated using a local linear smoother that aggregates data from all individuals. The smooth covariance is estimated from the individual "raw" covariances. Let $\Sigma_i(t_{ij}, t_{il}) = (Y_{ij} - \widehat{\mu}(t_{ij}))(Y_{il} - \widehat{\mu}(t_{il}))$ be the ith individual's raw covariance and it can be seen that $E(\Sigma_i(t_{ij}, t_{il})) = Cov(X(t_{ij}), X(t_{il})) + \sigma^2\delta_{jl}$, where $\delta_{jl} = 1$ if $j = l$ else 0. Hence, the off-diagonal elements of the individual raw covariances $\Sigma_i(t_{ij}, t_{il})$ are used for estimating the smooth covariance $\widehat{\Sigma}(s,t)$. Since the covariance of $X(t)$ achieves its highest values along the diagonal, a local linear fit is used along the direction of the diagonal while a local quadratic fit is used along the direction orthogonal of the diagonal to approximate the surface. The estimates of the FPCs (eigenfunctions), $\widehat{\phi}_k$, and its eigenvalues, $\widehat{\lambda}_k$, are solutions to the eigenequations

$$\int_T \widehat{\Sigma}(s,t)\widehat{\phi}_k(s)ds = \widehat{\lambda}_k\widehat{\phi}_k(t)$$

satisfying the constraints $\int_T \widehat{\phi}_k(t)^2 dt = 1$ and $\int_T \widehat{\phi}_k(t)\widehat{\phi}_m(t)dt = 0$ for $m < k$. The FPCs are estimated by discretizing the smooth covariance $\widehat{\Sigma}(s,t)$. The estimated kth FPC score for the ith individual is acquired through conditional expectation

$$\widehat{\xi}_{ik} = \widehat{E}(\xi_{ik}|Y_i) = \widehat{\lambda}_k \widehat{\phi}_{ik}^T \widehat{\Sigma}_{Y_i}^{-1}(Y_i - \widehat{\mu}_i)$$

such that $Y_i = (Y_{i1}, \ldots, Y_{im_i})^T$. The components $\widehat{\phi}_{ik}$ and $\widehat{\mu}_i$ are vector variants of $\widehat{\phi}_k(t)$ and $\widehat{\mu}(t)$, respectively, evaluated on a grid of time points $t_{ij}, j = 1, \ldots, m_i$. Finally, $\widehat{\Sigma}_{Y_i} = \widehat{\Sigma} + \widehat{\sigma}^2 I_{m_i}$ is also obtained by evaluating $\Sigma(s,t)$ on the same grid of time points. Noteworthy, the PACE method gives estimates for the best prediction of these individual FPC scores.

Since the outcome trajectory is often well approximated by the top K FPCs and their associated scores, we select K as the number of FPCs that explained at least 95% of the total variance in the outcome of interest. Compiling the above altogether for the PACE algorithm, we obtain the predicted outcome trajectory for the ith individual as

$$\widehat{X}_i(t) = \widehat{\mu}(t) + \sum_{k=1}^{K} \widehat{\phi}_k(t) * \widehat{\xi}_{ik}$$

4.3.2 Testing equality of mean functions

We provide a simplified overview of Gorecki-Smega et al.'s permutation test [89] based on a basis function presentation to test the equality of mean functions between G groups. The global null hypothesis and its alternative are

$$H_0 : \mu_1(t) = \mu_2(t) = \cdots = \mu_G(t) \text{ vs. } H_1 : \exists\, u \neq v \text{ s.t. } \mu_u(t) \neq \mu_v(t),$$

respectively. Suppose that we have individual smooth random functions $X_{gi} \in L_2(T)$ indexed by G groups, where $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$ such that $\sum_{g=1}^{G} n_g = N$.

The trajectories $X_{gi}(t)$ can be represented as a linear combination of orthonormal basis functions $\{\psi_l\}$ of $L_2(T)$, e.g., Fourier,

$$X_{gi}(t) = \sum_{l=1}^{L} c_{gil} * \psi_l(t)$$

where $t \in T$ and c_{gil} are random variables with finite variance. Defining $\boldsymbol{\psi}(t) = (\psi_1(t), \psi_2(t), ..., \psi_L(t))^T$ and $\mathbf{c}_{gi} = (c_{gi1}, c_{gi2}, \ldots, c_{giL})^T$, we can represent the individual, sample group mean, and sample grand mean trajectories as

$$X_{gi}(t) = \mathbf{c}_{gi}^T \boldsymbol{\psi}(t), \quad \overline{X}_g(t) = \frac{1}{n_g} \sum_{i=1}^{n_g} \mathbf{c}_{gi}^T \boldsymbol{\psi}(t), \quad \overline{X}(t) = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \mathbf{c}_{gi}^T \boldsymbol{\psi}(t),$$

respectively. The F-test statistic for the one-way ANOVA problem can be adapted for the functional data setting as

$$FP = \frac{\frac{1}{G-1} \sum_{g=1}^{G} n_g ||\overline{X}_g - \overline{X}||_2^2}{\frac{1}{N-G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} ||X_{gi} - \overline{X}_g||_2^2}$$

where $||f||_2^2 = \int_T f^2(t)dt$ for $f \in L_2(T)$.

Here, FP serves as the test statistic for a permutation-based p-value in testing the global null hypothesis. Gorecki-Smaga et al. [89] noted that the statistic FP can be written in a form less computationally burdensome for computer programs as

$$\frac{\frac{1}{G-1}(a-b)}{\frac{1}{N-G}(c-a)}$$

where

$$a = \sum_{g=1}^{G} \frac{1}{n_g} \mathbf{1}_{n_g}^T \mathbf{C}_g^T \mathbf{C}_g \mathbf{1}_{n_g}, \quad b = \frac{1}{N} \sum_{g=1}^{G} \sum_{h=1}^{G} \mathbf{1}_{n_g}^T \mathbf{C}_g^T \mathbf{C}_h \mathbf{1}_{n_h}, \quad c = \sum_{g=1}^{G} tr(\mathbf{C}_g^T \mathbf{C}_g)$$

here $\mathbf{C}_g = (\mathbf{c}_{g1}, \mathbf{c}_{g2}, ..., \mathbf{c}_{gn_g})$ and $\mathbf{1}_{n_g}$ is a vector of ones with length n_g. As such, the statistic FP only depends on the coefficient vectors $\mathbf{c}_{gi}$ and not on the basis functions ψ_l. Furthermore, we can see that random permutations of the trajectories $X_{gi}(t)$ will only change the value of a. We set 1000 permutation replicates for this test and the significance level at 5%.

4.3.3 Testing equality of correlation functions

To test the equality of correlation functions between G groups, we proceed in two steps. First, we center the individual trajectories by subtracting them by their group mean trajectories, estimated by taking the sample means of the FPCA-predicted eGFR at each time grid point (years) of the group-specific patients and connecting these sample means to form that group's trajectory. Second, we scale the individual trajectories by dividing them by the square root of the diagonal of the smooth covariance estimates (standard deviations) from our overall model. Finally, we apply the multiple-group permutation test developed by Cabassi et al. [90] to test the equality of the covariance functions of our standardized trajectories between G groups. The application of this test is feasible since the covariance of our standardized trajectories is equivalent to the correlation of our un-standardized trajectories.

The global null hypothesis and its alternative are

$$H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_G \text{ vs. } H_1 : \exists\, u \neq v \text{ s.t. } \Sigma_u \neq \Sigma_v,$$

respectively. The permutation test procedure combines the G(G-1)/2 partial tests into a single global test by the non-parametric combination algorithm of Pesarin and Salmaso [91]. The partial test statistics are evaluated as the distances between the covariance functions of two groups, for a pre-defined distance function. Like the previous notation, let

X_{gi} be the outcome trajectory for the ith patient in the gth group where $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$. The permutation test procedure is described in detail by Cabassi et al. [90]:

1. Center individual trajectories by subtracting them from their group mean trajectories, estimated by taking the sample means of the individual eGFR values at each time grid point $i = 1, \ldots, n_g$, i.e. $X_{gi}^{(0)} = X_{gi} - \overline{X}_g$.

2. Let $\mathbf{T}^{(0)}$ be the vector containing the pairwise distances between the sample covariance functions of the centered groups $d(\widehat{\Sigma}_u^{(0)}, \widehat{\Sigma}_v^{(0)})$, for all $1 \leq u < v \leq G$.

3. For $b = 1, 2, \ldots, B$, conduct random permutations of the data group labels and compute $\mathbf{T}^{(b)}$, the vector containing distances between the sample covariance functions of two groups in the permuted dataset, $d(\widehat{\Sigma}_u^{(b)}, \widehat{\Sigma}_v^{(b)})$, for all $1 \leq u < v \leq G$.

4. Compute the estimated pairwise p-values of the test as $\widehat{p}_{u,v}(d) = \frac{\sum_{b=1} \mathbf{1}[d(\widehat{\Sigma}_u^{(b)}, \widehat{\Sigma}_v^{(b)}) \geq d]}{B}$, for $d = d(\widehat{\Sigma}_u^{(0)}, \widehat{\Sigma}_v^{(0)})$.

5. Aggregate the pairwise p-values using the combining function ω by Pesarin and Salmaso (2010) to form the global test statistic $T_\omega^{(0)} = \omega(\widehat{p}_{1,2}, \widehat{p}_{1,3}, ..., \widehat{p}_{G,G-1})$.

6. For the $b = 1, 2, \ldots, B$ random permutations, compute the bth combined test statistic as

$$T_\omega^{(b)} = \omega(\widehat{p}_{1,2}^{(b)}, \widehat{p}_{1,3}^{(b)}, ..., \widehat{p}_{G,G-1}^{(b)})$$

   where $\widehat{p}_{u,v}^{(b)} = \widehat{p}_{u,v}(d(\widehat{\Sigma}_u^{(b)}, \widehat{\Sigma}_v^{(b)}))$.

7. Compute the estimated global p-value of the combined test

$$\widehat{p}_\omega = \frac{\sum_b \mathbf{1}[T_\omega^{(b)} \geq T_\omega^{(0)}]}{B}.$$

8. Reject H_0 if $\widehat{p}_\omega \leq \alpha$.

Based on the simulation studies and application results of Cabassi et al. [90], we used the square root distance for $d(\cdot, \cdot)$ and the max T combining function for T_ω. Here, the square root distance can be defined as the square root mapping of two covariance operators Σ_1 and Σ_2 to the space of Hilbert-Schmidt operators

$$d(\Sigma_1, \Sigma_2) = ||\Sigma_1^{1/2} - \Sigma_2^{1/2}||_{HS}$$

In the case of imbalance between groups, the permutation test procedure would not yield accurate partial p-values when performing permutations for the whole dataset. Therefore, we conducted paired permutations, or permutation tests for each pair of groups independently. Here, the partial tests would be exact, which would allow for two-sample inference of covariance functions. We set 1000 permutation replicates for this test and the significance level at 5%.

4.3.4 Comparing goodness-of-fit between models

It is of interest to determine if an overall model, trained on all individuals irrespective of subgroup (e.g., albuminuria category), is sufficient for predicting the long-term outcome trajectory for individuals. If not, we consider separate group-level models, fitted using only individuals belonging to the same group, to prospectively predict group-specific outcome trajectories for new individuals.

To formally test whether the group-level approach is needed, we propose a goodness-of-fit procedure, inspired by 5-fold cross-validation, to compare the prediction error of our overall model and multiple group-level models. Specifically,

1. Divide the overall cohort of individuals into 5 folds of approximately equal size, ensuring that each fold contains approximately the same proportions of individuals in each group as that of the overall cohort.

2. For the overall model, treat the 1st fold as the test set, and fit the model on the other 4 folds, the training data.

3. The model will predict the trajectories for individuals in the test set and we calculate the average curve squared error for the ith test individual as

$$ACSE_i = \sum_{j=1}^{n_i} \frac{[X_{ij} - \widehat{X}_i(t_{ij})]^2}{n_i}$$

where

- n_i is the number of outcome observations for ith individual
- X_{ij} is the observed outcome for the jth observation of the ith individual
- $\widehat{X}_i(t_{ij})$ is the predicted outcome value at t_j where t_j is the time grid point closest in proximity to the individual's jth observation time. The (FPCA) predicted outcome $\widehat{X}_i(t_{ij})$ is calculated as the sum of both the trained model's mean function and the product of the trained model's FPCs and predicted FPC scores for the ith individual.

4. The mean average curve squared error $MACSE_d$ is then computed as the arithmetic mean of $ACSE_i$ values of the test set.

5. We repeat Steps 3-5 for each of the other 4 folds as the test set.

6. Our goodness-of-fit test statistic for the model is computed as the average of the $MACSE_d$ estimates from each fold:

$$GoF = \frac{1}{5} \sum_{d=1}^{5} MACSE_d$$

7. We repeat Steps 2-6 to calculate the goodness-of-fit test statistic for each of the group-level models. Except each of these group-level models will only be trained on the portion of individuals belonging to their respective groups across all 4 folds and; likewise, the mean average squared curve error $MACSE_d$ is calculated for only the individuals belonging to the model's group in the test set.

To estimate variability in the mean average curve squared error $MACSE_d$, we repeat the above procedure to calculate 100 iterations of the goodness-of-fit statistic for the overall and group-level models. Lower goodness-of-fit statistic values are indicative of better model fit.

4.3.5 Study Cohort and Outcome

Our study cohort consisted of 2641 participants with diabetes enrolled in the Chronic Renal Insufficiency Cohort (CRIC) Study. Details on the rationale and design of the CRIC Study have been previously published [25, 29, 30, 31]. To summarize, the CRIC Study recruited a racially and ethnically diverse patient population aged 21 to 74 years with varying CKD stages 3a (eGFR 45-60), 3b (eGFR 30-45) and 4 (eGFR 20-30). Participants underwent annual assessments on medical and family history, cognitive and behavioral health, anthropometric measures (weight, height), resting blood pressure, and heart rate. Blood specimens and 24-h urine samples were also collected. Baseline characteristics of our overall CRIC study cohort is displayed in Table 4.1.

Our main outcome is repeated measures of eGFR over time, calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) creatinine equation [13]. The measures were observed to be sparse and irregularly spaced, with a median of 3.92 (range: 0-15.29, IQR: 2-6.98) years among participants.

Table 4.1: Baseline clinical characteristics of 2641 participants with diabetes in the Chronic Renal Insufficiency Cohort (CRIC) Study.

Age (years)	60.67 ± 9.48
Race	
White	1105 (42)
Black	1248 (47)
Other	288 (11)
Sex	
Male	1551 (59)
Female	1090 (41)
Smoked >100 cigarettes	
Yes	1500 (57)
No	1141 (43)
BMI (kg/m^2)	34.08 ± 7.77
HbAlc (%)	7.62 ± 1.63
Diastolic BP (mmHg)	69.12 ± 12.36
Systolic BP (mmHg)	132.31 ± 21.76
Mean Arterial Pressure (mmHg)	90.19 ± 13.43
Serum Creatinine (mg/dL)	1.74 ± 0.61
Urine Creatinine (mg/dL)	81.1 ± 51.76
Urine PCR (mg/g)	1294.33 ± 2696.16
Predicted Urine ACR (mg/g)*	
<30	965 (37)
30-300	768 (29)
>300	908 (34)
Baseline eGFR ($ml/min/1.73^2$)	46.95 ± 15.23
Hypertension	
Yes	2445 (93)
No	194 (7)
ACE Inhibitor or ARB use	
Yes	2089 (79)
No	537 (20)

Values are expressed as mean ± SD or N (%).
BMI, body mass index; HbA1c, hemoglobin A1c; BP, blood pressure;
PCR, protein-creatinine ratio; ACR, albumin-creatinine ratio;
eGFR, estimated glomerular filtration rate;
ACE, angiotensin-converting enzyme; ARB, angiotensin-receptor blocker.
*Converted from Urine PCR by using the crude model in Sumida et al. (2020).

Table 4.2: Correspondence between predicted ACR and PCR for CRIC participants.

	Predicted ACR (mg/g)		
PCR (mg/g)	$<$ 30 (Normo)	30-300 (Micro)	$>$ 300 (Macro)
$<$ 150	965	34	0
150-500	0	620	0
$>$ 500	0	114	908

PCR, protein-creatinine ratio; ACR, albumin-creatinine ratio

4.3.6 Albuminuria Groups

We assign the CRIC participants to albuminuria groups in CKD based on their urine albumin-creatinine ratio (ACR) values at baseline, the point of study entry. The albuminuria groups are defined to be normo (ACR $<$ 30 mg/g), micro (30-300 mg/g), and macro ($>$ 300 mg/g). However, 842 of the 2641 individuals did not have observed baseline ACR values, while all had urine protein-creatinine ratio (PCR) baseline data. Thus, to avoid omitting a quarter of our sample, we propose employing a novel equation for converting urine PCR to urine ACR using mg/g units [92]. In particular, we use their crude model for calculating predicted ACR

$$pACR = \exp\Bigg(5.3920 + 0.3072 * \log\Big(\min\Big(\frac{PCR}{50}, 1\Big)\Big) + 1.5793 * \log\Big(\max\Big(\min\Big(\frac{PCR}{500}, 1\Big), 0.1\Big)\Big) + 1.1266 * \Big(\min\Big(\frac{PCR}{500}, 1\Big)\Big)\Bigg)$$

Hence, the albuminuria groups for our study cohort of 2641 individuals will be based on predicted ACR values converted from their observed PCR values. The numbers in each group are listed in Table 4.2 as a cross-tabulation with their respective PCR categories noted in the Kidney Disease Improving Global Outcomes (KDIGO) guidelines [88]. Here, we see that a total of 2493 (94%) of our study cohort are in albuminuria groups based on predicted ACR values that correspond to their observed PCR categories.

4.3.7 Implementation of statistical methods

All statistical analysis was conducted using the R (version 4.0.3) programming environment [93]. The R package fdapace [94] was used for the estimation of the model components of FPCA. We use the fanova.tests function from the fdANOVA package [95] to test for equality of mean functions and the ksample.perm function from the fdcov package [96] to test for equality of correlation functions.

4.4 Results

The overall (FPCA) model, fitted to our entire cohort (N=2641) of CRIC patients, estimated three leading FPCs were optimal, together capturing 98.5% of the variation in eGFR trajectories. These estimated FPCs and their associated scores are displayed Figure 4.2. The first FPC determined that 81.3% of the variation is explained by a magnitude shift from the overall model's mean eGFR trajectory. Thus, a large majority of the variation in eGFR trajectories can be traced to patients' measured eGFR at study entry, which lends to the idea that the first FPC behaves like a model intercept. The first FPC scores significantly differed by albuminuria group both globally and pairwise (all $p < 0.001$). Patients in the normo group tended to have higher scores and a greater proportion of patients in this group have positive valued scores than the other two groups. Hence, resting solely on the first FPC, the normo patients are more inclined to have consistently increased eGFR compared to micro and macro patients. A notable proportion of macro patients have negative valued first FPC scores along with mostly lower scores than those of normo and micro patients. Accordingly, the first FPC exerted that macro patients are more inclined to have consistently decreased eGFR.

The second FPC accounted for 13.5% of the variation and captured varying rates

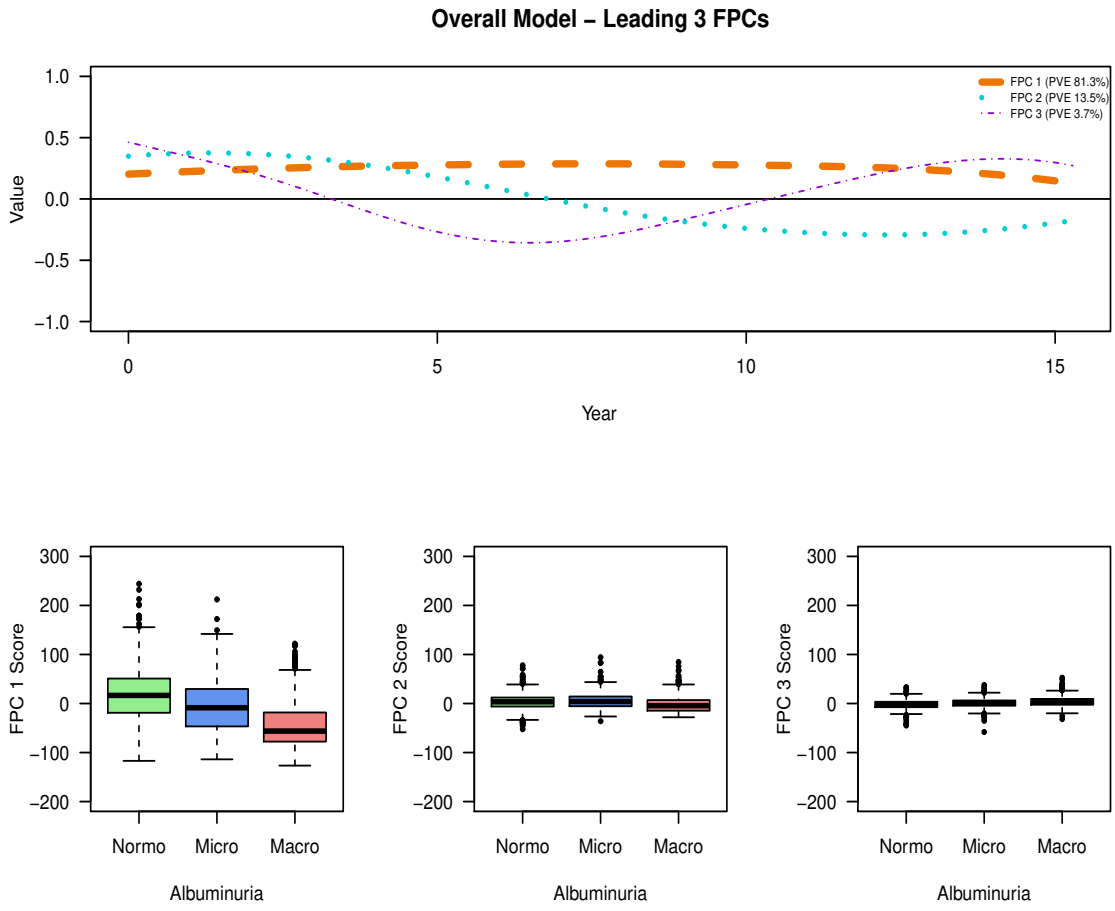

Figure 4.2: (a) Leading three FPCs for our overall model with proportion of eGFR variance explained (PVE %). (b) Box plots of the scores for the leading FPCs by albuminuria group. The three FPC scores significantly differed by albuminuria group both globally and pairwise (all $p < 0.001$, except normo vs micro for FPC 2 scores has $p = 0.013$).

of change in eGFR in the first and second halves of our follow-up time frame. Similar to how the first FPC is like a surrogate to a model intercept, the behavior of the second FPC is akin to that of a model slope. The second FPC scores significantly differed by albuminuria group both globally and pairwise (all $p < 0.001$, except normo vs micro $p = 0.013$). Here, the distribution of the second FPC scores for the normo and micro groups are alike with one another, while the macro patients tended to have slightly lower scores. Therefore, the rate of change in eGFR for macro patients are more differentiated from that of normo and micro patients. Finally, the third FPC explained 3.7% of the variation and also captured varying rates of change in eGFR although with different inflection points. The third FPC scores significantly differed by albuminuria group both globally and pairwise (all $p < 0.001$). Much like the second FPC scores, the distribution of the scores here for the normo and micro groups are more alike with one another. In contrast to before, now the macro patients tended to have higher third FPC scores than normo and micro patients, although the difference is less notable compared to the second FPC scores comparison. Taking both the second and third FPCs and their scores into account, the eGFR trajectories of macro patients are more prone to having various rate of changes throughout the follow-up period than those of normo and micro patients.

The predicted eGFR trajectories for all the CRIC patients from the overall model are illustrated in Figure 4.3. The mean eGFR trajectories for the albuminuria groups were calculated by taking the sample means of the FPCA-predicted eGFR at each time grid point (years) of the group-specific patients and connecting these sample means to form that group's trajectory. By Gorecki-Smaga et al.'s permutation test [89], the mean eGFR trajectories for the albuminuria groups differed both globally ($FP = 318.86$, $p < 0.001$) and pairwise (normo vs micro $FP = 74.13$, normo vs macro $FP = 641.34$, micro vs macro $FP = 239.27$, all $p < 0.001$). The normo group's mean eGFR trajectory was consistently greater than those of the other groups by an increased magnitude shift, largely attributed to the first

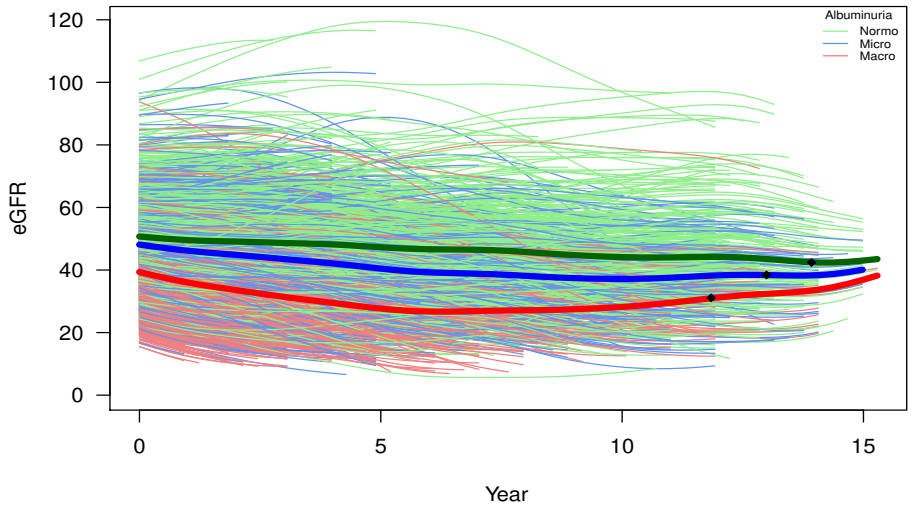

Figure 4.3: Predicted eGFR trajectories for our sample of diabetic patients from our overall model. The thick curves are the albuminuria group mean trajectories with a black diamond corresponding to when that group has < 50 eGFR samples after that time point (Normo – 13.92 years, Micro – 12.99 years, Macro – 11.84 years).

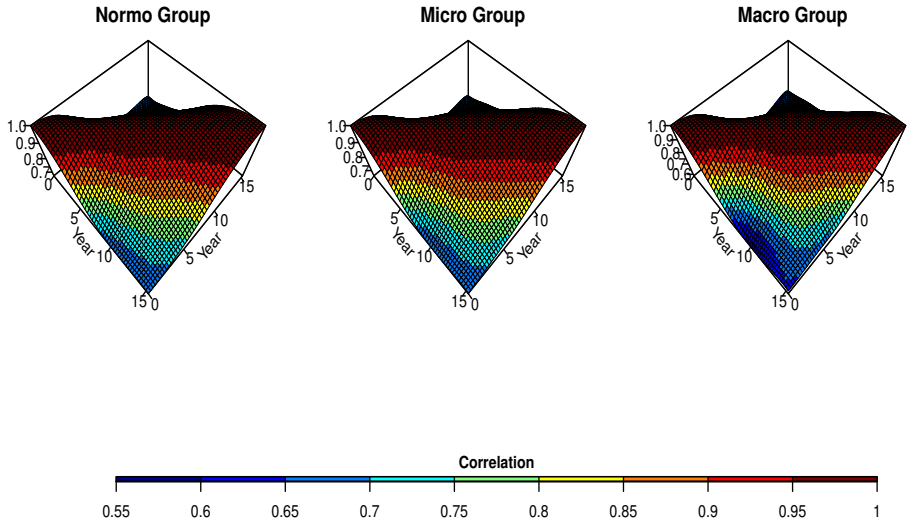

Figure 4.4: Correlation functions for the normo, micro, and macro albuminuria groups of diabetic patients in the CRIC study.

FPC and the associated scores for the normo patients being notably higher than the other groups of patients. The micro group's mean eGFR trajectory, although a downward shift from that of the normo group, had a similar rate of decline over time. As mentioned before, the second and third FPCs captured the varying rates of change in eGFR for our overall cohort and both their associated scores are extremely similar in distribution for the normo and micro groups, which led to similar observed rates of change in eGFR for both groups. The macro group's mean eGFR trajectory differed from the other group's mean trajectories in both magnitude shift and rates of change. In particular, the trajectory has a steeper initial decline than the normo and micro groups and a noticeable rebound of the trajectory for later follow-up period is owed to both the second and third FPCs. Here, the second FPC was negative along much of the same late follow-up period as the eGFR rebound, and with the macro patients having more negative valued scores, contributed to a positive rate of change in eGFR. The third FPC was positive for the similar follow-up period as the eGFR rebound and the macro patients having more positive valued scores further pitched in to the increase in eGFR towards the end of follow-up.

The estimated correlation functions of the eGFR trajectories for each albuminuria group are presented in Figure 4.4. Here, the correlation at any two time points is not lower than 0.55 for any group. By Cabassi et al.'s permutation test [90], the correlation functions of eGFR for the albuminuria groups differed both globally ($p < 0.001$) and pairwise (normo vs micro $p = 0.028$, normo vs macro $p < 0.001$, micro vs macro $p < 0.001$), although accounting for multiple comparisons via a Bonferroni correction, the normo and micro groups would not differ. We observed that the macro group contained a wider area of correlations < 0.7 than the other groups. This particular area is largely concentrated on the correlations between eGFR at the earlier time period < 5 years and that of the later time period > 5 years. Within this area, the macro group correlations between baseline, year 0, and > 5 years was mostly lower than those of the normo and micro groups, indicated by

values < 0.65 in the dark blue region.

The albuminuria-specific (group) models, each fitted using only data from CRIC patients of a particular albuminuria group, estimated varying leading FPCs for capturing at least 95% of the variation in the eGFR trajectories of each group and are displayed in Figure 4.5. In addition, the same estimated FPCs from the overall model, displayed in panel (a) of Figure 4.2, are overlaid on these plots for comparison with those of the group models. The first FPC for the normo model was similar to that of the overall model; however, the micro and macro models did not have as consistent a magnitude shift over time from their mean eGFR trajectories. In fact, the first FPCs for the micro and macro models did not account for as much variation in their patients' trajectories ($< 80\%$) and are also approaching the value 0 with greater follow-up time, which could simply reflect fewer samples at later time-points. Inversely, the second FPC for these models explained more variation ($> 17\%$) compared to the overall and normo models ($< 14\%$). The second FPCs for the four models captured varying rates of eGFR change, although those of the micro and macro models closely resemble reflections of the overall and normal models about the value 0, with the macro model having a more steady rate of change towards the last few years of follow-up, again possibly simply reflecting sparse data in this group at later follow-up years. If we take a look at the second FPC plots around the mean for our models in Appendix Figures A.8-A.11, we can deduce that except to the reflection of the second FPCs, the micro and macro models yield similar rates of change in eGFR as that of the overall and micro models with their second FPCs. Finally, we note (Figure 4.5) that the normo model only estimated two leading FPCs while the other models contained three, with the macro model's third FPC undergoing a sharp decline near the last years of follow-up.

As a preliminary visual assessment of model fit, we compared the mean fitted trajectories for the overall and group-specific models (Figure 4.6). The predicted mean eGFR trajectories of normo patients were similar when using an overall or normo model; while

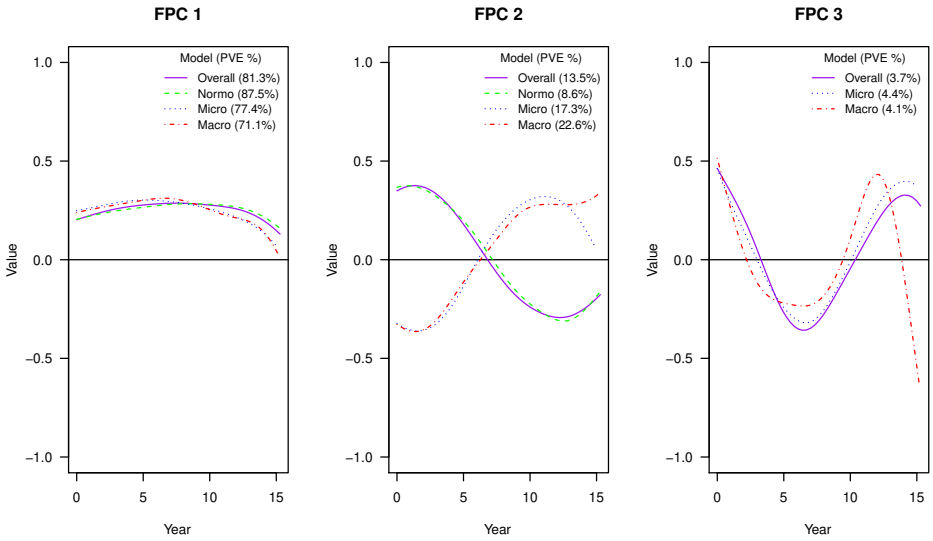

Figure 4.5: Leading three FPCs for our overall and group-specific models along with proportion of eGFR variance explained (PVE %). Notably, the normo model only required the leading two FPCs to explain at least 95% of the total eGFR variation; therefore, only the leading two FPCs for this model are plotted.

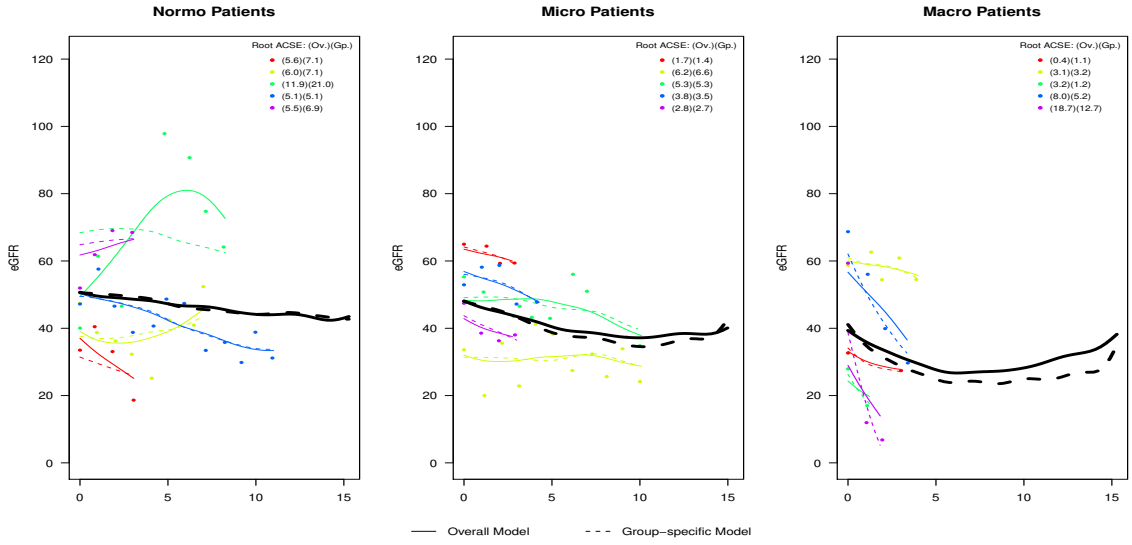

Figure 4.6: Comparison of predicted individual eGFR trajectories from the overall and albuminuria group-specific models for N=5 (a) normo patients, (b) micro patients, and (c) macro patients, with albuminuria group mean trajectories overlaid. The root of ACSE (average curve squared error), described in Section 4.3.4, for both the overall (ov.) and group-specific (gp.) predicted trajectories is displayed for each patient.

the mean trajectories of micro and macro patients differed with various degrees when using an overall or that respective group model. Specifically, the micro model predicted lower mean eGFR than that of the overall model after a few years of follow-up, while, the macro model showed a more glaring difference in mean eGFR than that of the overall model. The mean trajectories for micro and macro groups rebounded over the last few years, likely due to sparse data at follow-up.

To further explore these differences at the individual level, a sample of patients in each albuminuria group (N=5) were selected for comparison of their predicted eGFR trajectories to their observed eGFR measures from the overall model vs their albuminuria group-specific model as displayed in Figure 4.6. While the overall and group-specific models tracked the observed data reasonable well, there are, nevertheless, notable differences which are apparent through visual inspection as well as via the root average curve squared error (ACSE), introduced in Section 4.3.4. Compared to the group-specific models, we see, based on the root ACSE metric, the overall model tends to predict better for the normo patients, predict similarly for the micro patients, and can predict worse for macro patients. However, these results may be over optimistic, since these patients were also used to train the models. Thus, we next investigated cross-validated prediction error using the goodness-of-fit statistic as outlined in Section 4.3.4.

The results of 100 iterations of our proposed goodness-of-fit procedure for comparing between the single overall model and three albuminuria-specific models are presented in Figure 4.7. Our goodness-of-fit procedure is based on prediction error of eGFR and lower values are indicative of better model fit. The overall model (OV-NO) displayed largely better prediction performance for normo patients than the normo model (NO). With more severe albuminuria groups, we see the micro and macro models (MI, MA) had better prediction performance for their respective group patients than using the overall model (OV-MI, OV-MA). Despite the MA model mostly having the least prediction error, the MI

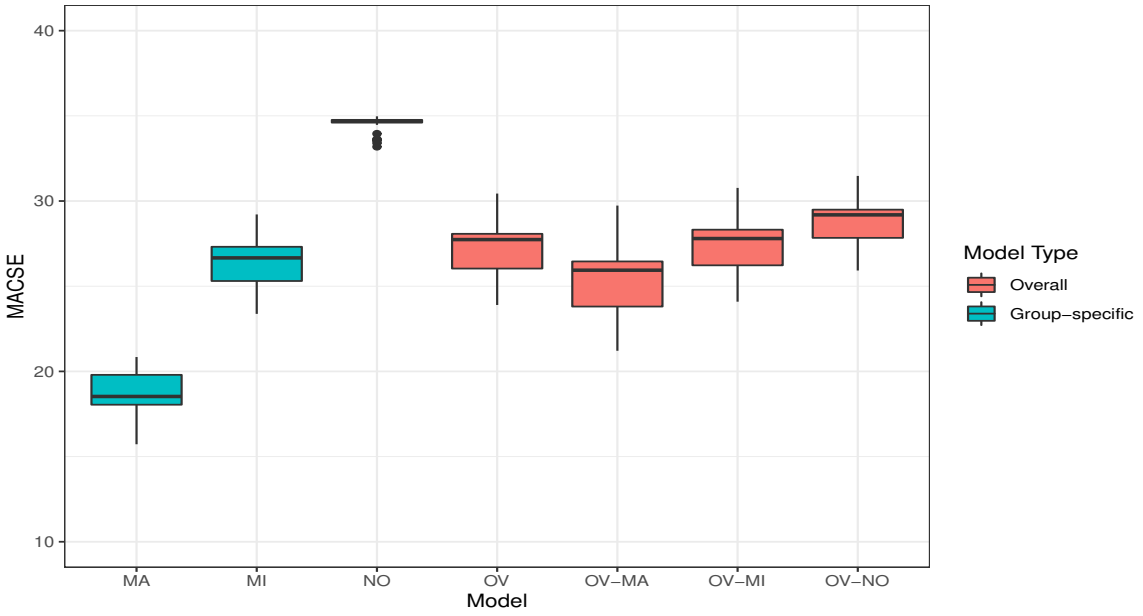

Figure 4.7: Box plots of model prediction performance: 100 repeats of 5-fold cross-validated MACSE (mean average curve squared error) for FPCA-estimated eGFR trajectories. Model: MA, Macro; MI, Micro; NO, Normo; OV, Overall; OV-MA, OV-MI, OV-NO; Overall prediction for the respective group-specific patients.

and MA models' prediction estimates also have the most variability. Furthermore, the gap in prediction performance is largest between the OV-NO and NO models, with the OV-NO model showing a clear advantage. Finally, the range of prediction estimates across all three groups of patients in using the single overall model [21.21-31.47] was also less than using three albuminuria-specific models [15.72-34.97].

4.5 Discussion

We applied FPCA methodology to model long-term eGFR trajectories for diabetic patients that accounted for nonlinear, sparse, and irregularly spaced eGFR time series. The first two leading FPCs, accounted for 95% of the variation in eGFR patterns, and behaved similarly to a model's intercept and slope respectively. To examine whether the longitu-

dinal eGFR patterns significantly varied between albuminuria groups, a clinical subgroup of interest in the renal disease context, we tested for differences in mean and correlation functions. As mentioned in the discussion of Dong et al. [87], it may be of great interest to predict the future eGFR for patients of different clinical subgroups. One of their proposed solutions to this is fitting separate FPCA for each subgroup much like our group-level approach. However, we further extend this idea by incorporating a goodness-of-fit procedure that uses cross-validated leave-a-curve-out prediction errors to decide between fitting separate albuminuria group-specific FPCA models and using a single overall FPCA model.

We found that the most dominant mode of eGFR variation was a magnitude shift from the mean eGFR for all groups whether fitting an overall or group model, having explained $> 70\%$ of the variability. Lesser dominant modes involved varying rates and time-intervals of eGFR change; these modes were pertinent for modeling trajectories for patients in the micro and macro groups than those in the normo group, with the second FPCs explaining $> 17\%$ of the eGFR variation for just the micro and macro groups versus $\approx 9\%$ for the normo group. This is further compounded by a third FPC not required for the normo group, in contrast to the micro and macro groups, in which the third FPC explained 4% of the variation in eGFR trends. Mean and correlation eGFR functions significantly differed between albuminuria groups. Upon further inspection, we found that the correlations between eGFR at the earlier time period < 5 years and that of the later time period > 5 years are noticeably lower for the macro patients, suggesting more diffuse, i.e., less tightly linked long-term eGFR trajectories in patients with more severre kidney disease. Results from our goodness-of-fit procedure of prediction performance comparison between FPCA models indicated that the choice of a single overall model is a viable option to predicting long-term eGFR trajectories for different albuminuria groups. In particular, although micro and macro specific models had lower prediction error compared to the overall model, these

group-specific models also had more variability in prediction error suggesting the class tradeoff between low bias (group-specific models) versus low variance (overall model).

We acknowledge the strengths and limitations in our work and potential future directions to explore. First, for strengths. Our CRIC study cohort of diabetic patients is one of the largest in the U.S., with extensive clinical profiles and extended longitudinal data of kidney function. FPCA has the advantage of predicting non-linear trajectories and investigating for their leading modes of variation, all the while overcoming sparsity and irregularly spaced trends. This approach, thus permits a nuanced assessment of long-term disease progression as evidenced in our findings. In addition, inference for mean and correlation functions between groups were based on permutation tests and are therefore robust to misspecification of distribution. Our goodness-of-fit procedure gauges the variability of curve prediction error via repeated cross-validated measures which reduces overfitting and avoids optimistic prediction performance for the overall and group-level models. Now, for limitations. From a clinical perspective, assignment of albuminuria groups to our CRIC patients was based on converting urine PCR to urine ACR. Although the equation for conversion is cutting-edge [92] and requires further testing to assess its robustness and utility in predicting ACR, we found that close to 94% of our CRIC patients were in albuminuria groups that corresponded to their PCR categories by the KDIGO guidelines [88]. Furthermore this conversion is easy to implement, and importantly allowed us to retain our full sample of 2641 CRIC patients. We conducted sensitivity analyses using the PCR categories and achieved similar results. Second, data were sparse at later follow-up times, especially for the macro group, hence any conclusions for later times need to be interpreted with caution. Third, although our proposed goodness-of-fit procedure for FPCA model comparison in predicting long-term outcome trajectories is relatively novel, more work in this area is needed. We aim to further its methodological developments. For instance, to investigate the precision of our model-driven estimate for prediction error, we would like to explore

bootstrap resampling techniques for deriving an empirical distribution of the standard error of our estimates, rather than using cross-validated estimates which could suffer from small test sample-size.

To our knowledge, ours is the first application of functional data methods to examine kidney disease progression in diabetic patients. Dong et al. [87] had previously applied FPCA to model kidney disease progression and investigate for dominant modes of variation in kidney transplant recipients; however, our study different in that we examined and compared longitudinal patterns and different model fits for long-term DKD progression in different albuminuria groups. More specifically, the first two leading FPCs essentially mimic intercept (i.e., starting eGFR value) and slope (i.e. linear rate of change of eGFR), suggesting that standard statistical approaches such as linear mixed models would be sufficient for capturing the majority of variation in kidney disease patterns. Nevertheless, the more subtle variations captured by the third FPC, albeit only explaining 4% of the variation, may still be important for individual patients, especially those with more advanced albuminuria stage. Similarly, while the full cohort model fit the data well and had low prediction error overall, there were differences in performance based on albuminuria severity, with group-specific models being more advantageous in the macro group. The decision of training a single model over multiple models to model long-term trajectories is a question of parsimony versus complexity, i.e., the overall model is simpler, computationally less demanding, and more generalizable since it applied to the entire spectrum of albuminuria. Thus a single FPCA model, or even a more traditional intercept and slope model, can be recommended in population settings where the goal is to model population level trends, and test broad brush differences between groups. However, in a clinical setting group-specific models may offer better predictive performance and offer the opportunity for personalized recommendations for individuals presenting with more advanced disease.

4.6 Acknowledgements

Chapter 4, in part is currently being prepared for submission for publication of the material. Kwan, Brian; Anderson, Amanda; Anderson, Cheryl; Chen, Jing; Fuhrer, Tobias; Montemayor, Daniel; Ricardo, Ana; Rosas, Sylvia; Yang, Wei; Zhang, Jing; Natarajan, Loki. "Inference for Functional Principal Components of Kidney Disease Progression in Diabetic Patients Across Albuminuria Groups in the Chronic Renal Insufficiency Cohort (CRIC) Study". The dissertation author was the primary investigator and author of this material.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1650112. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Chapter 5

Conclusions and Future Work

In this dissertation, we implemented, developed and compared statistical approaches for (i) modeling disease progression using longitudinal biomarkers and (ii) building useful and interpretable prognostic models in the context of diabetic kidney disease. The human metabolome holds high promise for uncovering biomarkers sensitive to detection of rapid DKD progression. In assuming linear rate of kidney function decline, the linear mixed model and two-stage methods are classic candidates for developing prognostic models for DKD progression using metabolite predictors. Notably, we also proposed using TSP-based methods to identify metabolite-pair markers that best discriminate between DKD severity stages. FPCA accounts for nonlinear and other complexities in curve estimation that linear mixed model and two-stage methods otherwise could not. By predicting long-term trajectories and detecting major modes of curve variation, the use of FPCA strengthens our understanding of patterns DKD progression, especially across different diabetic subpopulations. Thus, metabolomic markers and functional data methods exhibit strong potential for uncovering the characteristics of and predicting DKD progression.

The choice between linear mixed model and two-stage methods is ultimately dictated by the goals of the analysis, i.e., whether to minimize overall prediction error vs

unbiased estimation of associations between metabolite predictor and eGFR slope outcome. While the linear mixed model is the more conventional modeling approach of the two, two-stage methods proved to be viable alternatives for several study design scenarios, which would allow applied researchers to use slope outcomes. TSP-based methods identified the top-scoring metabolite-pairs for which an observed ordering of the two features is more common in one class than in the other, i.e., early-stage vs advanced-stage DKD, which makes them not only easily biologically interpretable as a "reversal" from one class to another based on feature ordering, but also has TSP as parsimonious models for discriminating severity in disease. We proposed our residualizing approach for TSP that helps in identify top-scoring pairs largely liberated from clinical covariate information, which could motivate further follow-up studies on these largely "cleaner" biomarker pairs. In applying FPCA to model long-term DKD progression via eGFR trajectories, the leading dominant modes of eGFR behaved much like a model's intercept (i.e., starting eGFR value) and slope (i.e., linear rate of change of eGFR), which raises the appeal of using a linear mixed model or two-stage methods; however, FPCA also accounts for sparse and irregularly spaced eGFR time series. Longitudinal eGFR patterns significantly varied between albuminuria groups of diabetic patients and our proposed goodness-of-fit procedure indicated that the choice between fitting separate albuminuria group-specific models or a single overall model for predicting long-term eGFR trajectories resembles a tradeoff between low bias (group-specific models) versus low variance (overall model).

Our dissertation opens several opportunities for future methodological and applied work. The simulations comparing the linear mixed model and two-stage approaches are based on N=200 subjects and additional analysis with different N could provide further insight on the optimal choice of methods for smaller and larger sized cohorts. These approaches were also compared only under a single missing data mechanism, MCAR, and other mechanisms, MAR and MNAR, require further investigation. The results for our

TSP-based methods were from using a binary outcome of DKD severity stage and multi-class approaches can be considered. We would also like to include metabolite ions highly associated with DKD and part of pathways informative of therapeutic targets for DKD back into the pool of metabolites after statistical filtering for top-scoring pair selection. Furthermore, more complex statistical models could be fitted to obtain the residuals although our usage of linear regression models have ease of implementation for having the residuals independent of clinical covariates. Our proposed goodness-of-fit procedure for comparing cross-validated prediction error of FPCA models is relatively novel and, to investigate the precision, we would like to incorporate bootstrap resampling techniques for deriving an empirical distribution of the standard error of our estimates, rather than using cross-validated estimates which could suffer from small test sample-size.

In summary, our works provide a framework of statistical approaches for modeling kidney function decline for diabetic patients using metabolomic markers and functional methods. Our methods and results are easily generalizable to other disease prognostic modeling studies which offer opportunities for future methodological research and clinical applications beyond the metabolite-DKD setting.

Appendix

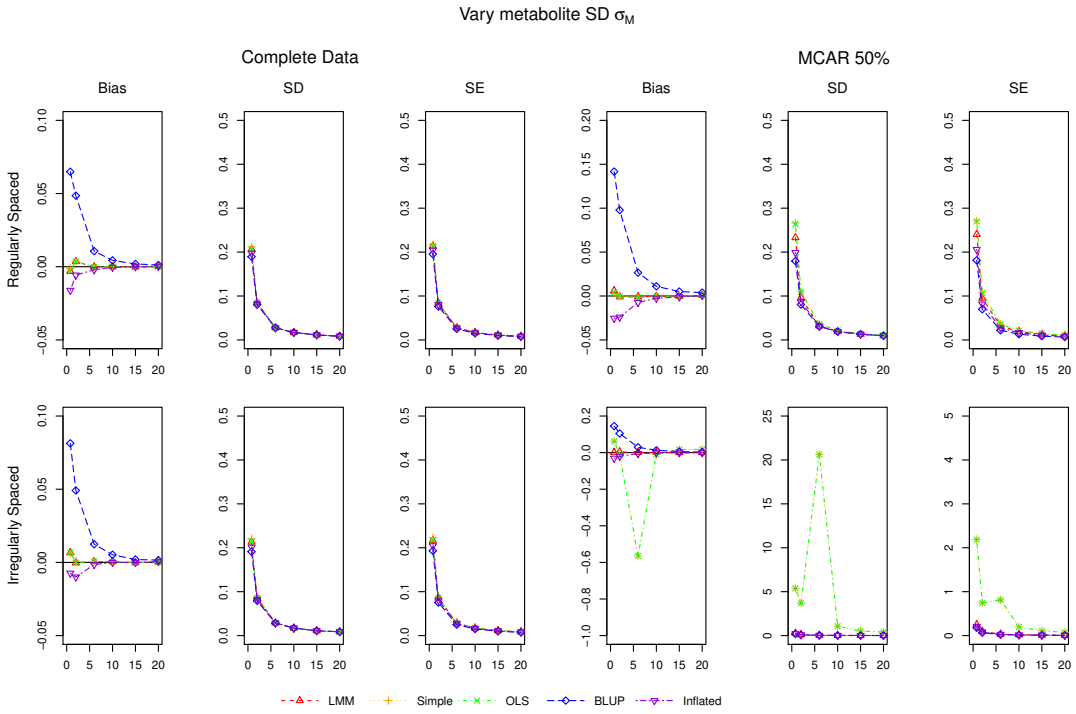

Figure A.1: Performance in bias, SD, and SE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of metabolite SD for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

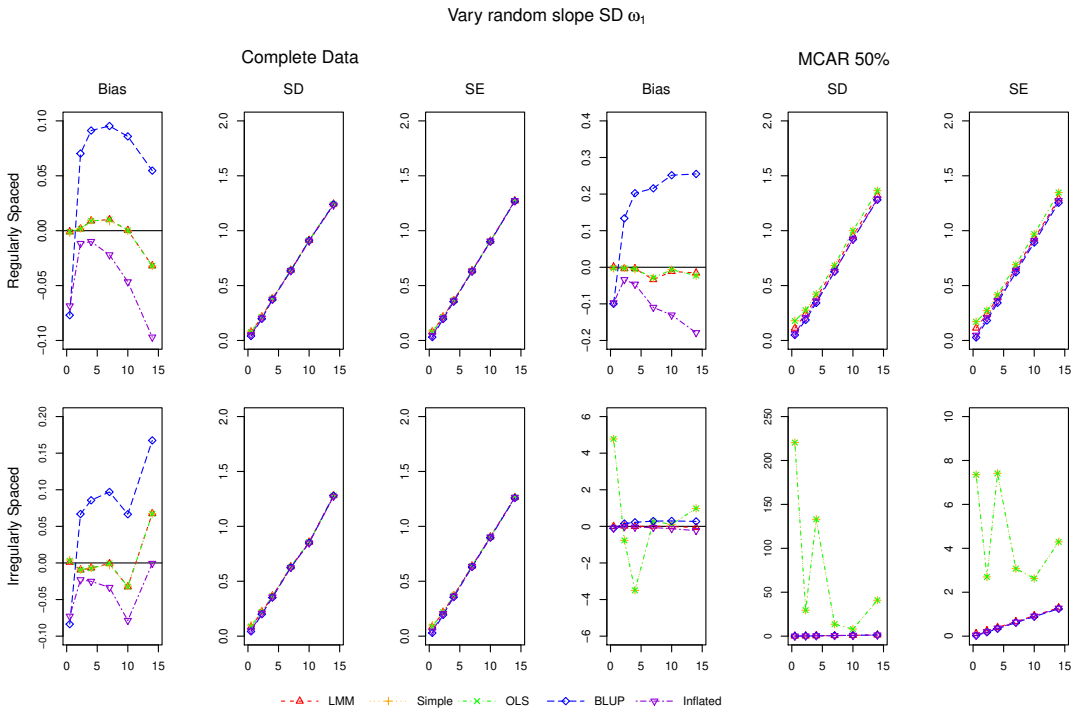

Figure A.2: Performance in bias, SD, and SE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of random slope SD for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

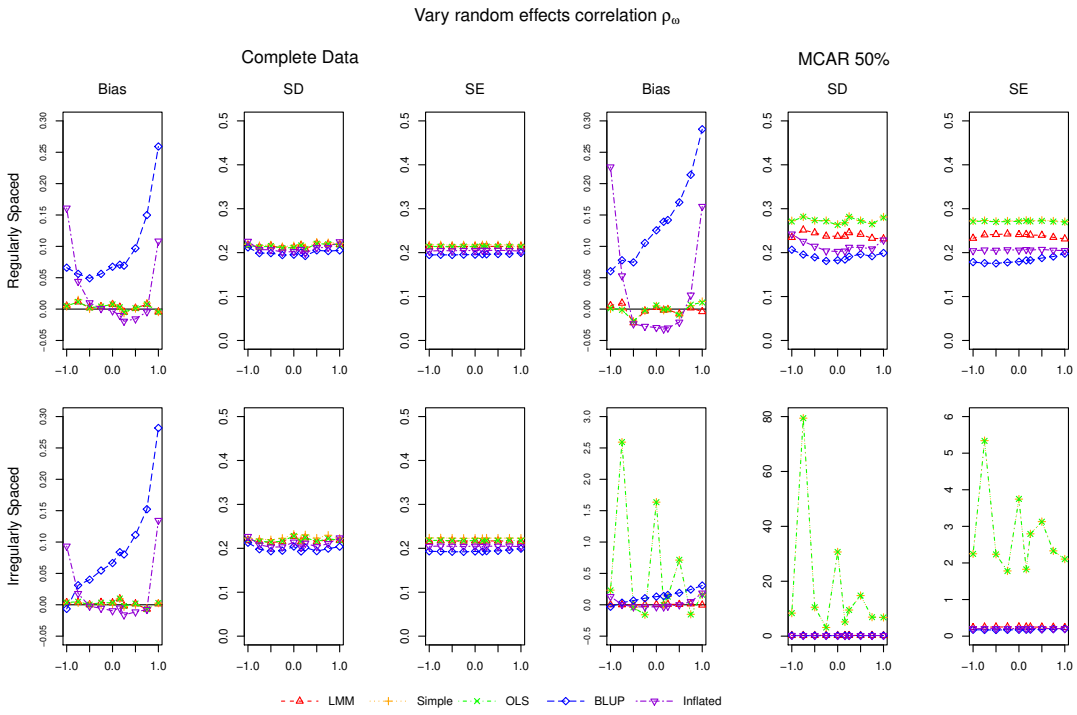

Figure A.3: Performance in bias, SD, and SE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of the correlation between random intercept and slope for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

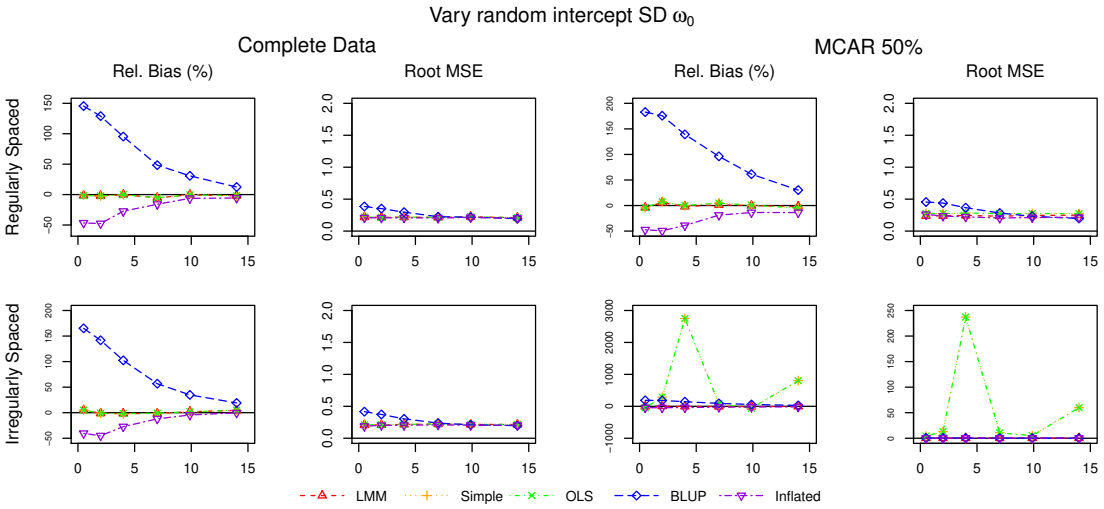

Figure A.4: Performance in relative bias (%) and root MSE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of random intercept SD for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

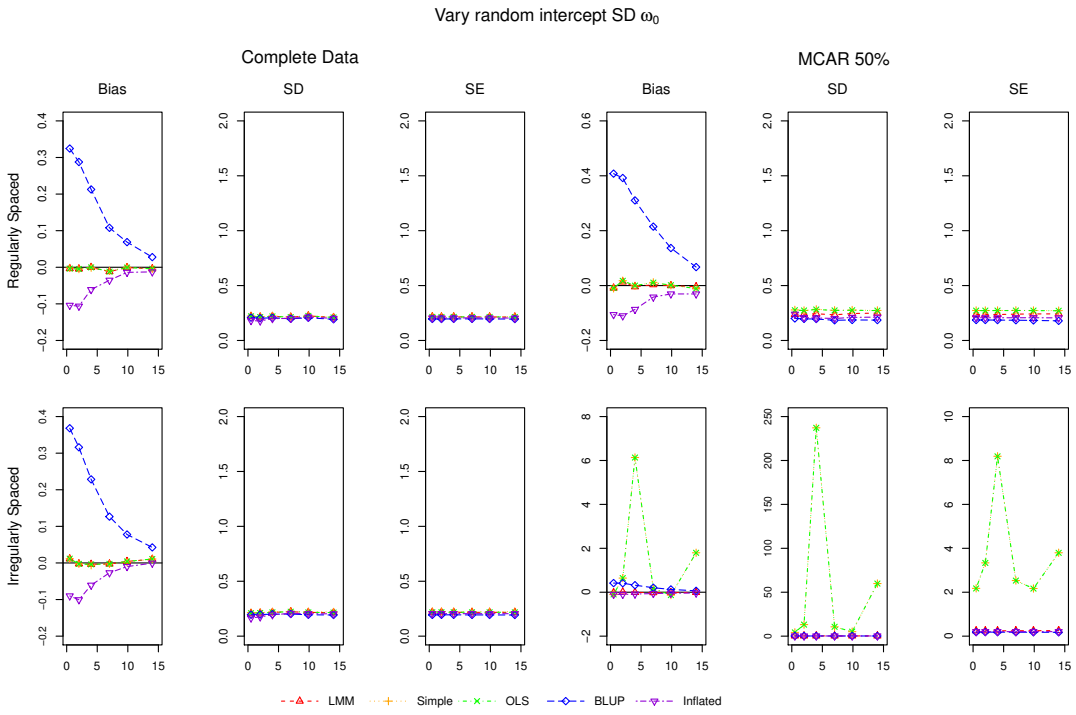

Figure A.5: Performance in bias, SD, and SE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of random intercept SD for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

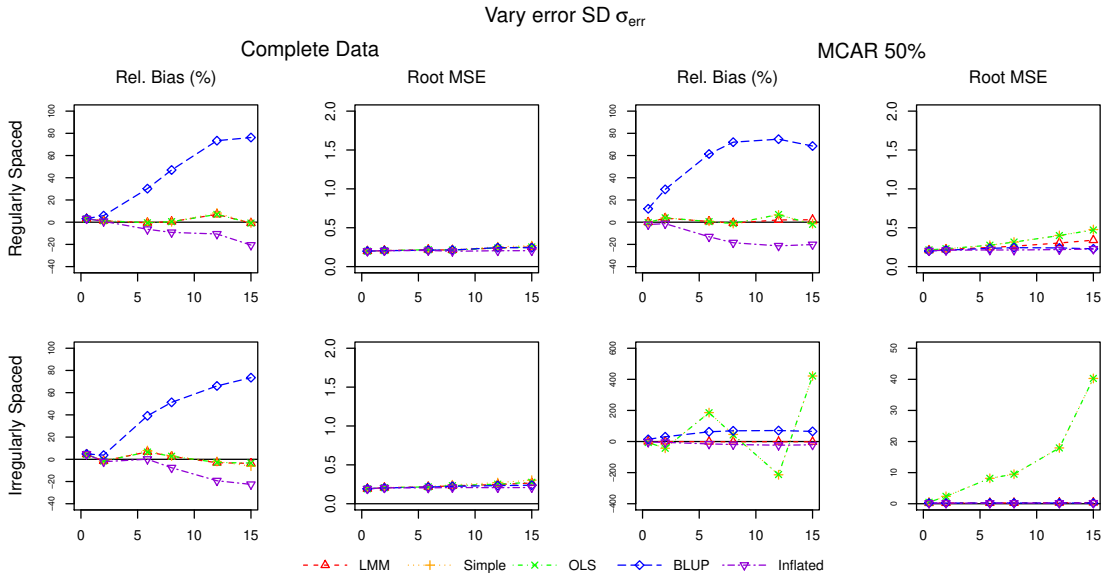

Figure A.6: Performance in relative bias (%) and root MSE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of error SD for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

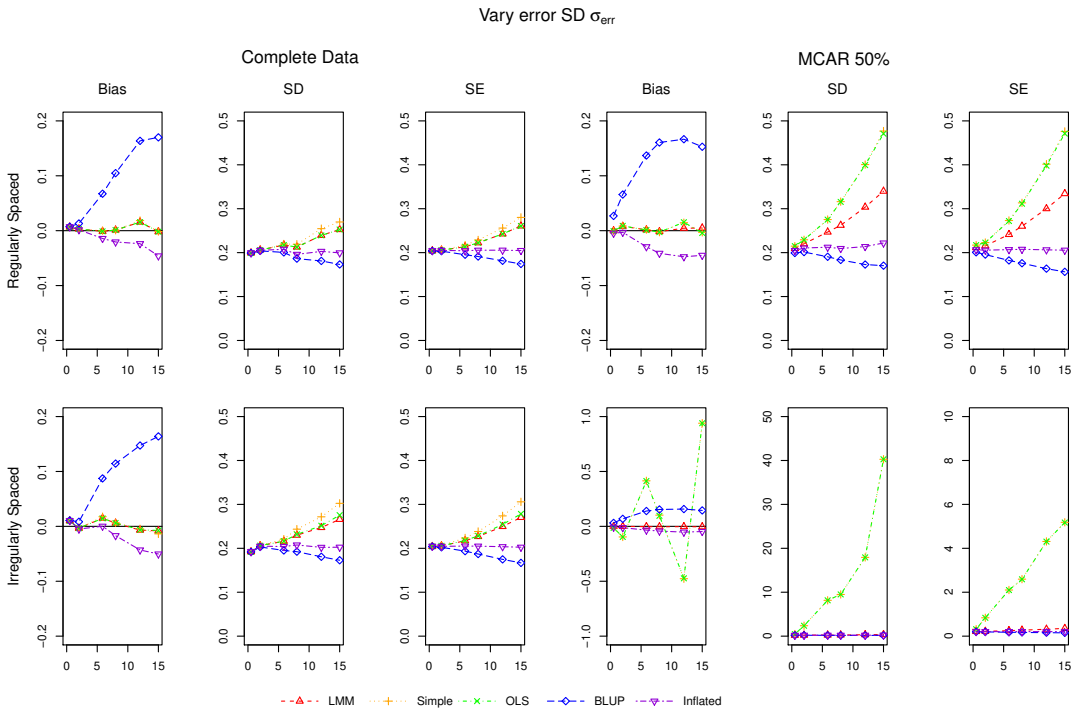

Figure A.7: Performance in bias, SD, and SE of our methods in estimating the association between annual rate of eGFR change and metabolite for the linear mixed model ($\widehat{\beta}_3$) versus two-stage methods ($\widehat{\alpha}_1$) as a function of error SD for the regularly and irregularly spaced cases of Complete Data and MCAR 50%.

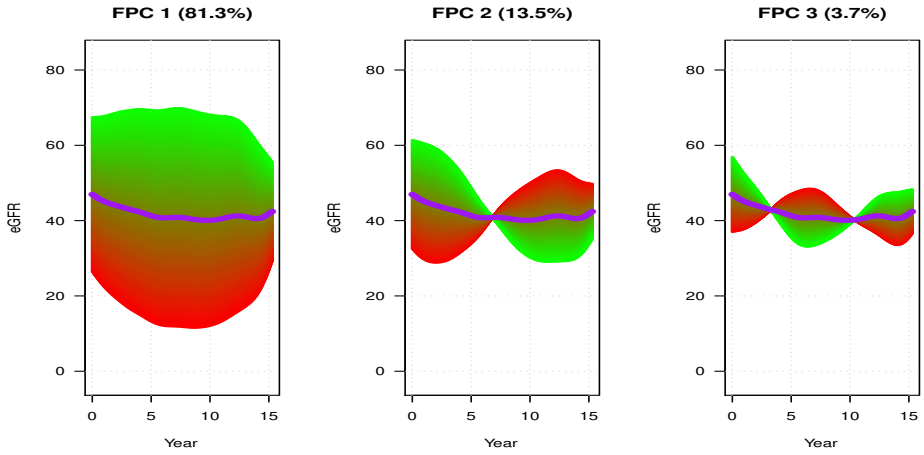

Figure A.8: The FPCs, with their proportion of eGFR variance explained (%), around the mean eGFR trajectory (black) for the Overall model. The green and red shaded areas show the range of variation around the mean: $+/-Q \times \sqrt{\widehat{\lambda}_k} \times \widehat{\phi}_k(t)$, where $\widehat{\phi}_k(t)$ is the estimated kth FPC and $\widehat{\lambda}_k$ is its eigenvalue. The green edge corresponds to $Q = 2$ and the red edge corresponds to $Q = -2$.

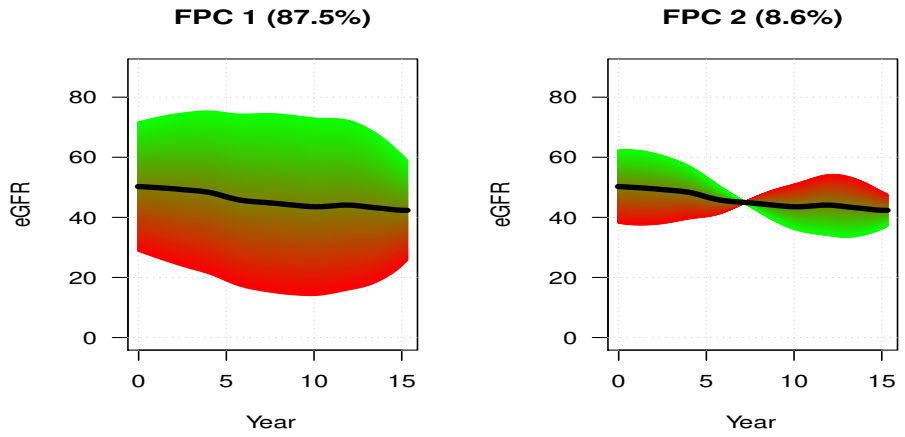

Figure A.9: The FPCs, with their proportion of eGFR variance explained (%), around the mean eGFR trajectory (black) for the Normo model. The green and red shaded areas show the range of variation around the mean: $+/-Q \times \sqrt{\widehat{\lambda}_k} \times \widehat{\phi}_k(t)$, where $\widehat{\phi}_k(t)$ is the estimated kth FPC and $\widehat{\lambda}_k$ is its eigenvalue. The green edge corresponds to $Q = 2$ and the red edge corresponds to $Q = -2$.

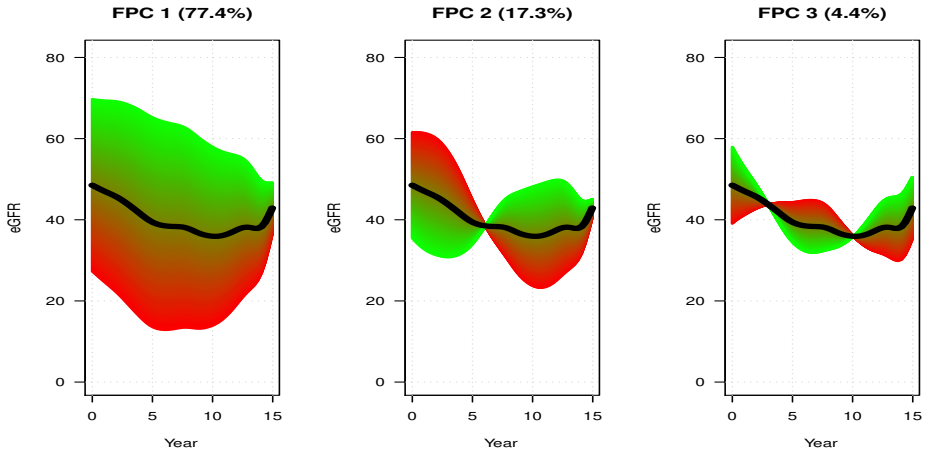

Figure A.10: The FPCs, with their proportion of eGFR variance explained (%), around the mean eGFR trajectory (black) for the Micro model. The green and red shaded areas show the range of variation around the mean: $+/-Q \times \sqrt{\widehat{\lambda}_k} \times \widehat{\phi}_k(t)$, where $\widehat{\phi}_k(t)$ is the estimated kth FPC and $\widehat{\lambda}_k$ is its eigenvalue. The green edge corresponds to $Q = 2$ and the red edge corresponds to $Q = -2$.

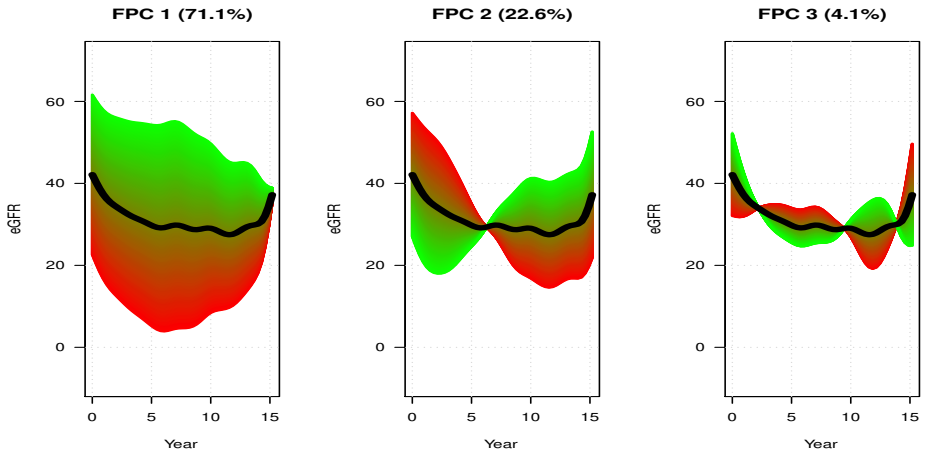

Figure A.11: The FPCs, with their proportion of eGFR variance explained (%), around the mean eGFR trajectory (black) for the Macro model. The green and red shaded areas show the range of variation around the mean: $+/-Q \times \sqrt{\widehat{\lambda}_k} \times \widehat{\phi}_k(t)$, where $\widehat{\phi}_k(t)$ is the estimated kth FPC and $\widehat{\lambda}_k$ is its eigenvalue. The green edge corresponds to $Q = 2$ and the red edge corresponds to $Q = -2$.

Bibliography

[1] United States Renal Data System. 2018 USRDS Annual Data Report: Epidemiology of kidney disease in the United States. Technical report, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2018.

[2] Centers for Disease Control and Prevention. National Chronic Kidney Disease Fact Sheet, 2017. Technical report, US Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA, 2017.

[3] Robert A. Bailey, Yiting Wang, Vivienne Zhu, and Marcia Ft Rupnow. Chronic kidney disease in US adults with type 2 diabetes: An updated national estimate of prevalence based on Kidney Disease: Improving Global Outcomes (KDIGO) staging. *BMC Research Notes*, 7(1):1–7, 2014.

[4] Carol E. Koro, Bo Hyen Lee, and Steve J. Bowlin. Antidiabetic medication use and prevalence of chronic kidney disease among patients with type 2 diabetes mellitus in the United States. *Clinical Therapeutics*, 31(11):2608–2617, 2009.

[5] Digsu N. Koye, Dianna J. Magliano, Robert G. Nelson, and Meda E. Pavkov. The Global Epidemiology of Diabetes and Kidney Disease. *Advances in Chronic Kidney Disease*, 25(2):121–132, 2018.

[6] Kai Mckeever Bullard, Catherine C Cowie, Sarah E Lessem, Sharon H Saydah, Andy Menke, Linda S Geiss, Trevor J Orchard, Deborah B Rolka, and Giuseppina Imperatore. Morbidity and Mortality Weekly Report Prevalence of Diagnosed Diabetes in Adults by Diabetes Type-United States, 2016. Technical Report 12, 2016.

[7] Morgan E. Grams, Wei Yang, Casey M. Rebholz, Xue Wang, Anna C. Porter, Lesley A. Inker, Edward Horwitz, James H. Sondheimer, L. Lee Hamm, Jiang He, Matthew R. Weir, Bernard G. Jaar, Tariq Shafi, Lawrence J. Appel, Chi yuan Hsu, Harold I. Feldman, Alan S. Go, John W. Kusek, James P. Lash, Akinlolu Ojo, Mahboob Rahman, and Raymond R. Townsend. Risks of Adverse Events in Advanced CKD: The Chronic Renal Insufficiency Cohort (CRIC) Study. *American Journal of Kidney Diseases*, 70(3):337–346, 2017.

[8] Edward W. Gregg, Yanfeng Li, Jing Wang, Nilka Rios Burrows, Mohammed K. Ali, Deborah Rolka, Desmond E. Williams, and Linda Geiss. Changes in Diabetes-Related Complications in the United States, 1990–2010. *New England Journal of Medicine*, 370(16):1514–1523, 2014.

[9] Ravi Retnakaran, Carole A. Cull, Kerensa I. Thorne, Amanda I. Adler, and Rury R. Holman. Risk factors for renal dysfunction in type 2 diabetes: U.K. Prospective Diabetes Study 74. *Diabetes*, 55(6):1832–1839, 2006.

[10] Richard J. MacIsaac and George Jerums. Diabetic kidney disease with and without albuminuria. *Current Opinion in Nephrology and Hypertension*, 20(3):246–257, 2011.

[11] Andrzej S. Krolewski, Monika A. Niewczas, Jan Skupien, Tomhito Gohda, Adam Smiles, Jon H. Eckfeldt, Alessandro Doria, and James H. Warram. Early progressive renal decline precedes the onset of microalbuminuria and its progression to macroalbuminuria. *Diabetes Care*, 37(1):226–234, 2014.

[12] A S Levey, J P Bosch, J B Lewis, T Greene, N Rogers, and D Roth. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Annals of internal medicine*, 130(6):461–70, 1999.

[13] Andrew S. Levey, Lesley A. Stevens, Christopher H. Schmid, Yaping Zhang, Alejandro F. Castro, Harold I. Feldman, John W. Kusek, Paul Eggers, Frederick Van Lente, Tom Greene, and Josef Coresh. A new equation to estimate glomerular filtration rate. *Annals of Internal Medicine*, 150(9):604–612, 2009.

[14] Lesley A. Inker, John Eckfeldt, Andrew S. Levey, Catherine Leiendecker-Foster, Gregory Rynders, Jane Manzi, Salman Waheed, and Josef Coresh. Expressing the CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) cystatin C equations for estimating GFR with standardized serum cystatin C Values. *American Journal of Kidney Diseases*, 58(4):682–684, 2011.

[15] Lesley A. Stevens, Josef Coresh, Christopher H. Schmid, Harold I. Feldman, Marc Froissart, John Kusek, Jerome Rossert, Frederick Van Lente, Robert D. Bruce, Yaping (Lucy) Zhang, Tom Greene, and Andrew S. Levey. Estimating GFR Using Serum Cystatin C Alone and in Combination With Serum Creatinine: A Pooled Analysis of 3,418 Individuals With CKD. *American Journal of Kidney Diseases*, 51(3):395–406, 2008.

[16] Lesley A. Inker, Christopher H. Schmid, Hocine Tighiouart, John H. Eckfeldt, Harold I. Feldman, Tom Greene, John W. Kusek, Jane Manzi, Frederick Van Lente, Yaping Lucy Zhang, Josef Coresh, and Andrew S. Levey. Estimating Glomerular

Filtration Rate from Serum Creatinine and Cystatin C. *New England Journal of Medicine*, 367(1):20–29, 2012.

[17] Michelle J. Pena, Dick De Zeeuw, Harald Mischak, Joachim Jankowski, Rainer Oberbauer, Wolfgang Woloszczuk, Jacqueline Benner, Guido Dallmann, Bernd Mayer, Gert Mayer, Peter Rossing, and Hiddo J. Lambers Heerspink. Prognostic clinical and molecular biomarkers of renal disease in type 2 diabetes. *Nephrology Dialysis Transplantation*, 30:iv86–iv95, 2015.

[18] Majda Haznadar, Padma Maruvada, Eliza Mette, John Milner, Steven C. Moore, Holly L. Nicastro, Joshua N. Sampson, L. Joseph Su, Mukesh Verma, and Krista A. Zanetti. Navigating the road ahead: addressing challenges for use of metabolomics in epidemiology studies. *Metabolomics*, 10(2):176–178, apr 2014.

[19] Hayley Abbiss, Garth L. Maker, and Robert D. Trengove. Metabolomics approaches for the diagnosis and understanding of kidney diseases. *Metabolites*, 9(2), 2019.

[20] Helen M. Colhoun and M. Loredana Marcovecchio. Biomarkers of diabetic kidney disease. *Diabetologia*, 61(5):996–1011, 2018.

[21] Akiyoshi Hirayama, Eitaro Nakashima, Masahiro Sugimoto, Shin Ichi Akiyama, Waichi Sato, Shoichi Maruyama, Seiichi Matsuo, Masaru Tomita, Yukio Yuzawa, and Tomoyoshi Soga. Metabolic profiling reveals new serum biomarkers for differentiating diabetic nephropathy. *Analytical and Bioanalytical Chemistry*, 404(10):3101–3109, 2012.

[22] Sahir Kalim and Eugene P. Rhee. An overview of renal metabolomics. *Kidney International*, 91(1):61–69, 2017.

[23] Yumin Zhang, Siwen Zhang, and Guixia Wang. Metabolomic biomarkers in diabetic kidney diseases - A systematic review. *Journal of Diabetes and its Complications*, 29(8):1345–1351, 2015.

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

[25] Brian Kwan, Tobias Fuhrer, Jing Zhang, Manjula Darshi, Benjamin Van Espen, Daniel Montemayor, Ian H. de Boer, Mirela Dobre, Chi yuan Hsu, Tanika N. Kelly, Dominic S. Raj, Panduranga S. Rao, Santosh L. Saraf, Julia Scialla, Sushrut S. Waikar, Kumar Sharma, Loki Natarajan, Lawrence J. Appel, Harold I. Feldman, Alan S. Go, Jiang He, James P. Lash, Mahboob Rahman, and Raymond R. Townsend. Metabolomic Markers of Kidney Function Decline in Patients With Diabetes: Evidence From the Chronic Renal Insufficiency Cohort (CRIC) Study. *American Journal of Kidney Diseases*, 76(4):511–520, 2020.

[26] Kumar Sharma, Bethany Karl, Anna V. Mathew, Jon A. Gangoiti, Christina L. Wassel, Rintaro Saito, Minya Pu, Shoba Sharma, Young Hyun You, Lin Wang, Maggie Diamond-Stanic, Maja T. Lindenmeyer, Carol Forsblom, Wei Wu, Joachim H. Ix, Trey Ideker, Jeffrey B. Kopp, Sanjay K. Nigam, Clemens D. Cohen, Per Henrik Groop, Bruce A. Barshop, Loki Natarajan, William L. Nyhan, and Robert K. Naviaux. Metabolomics reveals signature of mitochondrial dysfunction in diabetic kidney disease. *Journal of the American Society of Nephrology*, 24(11):1901–1912, 2013.

[27] Timothy J. Lyons and Arpita Basu. Biomarkers in diabetes: Hemoglobin A1c, vascular and tissue markers. *Translational Research*, 159(4):303–312, 2012.

[28] M. Frank O'Brien, Angel M. Cronin, Paul A. Fearn, Caroline J. Savage, Brandon Smith, Jason Stasi, Peter T. Scardino, Gabrielle Fisher, Jack Cuzick, Henrik Møller, R. Timothy Oliver, Daniel M. Berney, Christopher S. Foster, James A. Eastham, Andrew J. Vickers, and Hans Lilja. Evaluation of prediagnostic prostate-specific antigen dynamics as predictors of death from prostate cancer in patients treated conservatively. *International Journal of Cancer*, 128(10):2373–2381, 2011.

[29] Matthew Denker, Suzanne Boyle, Amanda H. Anderson, Lawrence J. Appel, Jing Chen, Jeffrey C. Fink, John Flack, Alan S. Go, Edward Horwitz, Chi Yuan Hsu, John W. Kusek, James P. Lash, Sankar Navaneethan, Akinlolu O. Ojo, Mahboob Rahman, Susan P. Steigerwalt, Raymond R. Townsend, and Harold I. Feldman. Chronic renal insufficiency cohort study (CRIC): Overview and summary of selected findings. *Clinical Journal of the American Society of Nephrology*, 10(11):2073–2083, 2015.

[30] H. I. Feldman. The Chronic Renal Insufficiency Cohort (CRIC) Study: Design and Methods. *Journal of the American Society of Nephrology*, 14(90002):148S–153, jul 2003.

[31] James P. Lash, Alan S. Go, Lawrence J. Appel, Jiang He, Akinlolu Ojo, Mahboob Rahman, Raymond R. Townsend, Dawei Xie, Denise Cifelli, Janet Cohan, Jeffrey C. Fink, Michael J. Fischer, Crystal Gadegbeku, L. Lee Hamm, John W. Kusek, J. Richard Landis, Andrew Narva, Nancy Robinson, Valerie Teal, and Harold I. Feldman. Chronic Renal Insufficiency Cohort (CRIC) Study: Baseline Characteristics and Associations with Kidney Function. *Clinical Journal of the American Society of Nephrology*, 4(8):1302–1311, aug 2009.

[32] Amanda H. Anderson, Dawei Xie, Xue Wang, Robin L. Baudier, Paula Orlandi, Lawrence J. Appel, Laura M. Dember, Jiang He, John W. Kusek, James P. Lash, Sankar D. Navaneethan, Akinlolu Ojo, Mahboob Rahman, Jason Roy, Julia J. Scialla, James H. Sondheimer, Susan P. Steigerwalt, F. Perry Wilson, Myles Wolf, and Harold I. Feldman. Novel Risk Factors for Progression of Diabetic and Nondiabetic CKD: Findings From the Chronic Renal Insufficiency Cohort (CRIC) Study. *American Journal of Kidney Diseases*, 2020.

[33] A. de Hauteclocque, S. Ragot, Y. Slaoui, E. Gand, A. Miot, P. Sosner, J. M. Halimi, P. Zaoui, V. Rigalleau, R. Roussel, P. J. Saulnier, S. Hadjadj Samy, and Jeffrey Arsham. The influence of sex on renal function decline in people with Type 2 diabetes. *Diabetic Medicine*, 31(9):1121–1128, 2014.

[34] Andreas Heinzel, Michael Kammer, Gert Mayer, Roman Reindl-Schwaighofer, Karin Hu, Paul Perco, Susanne Eder, Laszlo Rosivall, Patrick B. Mark, Wenjun Ju, Matthias Kretzler, Peter Gilmour, Jonathan M. Wilson, Kevin L. Duffin, Moustafa Abdalla, Mark I. McCarthy, Georg Heinze, Hiddo L. Heerspink, Andrzej Wiecek, Maria F. Gomez, and Rainer Oberbauer. Validation of plasma biomarker candidates for the prediction of eGFR decline in patients with type 2 diabetes. *Diabetes Care*, 41(9):1947–1954, 2018.

[35] Takeshi Osonoi, Miyoko Saito, Yusuke Osonoi, Satako Douguchi, Kensuke Ofuchi, and Makoto Katoh. Liraglutide Improves Estimated Glomerular Filtration Rate Slopes in Patients with Chronic Kidney Disease and Type 2 Diabetes: A 7-Year Retrospective Analysis. *Diabetes Technology & Therapeutics*, 22(11):1–7, 2020.

[36] Afshin Parsa, W. H.Linda Kao, Dawei Xie, Brad C. Astor, Man Li, Chi Yuan Hsu, Harold I. Feldman, Rulan S. Parekh, John W. Kusek, Tom H. Greene, Jeffrey C. Fink, Amanda H. Anderson, Michael J. Choi, Jackson T. Wright, James P. Lash, Barry I. Freedman, Akinlolu Ojo, Cheryl A. Winkler, Dominic S. Raj, Jeffrey B. Kopp, Jiang He, Nancy G. Jensvold, Kaixiang Tao, Michael S. Lipkowitz, and Lawrence J. Appel. APOL1 risk variants, race, and progression of chronic kidney disease. *New England Journal of Medicine*, 369(23):2183–2196, 2013.

[37] A. Sayers, J. Heron, A. D.A.C. Smith, C. Macdonald-Wallis, M. S. Gilthorpe, F. Steele, and K. Tilling. Joint modelling compared with two stage methods for analysing longitudinal data and prospective outcomes: A simulation study of childhood growth and BP. *Statistical Methods in Medical Research*, 26(1):437–452, 2017.

[38] Garrett M. Fitzmaurice, Nan M. Laird, and James H. Ware. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA, aug 2011.

[39] James R. Carpenter, Harvey Goldstein, and Jon Rasbash. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(4):431–443, oct 2003.

[40] R Core Team. R: A Language and Environment for Statistical Computing, 2019.

[41] Morgan E. Grams, Tariq Shafi, and Eugene P. Rhee. Metabolomics Research in Chronic Kidney Disease. *Journal of the American Society of Nephrology*, 29(6):1588–1590, jun 2018.

[42] Bartolomé R. Celli, Nicola E. Thomas, Julie A. Anderson, Gary T. Ferguson, Christine R. Jenkins, Paul W. Jones, Jørgen Vestbo, Katharine Knobil, Julie C. Yates, and Peter M.A. Calverley. Effect of pharmacotherapy on rate of decline of lung function in chronic obstructive pulmonary disease: Results from the TORCH study. *American Journal of Respiratory and Critical Care Medicine*, 178(4):332–338, 2008.

[43] Liang Li, Brad C. Astor, Julia Lewis, Bo Hu, Lawrence J. Appel, Michael S. Lipkowitz, Robert D. Toto, Xuelei Wang, Jackson T. Wright, and Tom H. Greene. Longitudinal progression trajectory of GFR among patients with CKD. *American Journal of Kidney Diseases*, 59(4):504–512, 2012.

[44] John S. Tamaresis, Juan C. Irwin, Gabriel A. Goldfien, Joseph T. Rabban, Richard O. Burney, Camran Nezhat, Louis V. DePaolo, and Linda C. Giudice. Molecular classification of endometriosis and disease stage using high-dimensional genomic data. *Endocrinology*, 155(12):4986–4999, 2014.

[45] Haitian Wang, Shaw Hwa Lo, Tian Zheng, and Inchi Hu. Interaction-based feature selection and classification for high-dimensional biological data. *Bioinformatics*, 28(21):2834–2842, 2012.

[46] Jan Kalina. Classification methods for high-dimensional genetic data. *Biocybernetics and Biomedical Engineering*, 34(1):10–18, 2014.

[47] Yifeng Li, Fang Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2):325–340, 2018.

[48] Marcel Dettling and Peter Bühlmann. Boosting for tumor classification with gene expression data. *Bioinformatics*, 19(9):1061–1069, 2003.

[49] Heping Zhang, Chang Yung Yu, and Burton Singer. Cell and tumor classification using gene expression data: Construction of forests. *Proceedings of the National Academy of Sciences of the United States of America*, 100(7):4168–4172, 2003.

[50] Sihua Peng, Qianghua Xu, Xuefeng Bruce Ling, Xiaoning Peng, Wei Du, and Liangbiao Chen. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letters*, 555(2):358–362, 2003.

[51] Chen Hsiang Yeang, Sridhar Ramaswamy, Pablo Tamayo, Sayan Mukherjee, Ryan M. Rifkin, Michael Angelo, Michael Reich, Eric Lander, Jill Mesirov, and Todd Golub. Molecular classification of multiple tumor types. *Bioinformatics*, 17:S316–S322, 2001.

[52] Donald Geman, Christian D'Avignon, Daniel Q. Naiman, and Raimond L. Winslow. Classifying Gene Expression Profiles from Pairwise mRNA Comparisons. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–19, jan 2004.

[53] Aik Choon Tan, Daniel Q. Naiman, Lei Xu, Raimond L. Winslow, and Donald Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, 2005.

[54] Lei Xu, Aik Choon Tan, Daniel Q. Naiman, Donald Geman, and Raimond L. Winslow. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, 21(20):3905–3911, 2005.

[55] Lucas B Edelman, Giuseppe Toia, Donald Geman, Wei Zhang, and Nathan D Price. Two-transcript gene expression classifiers in the diagnosis and prognosis of human diseases. *BMC Genomics*, 10(1):583, 2009.

[56] Bahman Afsari, Ulisses M. Braga-Neto, and Donald Geman. Rank discriminants for predicting phenotypes from RNA expression. *The Annals of Applied Statistics*, 8(3):1469–1491, sep 2014.

[57] Bahman Afsari, Elana J. Fertig, Donald Geman, and Luigi Marchionni. SwitchBox: An R package for k-Top Scoring Pairs classifier development. *Bioinformatics*, 31(2):273–274, 2015.

[58] Tobias Fuhrer, Dominik Heer, Boris Begemann, and Nicola Zamboni. High-throughput, accurate mass metabolome profiling of cellular extracts by flow injection-time-of-flight mass spectrometry. *Analytical Chemistry*, 83(18):7074–7080, 2011.

[59] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[60] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[61] Dharmeshkumar Parmar, Nivedita Bhattacharya, Shanthini Kannan, Sangeetha Vadivel, Gautam Kumar Pandey, Avinash Ghanate, Nagarjuna Chary Ragi, Paramasivam Prabu, Thyparambil Aravindakshan Pramodkumar, Nagaraj Manickam, Viswanathan Mohan, Prabhakar Sripadi, Gokulakrishnan Kuppan, and Venkateswarlu Panchagnula. Plausible diagnostic value of urinary isomeric dimethylarginine ratio for diabetic nephropathy. *Scientific Reports*, 10(1):1–7, 2020.

[62] National Center for Biotechnology Information. PubChem Compound Summary for CID 22425, Pipazethate.

[63] David S. Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, Chris Fung, Lisa Nikolai, Mike Lewis, Marie Aude Coutouly, Ian Forsythe, Peter Tang, Savita Shrivastava, Kevin Jeroncic, Paul Stothard, Godwin Amegbey, David Block, David D. Hau, James Wagner, Jessica Miniaci, Melisa Clements, Mulu Gebremedhin, Natalie Guo, Ying Zhang, Gavin E. Duggan, Glen D. MacInnis, Alim M. Weljie, Reza Dowlatabadi, Fiona Bamforth, Derrick Clive, Russ Greiner, Liang Li, Tom Marrie, Brian D. Sykes, Hans J. Vogel, and Lori Querengesser. HMDB: The human metabolome database. *Nucleic Acids Research*, 35(SUPPL. 1):521–526, 2007.

[64] David S. Wishart, Craig Knox, An Chi Guo, Roman Eisner, Nelson Young, Bijaya Gautam, David D. Hau, Nick Psychogios, Edison Dong, Souhaila Bouatra, Rupasri Mandal, Igor Sinelnikov, Jianguo Xia, Leslie Jia, Joseph A. Cruz, Emilia Lim, Constance A. Sobsey, Savita Shrivastava, Paul Huang, Philip Liu, Lydia Fang, Jun Peng, Ryan Fradette, Dean Cheng, Dan Tzur, Melisa Clements, Avalyn Lewis, Andrea De souza, Azaret Zuniga, Margot Dawe, Yeping Xiong, Derrick Clive, Russ Greiner, Alsu Nazyrova, Rustem Shaykhutdinov, Liang Li, Hans J. Vogel, and Ian Forsythei. HMDB: A knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(SUPPL. 1):603–610, 2009.

[65] David S. Wishart, Timothy Jewison, An Chi Guo, Michael Wilson, Craig Knox, Yifeng Liu, Yannick Djoumbou, Rupasri Mandal, Farid Aziat, Edison Dong, Souhaila Bouatra, Igor Sinelnikov, David Arndt, Jianguo Xia, Philip Liu, Faizath Yallou, Trent Bjorndahl, Rolando Perez-Pineiro, Roman Eisner, Felicity Allen, Vanessa Neveu, Russ Greiner, and Augustin Scalbert. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Research*, 41(D1):801–807, 2013.

[66] David S. Wishart, Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, Kevin Liang, Rosa Vázquez-Fresno, Tanvir Sajed, Daniel Johnson, Carin Li, Naama Karu, Zinat Sayeeda, Elvis Lo, Nazanin Assempour, Mark Berjanskii, Sandeep Singhal, David Arndt, Yonjie Liang, Hasan Badran, Jason Grant, Arnau Serra-Cayuela, Yifeng Liu, Rupa Mandal, Vanessa Neveu, Allison Pon, Craig Knox, Michael Wilson, Claudine Manach, and Augustin Scalbert. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Research*, 46(D1):D608–D617, 2018.

[67] Josephine Asafu-Adjei. *Covariate Adjusted Discrimination with Applications to Neuroscience*. PhD thesis, University of Pittsburgh, 2012.

[68] Josephine K. Asafu-Adjei, Allan R. Sampson, Robert A. Sweet, and David A. Lewis. Adjusting for matching and covariates in linear discriminant analysis. *Biostatistics*, 14(4):779–791, 2013.

[69] Peter A. Lachenbruch. Covariance adjusted discriminant functions. *Annals of the Institute of Statistical Mathematics*, 29(1):247–257, dec 1977.

[70] X. M. Tu, J Kowalski, J Randall, J. Mendoza-Blanco, M K Shear, T. H. Monk, E Frank, and D J Kupfer. Generalized Covariance-Adjusted Discriminants: Perspective and Application. *Biometrics*, 53(3):900, sep 1997.

[71] Josephine K. Asafu-Adjei and Allan R. Sampson. Covariate adjusted classification trees. *Biostatistics*, 19(1):42–53, 2018.

[72] Trevor Hastie and Robert Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, apr 1998.

[73] Ting Fan Wu, Chih Jen Lin, and Ruby C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

[74] Ian H. de Boer, Ronit Katz, Linda F. Fried, Joachim H. Ix, Jose Luchsinger, Mark J. Sarnak, Michael G. Shlipak, David S. Siscovick, and Bryan Kestenbaum. Obesity and Change in Estimated GFR Among Older Adults. *American Journal of Kidney Diseases*, 54(6):1043–1051, 2009.

[75] Cassianne Robinson-Cohen, Alyson J. Littman, Glen E. Duncan, Noel S. Weiss, Michael C. Sachs, John Ruzinski, John Kundzins, Denise Rock, Ian H. De Boer, T. Alp Ikizler, Jonathan Himmelfarb, and Bryan R. Kestenbaum. Physical activity and change in estimated GFR among persons with CKD. *Journal of the American Society of Nephrology*, 25(2):399–406, 2014.

[76] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York, 2nd edition, 2002.

[77] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer New York, New York, NY, 2005.

[78] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, jun 2016.

[79] Minggao Shi, Robert E. Weiss, and Jeremy M. G. Taylor. An Analysis of Paediatric CD4 Counts for Acquired Immune Deficiency Syndrome Using Flexible Random Curves. *Applied Statistics*, 45(2):151, 1996.

[80] Joan G. Staniswalis and J. Jack Lee. Nonparametric Regression Analysis of Longitudinal Data. *Journal of the American Statistical Association*, 93(444):1403–1418, 1998.

[81] Gareth M. James, Trevor J. Hastie, and Catherine A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87(3):587–602, 2000.

[82] John A Rice and Colin O. Wu. Unequally Sampled Noisy Curves. *Biometrics*, 57(March):253–259, 2001.

[83] Fang Yao, Hans Georg Müller, and Jane Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[84] Fang Yao and Thomas C.M. Lee. Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68(1):3–25, 2006.

[85] Jie Peng and Debashis Paul. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, 18(4):995–1015, 2009.

[86] Debashis Paul and Jie Peng. Consistency of restricted maximum likelihood estimators of principal components. *Annals of Statistics*, 37(3):1229–1271, 2009.

[87] Jianghu J. Dong, Liangliang Wang, Jagbir Gill, and Jiguo Cao. Functional principal component analysis of glomerular filtration rate curves after kidney transplant. *Statistical Methods in Medical Research*, 27(12):3785–3796, 2018.

[88] Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Technical report, Kidney inter., Suppl, 2013.

[89] Tomasz Górecki and Łukasz Smaga. A comparison of tests for the one-way ANOVA problem for functional data. *Computational Statistics*, 30(4):987–1010, 2015.

[90] Alessandra Cabassi, Davide Pigoli, Piercesare Secchi, and Patrick A. Carter. Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *Electronic Journal of Statistics*, 11(2):3815–3840, 2017.

[91] F Pesarin and L Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, Inc., 2010.

[92] Keiichi Sumida, Girish N. Nadkarni, Morgan E. Grams, Yingying Sang, Shoshana H. Ballew, Josef Coresh, Kunihiro Matsushita, Aditya Surapaneni, Nigel Brunskill, Steve J. Chadban, Alex R. Chang, Massimo Cirillo, Kenn B. Daratha, Ron T. Gansevoort, Amit X. Garg, Licia Iacoviello, Takamasa Kayama, Tsuneo Konta, Csaba P. Kovesdy, James Lash, Brian J. Lee, Rupert W. Major, Marie Metzger, Katsuyuki Miura, David M.J. Naimark, Robert G. Nelson, Simon Sawhney, Nikita Stempniewicz, Mila Tang, Raymond R. Townsend, Jamie P. Traynor, José M. Valdivielso, Jack Wetzels, Kevan R. Polkinghorne, and Hiddo J.L. Heerspink. Conversion of Urine Protein-Creatinine Ratio or Urine Dipstick Protein to Urine Albumin-Creatinine Ratio for Use in Chronic Kidney Disease Screening and Prognosis : An Individual Participant-Based Meta-analysis. *Annals of internal medicine*, 173(6):426–435, 2020.

[93] R Core Team. R: A Language and Environment for Statistical Computing, 2020.

[94] C. Carroll, A. Gajardo, Y. Chen, X. Dai, J. Fan, P.Z. Hadjipantelis, K. Han, H. Ji, H. Mueller, and J. Wang. fdapace: Functional Data Analysis and Empirical Dynamics, 2021.

[95] Tomasz Górecki and Łukasz Smaga. fdANOVA: Analysis of Variance for Univariate and Multivariate Functional Data, 2018.

[96] A. Cabassi and A.B. Kashlak. fdcov: Analysis of Covariance Operators, 2017.